Reading Profiles in Multi-Site Data With Missingness

Mark A. Eckert[1]*, Kenneth I. Vaden Jr.[1], Mulugeta Gebregziabher[2] and Dyslexia Data Consortium

[1] *Hearing Research Program, Department of Otolaryngology-Head and Neck Surgery, Medical University of South Carolina, Charleston, SC, United States,* [2] *Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States*

Children with reading disability exhibit varied deficits in reading and cognitive abilities that contribute to their reading comprehension problems. Some children exhibit primary deficits in phonological processing, while others can exhibit deficits in oral language and executive functions that affect comprehension. This behavioral heterogeneity is problematic when missing data prevent the characterization of different reading profiles, which often occurs in retrospective data sharing initiatives without coordinated data collection. Here we show that reading profiles can be reliably identified based on Random Forest classification of incomplete behavioral datasets, after the missForest method is used to multiply impute missing values. Results from simulation analyses showed that reading profiles could be accurately classified across degrees of missingness (e.g., ~5% classification error for 30% missingness across the sample). The application of missForest to a real multi-site dataset with missingness ($n = 924$) showed that reading disability profiles significantly and consistently differed in reading and cognitive abilities for cases with and without missing data. The results of validation analyses indicated that the reading profiles (cases with and without missing data) exhibited significant differences for an independent set of behavioral variables that were not used to classify reading profiles. Together, the results show how multiple imputation can be applied to the classification of cases with missing data and can increase the integrity of results from multi-site open access datasets.

Keywords: missingness, multiple imputation, classification, dyslexia, big data, reading profiles

INTRODUCTION

Not all reading disabilities are the same. Children with reading disability can exhibit: (1) relatively specific phonological and spelling impairments (Stanovich, 1988; Fletcher, 2009; Lauterbach et al., 2017); (2) general language or language-learning impairments (Bishop and Snowling, 2004; Berninger and May, 2011), and (3) comprehension-specific impairments (Catts et al., 2012). These observations are generally consistent with hypotheses that phonological decoding and non-phonological language skills are unique and additive predictors of reading (Gough and Tunmer, 1986; Bishop and Snowling, 2004), although executive functions also appear to contribute to comprehension (Cain et al., 2004; Cutting et al., 2009; Locascio et al., 2010). The inclusion of children with these different reading disability profiles can be problematic for the outcomes of a study without a way to identify and account for different reading profiles that may have different etiologies (Samuelsson et al., 2007; Naples et al., 2009). For example, the study of retrospective

multi-site datasets can include data from children with different reading profiles, particularly if sampling or ascertainment criteria were different across studies. This problem can be compounded if there is completely missing data from contributing research sites that could help differentiate reading disability profiles. Here we examined methods for dealing with missing data to identify reading profiles in a retrospective multi-site dataset.

Reading disability profiles have been characterized based on a variety of methods and measures that are sensitive to differences in reading skill, memory, perceptual, and general cognitive function (Korkman and Häkkinen-Rihu, 1994; King et al., 2007; Talcott et al., 2013; Miciak et al., 2015; Kornilov and Grigorenko, 2017). For example, trajectory analyses were used to differentiate reading disability cases with and without quite low general cognitive function (Kuppen and Goswami, 2016). Torppa et al. (2007) identified unique reading disability profiles, which included cases with and without low general cognitive function, using a latent class modeling approach. Or for example, Eckert et al. (2017) used a machine learning approach to verify clinician ratings of reading disability profiles that included relatively specific deficits in phonological decoding (Poor Decoder), low performance across reading and cognitive measures (Generally Poor Readers), and children with comprehension-specific reading disability (Poor Comprehenders).

While these reading disability profiles can be reliably classified on the basis of standardized test scores, missing data presents a major challenge for classification. This is particularly true for multi-site retrospective data analyses, when data collection was not coordinated before conducting studies. For example, missing a critical variable that is key for distinguishing between two profiles could prevent classification unless there is a method to deal with the missingness. On the other hand, a complete case analysis approach that ignores cases with missing scores could result in the exclusion of large numbers of children with a specific reading disability profile from multi-site datasets (e.g., Eckert et al., 2017). Such a strict inclusion policy introduces a false negative bias because of low power and this could be potentially mitigated by using multiply imputed data. The feasibility of identifying distinct reading profiles when there is missing data has not been examined, however.

In the current study, we examined three strategies for imputing missing data (mean replacement, multiple imputation with Predictive Mean Matching, multiple imputation with the missForest method) and the degree to which imputed data could then be used to reliably classify reading profiles. Multiple imputation is a statistically principled technique to estimate predictably missing values, usually by means of an informed regression model with an error term (Rubin, 1987; Little, 1988; Little and Rubin, 2002). The imputation model is used to produce multiple (m) simulated versions of the complete dataset that fill in missing values with new random draws that approximate the variance of the originally observed (i.e., non-missing) values. Each of the imputed datasets are separately analyzed and then the results across datasets are pooled (Little and Rubin, 2002). Multiple imputation produces low false negative and false positive rates when less than

30% of the data are missing across a sample (Vaden et al., 2012). Based on the intuition that a multivariate imputation model could be more effective at approximating multifactor patterns that are critical for classification analyses, we also tested the missForest algorithm for imputing missing data. The missForest method uses a Random Forest imputation model to estimate missing values based on decision trees instead of univariate regression models (Stekhoven and Bühlmann, 2011).

We examined the extent to which Poor Decoders, Generally Poor Readers, Poor Comprehenders, and Controls were consistently classified based on a retrospective multi-site dataset with missing data (Eckert et al., 2016, 2017). We first examined classification accuracy in a synthetic dataset to demonstrate the relative advantage of missForest with the current dataset compared to other imputation approaches. We then applied missForest to a dataset with real missingness and compared the behavioral profiles of cases with and without missingness. Finally, we examined the extent to which the reading disability profiles could be validated using a set of independent measures of spelling and spatial cognition from the multi-site dataset.

MATERIALS AND METHODS

The data used in the current study were collected as part of a research project to develop methods for retrospective multi-site datasets that are increasingly a focus of data sharing initiatives. These data were contributed by members of the Dyslexia Data Consortium and included data from studies of reading disability group differences and dimensional studies of reading development. Specifically, these were studies in which participants were recruited for neuroimaging studies of dyslexia because they had received a clinical diagnosis of dyslexia, were enrolled in schools specializing in the education of children with dyslexia, or they were recruited from the local community for studies on reading development. Thus, the Dyslexia Data Consortium dataset is composed of participants who were recruited with varied sampling approaches. Additional information about the Dyslexia Data Consortium, including how to request the data and how to access the code used in this study can be found at www.dyslexiadata.org. Institutional Review Board (IRB) approval was obtained at each contributing site to share de-identified data. In addition, Medical University of South Carolina IRB approval was obtained to receive and analyze the de-identified data. We specifically requested word and pseudoword reading accuracy, rapid naming, and Verbal Comprehension (Verbal IQ) to determine the extent to which unique reading disability profiles could be reliably identified with these data based on the theoretical framework described above. We also requested any additional data that investigators could share. This resulted in a predominance of word reading accuracy and Verbal IQ measures, and other behavioral measures that were used in this study to provide external validity for the reading profiles.

The current study consisted of simulation and real dataset analyses. The simulation analyses were performed based on data that included 198 children from 8 different research sites (78 female; ages 6.39–12.85) without any missingness. These data were used in Eckert et al. (2017) to define reading profiles and allowed for the creation of simulated missingness to compare missingness approaches. Secondly, Random Forest was used to classify reading profiles in a sample of 726 children from 11 different research sites (334 female; ages 6.02–12.94 years), which excluded any cases from the Random Forest training dataset. Importantly, the test dataset included cases with missing data that were imputed using the best method from the simulation results (missForest). The final analysis examined differences for behavioral measures that were not used to make the reading profile classifications from the 924 children whose data were included in the training and test datasets (11 research sites; 412 female; ages 6.02–12.94 years). Although there was extensive missingness in the combined validation dataset, this provided supporting evidence for the distinct reading profiles that were classified with observed and imputed data.

Reading Profile Labels

The method for classifying reading profiles was developed in Eckert et al. (2017). These methods are presented here to explain how the reading profile labels were obtained and the accuracy of this approach. The Random Forest reading profile classifications from Eckert et al. (2017) were used in the simulated and real data analyses for this study.

An expert-labeled training dataset was required to generate synthetic datasets and train a machine learning algorithm to classify reading profiles. The 198 children included in this training dataset had standardized scores from the (1) Woodcock–Johnson IIIR or Woodcock Reading Mastery Tests [Word Attack (pseudoword ID), Word Identification (real word ID), and Passage Comprehension; Woodcock, 1987; Woodcock et al., 2001], (2) Wechsler Intelligence Scales for Children or Wechsler Abbreviated Scales of Intelligence (Verbal Comprehension and Perceptual Reasoning or Verbal IQ; Wechsler, 1999, 2004), and (3) a standardized measure of Rapid Naming of letters or numbers (Wagner et al., 1999; Wolf and Denckla, 2005). Guided by the clinical and research expertise of a Neuropsychologist, the extant literature described above, and the age of the children, two clinically certified Speech-Language Pathologists labeled cases into the following theoretically defined profiles: (1) Poor Decoders, (2) Poor Comprehenders, (3) Generally Poor Readers, and (4) Controls. The resultant training dataset was composed of 58 Poor Decoders, 27 Poor Comprehenders, 14 Generally Poor Readers, and 99 Controls.

Random Forest was used to classify reading disability cases and controls, which matched the expert labels with 94% accuracy, based on leave-one-out cross validation tests (as reported in Eckert et al., 2017). The R statistics software (R version 3.3.1 with packages: caret v6.0.71, randomForest v4.6.12) was used to train the Random Forest classifier (parameters: *mtry* = 2; *number of trees* = 500) after a cross-validated tuning was performed (fivefold; 10 repeats) with the training dataset described above.

The same classification training procedure was used for the synthetic and real data analyses in the current study.
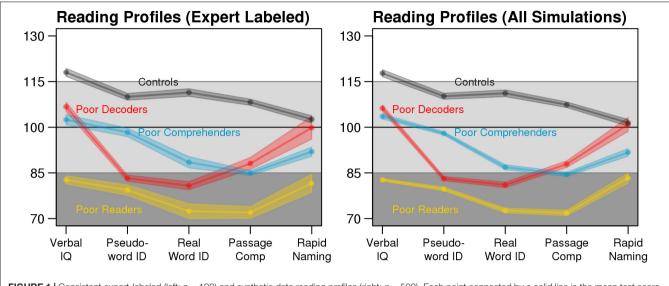
Simulation Analyses
Data Synthesis

A simulation study was performed to evaluate classification accuracy when imputation was used to deal with different levels of missingness. The data synthesis tool *synthpop* R-package (v1.3.1; Nowok et al., 2016) used in this study was originally developed to study epidemiologic, demographic, and health-related datasets without risk of re-identification as no case from an original dataset exactly re-occurs in the synthesized datasets. *synthpop* generates new datasets by estimating the value for each variable based on its association with the other variables in the data set. This regression modeling determines the conditional distributions separately for each variable and thus maintains the correlational structure of the original dataset. We generated 1,000 synthetic datasets based on the expert-labeled training dataset using the default *synthpop* settings. These datasets had $N = 500$ participants and an equal number of labeled cases for each reading profile (**Figure 1**). Because expert labels for the reading profiles were propagated from the training dataset into the synthetic dataset, the accuracy of reading profile classification with the synthetic data was calculated by comparing the synthesized expert labels to the Random Forest classifications. Furthermore, the synthetic dataset allowed us to systematically remove known values to simulate missingness and assess the performance of different imputation methods.

Simulated Missingness

To provide guidance about how the extent of missingness interacts with imputation performance and classification accuracy, we tested six levels of missingness within each synthetic dataset: 0, 10, 20, 30, 40, and 50%. Imputation methods are most appropriate when used to estimate values that are missing predictably in relation to known participant characteristics (missing at random), or when those values are missing completely at random (Little and Rubin, 2002). Imputation methods are less appropriate when missingness is not at random, which occurs when the values are systematically related to the missingness, because the true range of values is greater than what is represented in the dataset. An example of missingness not at random would be missing comprehension scores because children had such severe reading disability that they could not perform the task. Here we focused on missingness at random, which can be a significant limitation in retrospective multi-site datasets.

Three variables in each simulation had missing data, which could include any variable that was used for profile classification (i.e., each participant's age and five test scores). A missingness indicator (present = 0, missing = 1) represented the real or missing value and was used to ensure that each participant was missing no more than three values. Logistic regression was used to verify that sex and research site predicted missingness ($p < 0.01$). These demographic variables were selected to ensure missingness was predictable to evaluate these methods for missingness at random. Sex and age were selected, in particular,

FIGURE 1 | Consistent expert-labeled (left: $n = 198$) and synthetic data reading profiles (right: $n = 500$). Each point connected by a solid line is the mean test score and the shaded area depicts the standard error of the mean (SEM). For the synthetic data profiles, average test scores and SEM were calculated for each dataset with no missing data, then averaged across the 1,000 datasets.

because they could explain why there would missing data for studies that are disproportionately represented by one sex or across age and therefore provide results that are generalizable to other studies.

Comparing Imputation Methods

Three imputation methods were tested to determine the approach that resulted in the highest reading profile classification accuracy when imputed datasets were submitted to a Random Forest classifier. (1) First, mean replacement was used to substitute each missing test score with the average observed score for that test. Although mean replacement is a common approach for analyzing data with missingness, this method is known to inflate Type I error rates by failing to approximate the distribution of values and underestimating its variance (Schafer and Graham, 2002). (2) Next, multiple imputation with Predictive Mean Matching (Little, 1988) was tested, which imputes missing values by drawing test scores from similar cases with no missing scores (R-package: MICE, v2.30; Van Buuren and Groothuis-Oudshoorn, 2011). (3) Finally, multiple imputation via the missForest method (R-package: missForest, v1.4; Stekhoven and Bühlmann, 2011) was tested. missForest fits a Random Forest model (Breiman, 2001) on all of the observed data to predict missing values for each factor with missing scores. Both Predictive Mean Matching and missForest imputation methods were used to generate ten imputed datasets that were each classified into reading profiles using a trained Random Forest model. Next, statistics calculated for each of the $m = 10$ imputed datasets were pooled to form point and variance estimates (Little and Rubin, 2002). No such averaging was performed with mean replacement, because this approach can only generate one dataset. Critically, the synthetic reading profile labels were concealed from each of the imputation methods, so imputed test scores were

informed by covariance in the test scores, participant age, participant sex, and research site, but not the reading profile labels.

Efficiency and classification accuracy were used to assess the imputation methods. Efficiency was calculated by dividing the variance of the imputed scores by the variance of the original scores that had no missing data, for each test. Imputation methods that produce values that are more variable than the original data tend to produce false negative biases in univariate tests, while methods that reduce variance relative to the true values can produce false positive biases. Because we removed known values to simulate missingness, variance in the imputed test scores could be directly compared to the variance for the original test scores. The reading profile classification accuracy, sensitivity, and specificity were computed by comparing the Random Forest classification results to the reading profile labels in each synthetic dataset.

After determining that missForest resulted in the most accurate reading profile classifications, we assessed how missingness within each variable affected the classification accuracy for each reading profile when values were imputed with missForest. This involved selectively pooling profile-specific classification accuracy from each simulation that was missing for a specific variable. This allowed us to determine if each reading profile could be classified with confidence when a key variable that differentiated two profiles was missing. Each of these results was averaged across 1,000 simulations to produce reliable estimates for the effects of the amount of missingness and imputation methods on reading profile classification.

Real Data Analyses

Multiple imputation was performed with the real dataset using missForest based on the simulation results described

below. As noted above, independent classification tests were performed using a sample that shared no cases with the training dataset. These independent real data included 726 participants with missing values for Word Attack (0.4%), Word ID (0.1%), Passage Comprehension (16.1%), Verbal IQ (16.6%), and RAN (35.5%). Older children were more likely to be missing Passage Comprehension [age: $r_{(724)} = 0.28$, $p = 2.75E-14$] and Verbal IQ data [age: $r_{(724)} = 0.19$, $p = 3.63E-7$], while younger children were more likely to be missing RAN [age: $r_{(724)} = -0.07$, $p = 0.04$]. Children with higher Verbal IQ scores were more likely to be missing Passage Comprehension data [$r_{(603)} = 0.17$, $p = 4.23E-5$]. In addition, research site significantly predicted missingness for Passage Comprehension (*Nagelkerke R-square* = 0.76, $p = 1.90E-85$), Verbal IQ (*Nagelkerke R-square* = 0.74, $p = 1.96E-83$), and RAN (*Nagelkerke R-square* = 0.55, $p = 2.80E-74$). Because the missingness largely resulted from research site differences in assessment plan, the missingness was predictable and missing at random. While possible, we had no information to suggest that missing data were missing not at random.

To validate reading profiles with independent behavioral measures, participants in the Random Forest training data were combined with the participants in the test dataset. This combined validation dataset included 924 participants and with the exception of Performance IQ scores (23.3% missing; Wechsler, 1999, 2004), there was substantial missingness for each variable used to independently validate the reading profiles. Specifically, Woodcock–Johnson Spelling (47.3% missing; Woodcock et al., 2001), Comprehensive Test of Phonological Processing Elision (39.8% missing; Wagner et al., 1999), and Wechsler Digit Span (70.7% missing; Wechsler, 1999, 2004) measures exhibited too much missingness to reliably impute data. Caution is warranted in interpreting results from these data, but they were useful in illustrating distinct patterns of performance across the reading profiles.

RESULTS

Synthetic Data Results
Imputation Diagnostics
The expert labeled reading profiles from the original dataset were well-approximated by the synthetic datasets used in the simulations (**Figure 1**, above). We compared the efficiency of each imputation method by calculating the ratio of the variance for the imputed test scores to the variance of the original test scores. Multiple imputation by Predictive Mean Matching most closely approximated the variance of the original dataset across levels of missingness (**Figure 2**). In contrast, Mean Replacement imputation demonstrated a steep decrease in efficiency as a function of percent missingness, simply because this method substituted all missing values with the mean observed value (variance = 0). missForest also underestimated the true variance in the data, although to a lesser extent than Mean Replacement. While underestimating variance can inflate the false positive rate for univariate statistics, the classification results below indicated that the Random Forest multivariate

classifications were not biased when using the missForest imputed data.

Classification With Simulated Missingness
After simulated missing test scores were imputed with each method, each synthetic case was classified into a reading profile using Random Forest. The highest classification accuracy was observed for the missForest imputed data, relative to the other missingness approaches. Across levels of simulated missingness, the smallest declines in classification accuracy, sensitivity, and specificity were observed for the missForest imputed data compared to the original synthetic data (**Figure 3**). These results suggest that classifications with missForest imputed data were the least biased and that missForest was most sensitive to the multivariate patterns that were present in the observed data.

After determining that test scores imputed with missForest resulted in the highest classification accuracy, we examined classification accuracy for each reading profile when specific test scores were imputed using missForest (**Figure 4**). Poor Decoders and Poor Comprehenders were less accurately classified when pseudoword ID scores were among the missing data, while Generally Poor Readers were less accurately classified when Verbal IQ was missing. There was no single variable that negatively affected Control classification accuracy, which was high even for 50% missingness. **Figure 4** shows that 30% missingness would increase classification error by ∼5% across measures and reading profiles.

Real Data Results
Classification of Real Data With Missingness
Multiple imputation increased the available sample size from 327 participants with complete data to 726 participants (**Table 1**), excluding the 198 cases used to train the classifier. The reading profiles exhibited similar patterns of behavioral strengths and weaknesses when comparing the training dataset to the independent test dataset with no missing data and the test dataset with imputed scores (**Figure 5**), as well as the synthetic data (**Figure 1**). For example, Poor Decoder and Generally Poor Reader differences in Verbal IQ were significantly different from the cases whose data were imputed [ACA test dataset: $t_{(49)} = 8.41$, Cohen's $d = 2.54$, $p = 4.60E-11$; MI: $t_{(118.2)} = 11.66$, Cohen's $d = 2.58$, $p = 2.28E-21$]. These results indicate that missForest accurately estimated the values for missing variables that were critical for distinguishing the reading profiles (**Figure 5**).

Another demonstration of the effectiveness of missForest was the consistent reading profile classifications across the ten imputed datasets, as reflected in the low standard deviations for the classification labels (<2 in **Table 1**). In addition, most of the correlations between behavioral measures were consistent across ACA and MI datasets (Supplementary Figure S1). For example, Verbal IQ and real word ID were significantly and positively related in Controls [ACA: $r_{(232)} = 0.55$, $p = 3.76E-20$; MI: $r_{(526.2)} = 0.40$, $p = 3.69E-22$], significantly and negatively related in Poor Comprehenders [ACA: $r_{(41)} = -0.35$, $p = 0.02$; MI: $r_{(75.6)} = -0.33$, $p = 0.004$], and non-significantly related in Poor Decoders [ACA: $r_{(33)} = -0.22$, $p = 0.21$; MI: $r_{(92.0)} = 0.16$, $p = 0.13$] and Generally Poor Readers [ACA: $r_{(14)} = -0.45$,
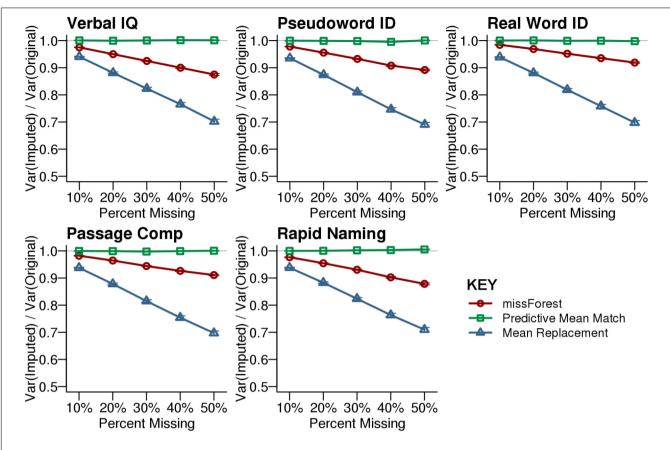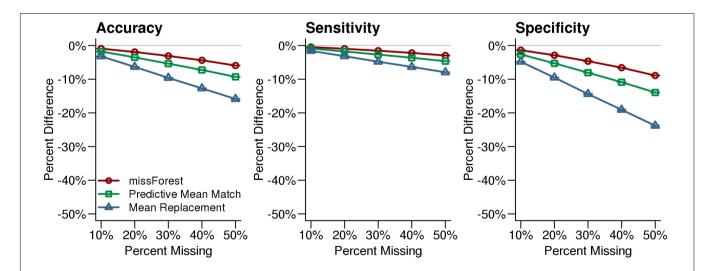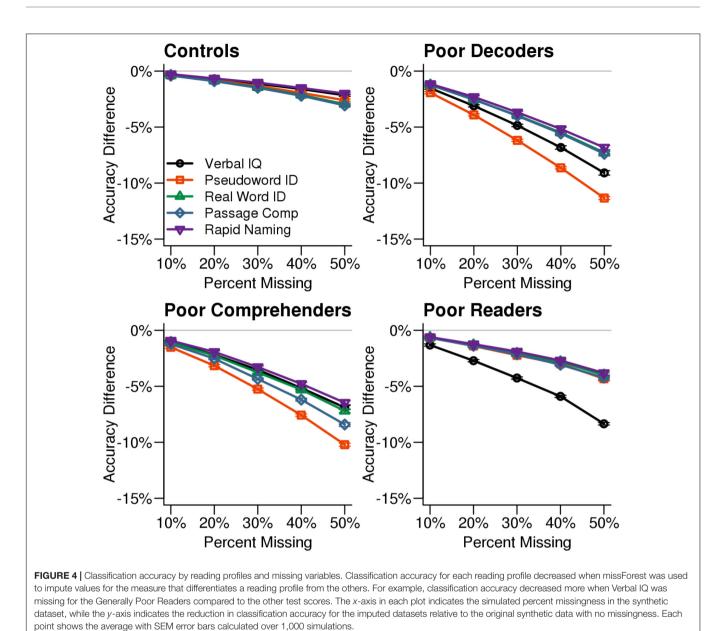
FIGURE 2 | Efficiency for the imputed synthetic data. Efficiency is calculated by dividing the variance of the imputed dataset by the variance of the original dataset that had no missing data. Predictive Mean Matching provided the best approximation of the original variance across the tested levels of simulated missingness at random. The missForest imputed dataset exhibited less variance than the original dataset. Mean Replacement demonstrated the lowest efficiency with increasing missingness. Each point shows the average with SEM error bars calculated across 1,000 simulations.

FIGURE 3 | Reading profile classification accuracy for the imputed synthetic datasets. Reading profile classifications based on synthetic datasets with no missing data were compared to classifications that were performed after imputing missing values with missForest, Predictive Mean Matching, and Mean Replacement. The missForest imputation (red line) resulted in the most accurate classifications. The *x*-axis in each plot indicates the simulated percent missingness, while the *y*-axis indicates the reduction in classification accuracy, sensitivity, and specificity for the imputed datasets compared with the complete synthetic dataset. Each point shows the average with SEM error bars calculated over 1,000 simulations.
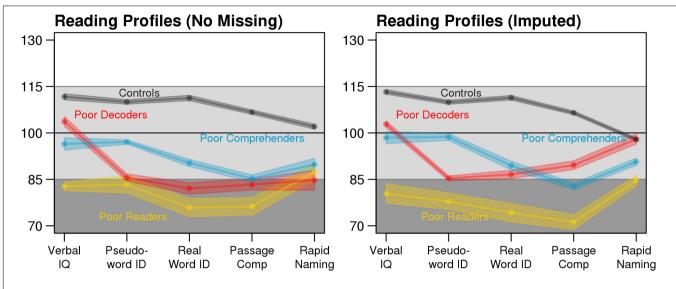
FIGURE 4 | Classification accuracy by reading profiles and missing variables. Classification accuracy for each reading profile decreased when missForest was used to impute values for the measure that differentiates a reading profile from the others. For example, classification accuracy decreased more when Verbal IQ was missing for the Generally Poor Readers compared to the other test scores. The *x*-axis in each plot indicates the simulated percent missingness in the synthetic dataset, while the *y*-axis indicates the reduction in classification accuracy for the imputed datasets relative to the original synthetic data with no missingness. Each point shows the average with SEM error bars calculated over 1,000 simulations.

TABLE 1 | Participant counts for each profile based on multiply imputed test scores or observed scores.

Analysis approach	Poor decoders	Poor comprehenders	Generally poor readers	Controls	Total *N*
Training Dataset	58	27	14	99	198
^MI (mean ± *SD*)	94 ± 1.1	77.6 ± 1.0	26.2 ± 0.8	528.2 ± 0.9	726
ACA	35	43	16	233	327

^MI: multiple imputation: Random Forest classifications based on missForest inputed data (pooled across ten imputed dataset). ACA, available case analysis. The independent test dataset for the MI and ACA analyses did not include participants from the reading profile training dataset.
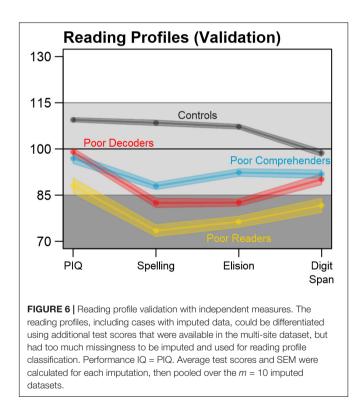
$p = 0.08$; MI: $r_{(24.2)} = -0.36$, $p = 0.07$]. These results show that missForest preserved the covariance structure between the behavioral variables within each reading profile.

Validation of Reading Profiles

Behavioral measures that were not used to classify reading profiles were also examined to validate the profiles and determine the extent to which consistent differences were observed across complete case and multiple imputation analysis approaches. **Figure 6** and **Table 2** show that there were significant and unique reading profile differences for Performance IQ, Spelling, Elision, and Digit Span measures. Reading disability cases exhibited consistently lower scores compared to controls across these measures. Generally Poor Readers were consistently the

FIGURE 5 | The same reading profiles are observed in the training dataset (*N* = 198) and MI test dataset (*N* = 726; pooled estimates from the *m* = 10 missForest imputations). Shaded area depicts the SEM.

FIGURE 6 | Reading profile validation with independent measures. The reading profiles, including cases with imputed data, could be differentiated using additional test scores that were available in the multi-site dataset, but had too much missingness to be imputed and used for reading profile classification. Performance IQ = PIQ. Average test scores and SEM were calculated for each imputation, then pooled over the *m* = 10 imputed datasets.

lowest performers across measures. Generally Poor Readers and Poor Decoders demonstrated significantly lower Elision performance compared to Poor Comprehenders. In addition, Poor Decoders could be differentiated from Generally Poor Readers because of significantly better Performance IQ and Digit Span performance than the Generally Poor Readers. These results were observed for both the complete case and multiply imputed datasets. Caution is appropriate in interpreting

the patterns of reading profile differences for these variables because of the significant extent of missingness, which prevented these data from being imputed or informing the Random Forest classification model. Nonetheless, the results indicate that the four reading profiles could be distinguished with an independent set of behavioral measures, thereby providing validation of the four reading profiles in this multi-site dataset.

DISCUSSION

Children with reading disability exhibit heterogeneous deficits in component reading skills that contribute to the presentation of unique reading profiles. These reading profiles could advance understanding for the etiologies of reading disability because they have been linked to different patterns of brain morphology (Leonard et al., 2006; Leonard and Eckert, 2008; Jednoróg et al., 2014; Bailey et al., 2016) and genetic findings (Samuelsson et al., 2007; Naples et al., 2009). Although we were not able to identify the etiology of the reading profiles examined in this study, or characterize the presence of any additional reading profiles in the current dataset (Torppa et al., 2007; Wolff, 2010; Tamboer et al., 2016; Kornilov and Grigorenko, 2017), we have demonstrated that the reading profiles described in the current study can be reliably and consistently classified when there is ∼30% missingness. Classification accuracy exhibited limited decline when missing data were imputed with missForest, including for key variables that differentiate reading profiles were missing. We also validated the reading profiles with behavioral data that were not used for classification. The methods used in this project could be leveraged to characterize or control for behavioral heterogeneity in reading studies, and can be applied in studies of other complex behavioral disorders.

TABLE 2 | Reading profile differences were observed for complete case (ACA) and multiply imputed (MI) data for behavioral measures that were not used to classify the reading profiles.

Variable	Missingness approach	n	F	Eta squared	p	^Reading profile differences
PIQ	ACA	488	36.46	0.18	2.95E-21	CTL > PR, PC, PD; PD, PC > PR
	MI	709	43.92	0.16	4.99E-26	CTL > PR, PC, PD; PD, PC > PR
Spelling	ACA	287	70.11	0.43	6.26E-34	CTL > PR, PC, PD; PC > PR
	MI	487	103.67	0.39	8.09E-52	CTL > PR, PC, PD; PC > PR (PD > PR)
Elision	ACA	415	75.89	0.36	4.45E-39	CTL > PR, PC, PD; PC > PD, PR
	MI	556	99.24	0.35	2.20E-51	CTL > PR, PC, PD; PC > PD, PR
Digit span	ACA	264	18.16	0.17	9.94E-11	CTL > PR, PC, PD; PD, PC > PR
	MI	271	18.80	0.17	4.30E-11	CTL > PR, PC, PD; PD, PC > PR

^Tukey HSD-corrected post hoc comparisons (p < 0.05). While there were age differences between profiles [all cases: $F_{(3,924)} = 14.86$, Eta squared = 0.05, p = 1.87E-09; Poor Decoders were 0.53 years younger than Poor Comprehenders, and 0.94 years younger than Controls], the test score differences here remained significant and the results did not change when covarying for the effects of age. The pairwise PC > PR difference in spelling scores was only significant in the MI results after controlling for age differences, shown in parentheses. PD, poor decoders; PC, poor comprehenders; PR, poor readers; CTL, controls.

Simulation Results

The simulation results provided critical information about the degree to which imputed data could accurately characterize variance compared to the real data, and allowed for comparisons of reading profile classification accuracy between imputation methods. Predictive Mean Matching was the most accurate imputation method based on variance approximation of the original dataset across percent levels of missingness, which indicates that it should provide relatively limited false positive results in univariate statistic tests. Nonetheless, the data imputed using missForest resulted in the highest reading profile classification accuracy across levels of missingness. The missForest imputed data also resulted in less biased classifications based on higher sensitivity and higher specificity compared to the other tested missing data methods. In the context of multivariate analyses such as Random Forest classification, missForest appears to have advantages over the other imputation methods even though the variability of imputed values was reduced, which could increase false positive rates for univariate tests. Thus, Predictive Mean Matching may be a preferable multiple imputation approach for univariate analyses, while missForest appeared to better capture the relations between variables and was preferable for multivariate classification analyses.

Perhaps not surprisingly, reading profile misclassification increased when a missing variable was critical for differentiating one reading profile from the others. For example, Generally Poor Reader misclassification was more likely when Verbal IQ was missing, while Poor Comprehender and Poor Decoder misclassifications were more likely when pseudoword ID was missing. However, there were only ~5% more classification errors with 30% simulated missingness for these critical variables. These results suggest that missForest can be applied to a real dataset to optimize reading profile classification accuracy.

Real Dataset Results

Again, we sought to develop methods that could help identify homogeneous cases of reading disability in a retrospective dataset, despite missing data. There is some evidence that reading disability classifications using cut-points can be inconsistent depending on the types of measures, in part because of inconsistent measurement reliability (Miciak et al., 2015). Measurement reliability and standardization could affect the validity of imputation and classification. The measures used for classification in the current study are widely used in the reading disability community and have demonstrated good concurrent and predictive validity (Fuchs et al., 2004; Schatschneider et al., 2004; Fuchs and Young, 2006).

We used a multivariate approach to identify patterns of performance across component reading skills, which was guided by a theoretical framework (Bishop and Snowling, 2004), clinical expertise, and empirical evidence for unique reading profiles. We used measures that are commonly administered in reading disability studies (e.g., Woodcock-Johnson Word Attack) to leverage as much of the multi-site dataset as possible. One drawback to using only real word and pseudoword accuracy measures is the absence of reading efficiency, orthography, and expressive or receptive syntax measures to help guide the classifications. These measures could have enhanced multiple imputation and reading profile classifications.

While our results support the premise for characterizing three different reading disability groups, the profile plots suggest that combinations of continuous effects within cognitive domains produces these different profiles. For example, the profile plots of the Generally Poor Readers and Poor Comprehenders in **Figures 1** and **5** are largely parallel. It is the magnitude of impairments that appears to differentiate these profiles. In addition, the pattern of associations between Verbal IQ and reading measures was similarly consistent for the Generally Poor Readers and Poor Comprehenders (Supplementary Figure S1). These results could indicate that Poor Comprehenders and Generally Poor Readers have a common latent impairment that was not assessed, such as poor organization or planning skills, which affects multiple functions and has a more exaggerated effect for challenging tasks that require support from multiple neural systems (i.e., Passage Comprehension). This idea is consistent with evidence that children exhibiting comprehension-specific reading disability

also exhibit relatively low executive function performance (Locascio et al., 2010), which may explain this unexpectedly low oral language performance (Cutting et al., 2009). Low oral language performance (e.g., receptive syntax understanding) would also be expected in the Generally Poor Readers (Oral and Written Language Learning Disability; Berninger, 2008; Berninger and May, 2011). We predict that the magnitude of executive function impairment creates differences across multiple behavioral measures between Generally Poor Readers and Poor Comprehenders. The Generally Poor Readers were not classified into the Poor Comprehenders profile, in part, because they also exhibited behavioral similarity to the Poor Decoders, as we discuss below.

Independent Validation

We leveraged data that were not used for imputation and reading profile classification to demonstrate that the reading profiles could be differentiated with Performance IQ, Elision, Spelling, and Digit Span measures. The same pattern of behavioral differences was observed for cases with complete data and for cases who were classified with imputed data. The only difference was the larger magnitude of *F* statistics for the imputed datasets, although not *eta squared* effect sizes, suggesting that the reading profile differences were stable.

The validation data and results were limited to the extent that there were too many cases with missing data for reliable imputation results. Nonetheless, these variables and results provided evidence of differences between the reading profiles that included: (1) Generally Poor Readers who had consistently low scores compared to Controls and lower Performance IQ and Digit Span compared to Poor Decoders and Poor Comprehenders; (2) Poor Decoders who had lower Elision than the Poor Comprehenders, but did not differ with Poor Comprehenders in Performance IQ and Digit Span performance; and (3) Poor Comprehenders who performed consistently lower than Controls, but as noted above, had better performance for the Spelling and Elision measures compared to the Generally Poor Readers and better Elision than the Poor Decoders.

In contrast to **Figures 1** and **5** in which Generally Poor Readers and Poor Comprehenders had similar patterns and differed in magnitude of impairment, **Figure 6** shows similar patterns for Generally Poor Readers and Poor Decoders that differed in the magnitude of impairment. In addition, the pattern of associations between reading measures was also consistent for the Generally Poor Readers and Poor Decoders (Supplementary Figure S1). These results suggest that impairments associated with Poor Decoders also contribute to the reading problems that Generally Poor Readers experience. Thus, Generally Poor Readers could experience executive function problems that affect Poor Comprehenders, as well as phonological processing problems that affect Poor Decoders. This interpretation is broadly consistent with the premise that Generally Poor Readers experience a combination of decoding and comprehension problems (Gough and Tunmer, 1986) or non-phonological language problems (Bishop and Snowling, 2004).

Limitations

The missingness and classification methods used in this study were effective despite an unbalanced number of participants in each reading profile. Users of these methods should be cautious in their implementation with relatively small datasets when it is possible that the cases belonging to the under-sampled profile could be misclassified because of a relatively limited contribution of these cases to the imputation model. We also acknowledge that the three reading disability profiles examined in this study may not be a comprehensive characterization of different reading profiles. Again, we did not have access to reading efficiency or fluency measures that might have further differentiated the profiles. We were able to independently validate the three reading profiles, but also note again that this study was limited by the more extensive missingness for the variables used to validate our results.

CONCLUSION

Researchers are increasingly motivated to share data in open access databases. These data sharing and scientific transparency initiatives can advance the pace of discoveries and enhance scientific integrity, but retrospective analysis of data collected without coordination can be challenging for many reasons. Specifically, children with different reading disability profiles can be included in multi-site datasets, particularly if different ascertainment or sampling criteria were used between research sites. This sampling heterogeneity can introduce variance and challenge our ability to make inferences from multi-site results. Ideally, investigators could control for this heterogeneity, but this is difficult when there is missing data that would help differentiate reading profiles. The current study demonstrates that consistent patterns of reading disability can be identified in a multi-site retrospective dataset, even when there is missing data and different numbers of children within each reading profile.

We used simulations and real data to show that reading profiles can be classified with a high degree of accuracy when missing data are imputed using the missForest method. The sample size of the real dataset was almost doubled by adding participants with missing data to the participants with complete data. This provides a clear example of how the missForest imputation method could increase statistical power and reduce the false negative rate. Including cases with missing data also potentially increases the quality of inference because of more representative sampling, as demonstrated in **Figure 5**.

An expanded set of behavioral measures that includes executive function, visual perception, and cognitive processing speed, as well as developmental and family histories, could provide additional validation for the reading profiles described here. We also emphasize that the methods used in the current study are not designed to classify a single participant for clinical reasons or for case studies. These methods should be used to enhance the validity and inference from group level statistics. An important and broader implication of the

findings is that the methods demonstrated here could be applied to the study of other complex developmental and age-related disorders where there is evidence for the expression of unique profiles or phenotypes. Thus, we anticipate that the combined multiple imputation and classification approach described here can be applied to open access datasets that are composed of heterogeneous cases.

AUTHOR CONTRIBUTIONS

ME conceptualized the study, performed a subset of the data analyses, and wrote the manuscript. KV contributed to the experimental and data analytic framework, performed data analyses, and contributed to manuscript preparation. MG contributed to the statistical methods and contributed to manuscript preparation. The Dyslexia Data Consortium is a group of investigators who contributed data for the research

described in the manuscript. Please see www.dyslexiadata.org for more information about and contributors to the Dyslexia Data Consortium.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00644/full#supplementary-material

REFERENCES

Bailey, S., Hoeft, F., Aboud, K., and Cutting, L. (2016). Anomalous gray matter patterns in specific reading comprehension deficit are independent of dyslexia. *Ann. Dyslexia* 66, 256–274. doi: 10.1007/s11881-015-0114-y

Berninger, V. W. (2008). "Defining and differentiating dyslexia, dysgraphia and language learning disability within a working memory model," in *Language Impairment and Reading Disability-Interactions Among Brain, Behavior, and Experience,* eds E. Silliman, and M. Mody (New York, NY: Guilford), 103–134.

Berninger, V. W., and May, M. O. (2011). Evidence-based diagnosis and treatment for specific learning disabilities involving impairments in written and/or oral language. *J. Learn. Disabil.* 44, 167–183. doi: 10.1177/0022219410391189

Bishop, D. V., and Snowling, M. J. (2004). Developmental dyslexia and specific language impairment: Same or different? *Psychol. Bull.* 130, 858–886. doi: 10.1037/0033-2909.130.6.858

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cain, K., Oakhill, J., and Bryant, P. (2004). Children's reading comprehension ability: concurrent prediction by working memory, verbal ability, and component skills. *J. Educ. Psychol.* 96, 31–42. doi: 10.1037/0022-0663.96.1.31

Catts, H. W., Compton, D., Tomblin, J. B., and Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *J. Educ. Psychol.* 104, 166–181. doi: 10.1037/a0025323

Cutting, L. E., Materek, A., Cole, C. A., Levine, T. M., and Mahone, E. M. (2009). Effects of fluency, oral language, and executive function on reading comprehension performance. *Ann. Dyslexia* 59, 34–54. doi: 10.1007/s11881-009-0022-0

Eckert, M. A., Berninger, V. W., Vaden, K. I., Gebregziabher, M., and Tsu, L. (2016). Gray matter features of reading disability: a combined meta-analytic and direct analysis approach. *eNeuro* 3:ENEURO.0103-15.2015. doi: 10.1523/ENEURO.0103-15.2015

Eckert, M. A., Vaden, K. I., Maxwell, A. B., Cute, S. L., Gebregziabher, M., and Berninger, V. W. (2017). Common brain structure findings across children with varied reading disability profiles. *Sci. Rep.* 7:6009. doi: 10.1038/s41598-017-05691-5

Fletcher, J. M. (2009). Dyslexia: the evolution of a scientific concept. *J. Int. Neuropsychol. Soc.* 15, 501–508. doi: 10.1017/S1355617709090900

Fuchs, D., and Young, C. L. (2006). On the irrelevance of intelligence in predicting responsiveness to reading instruction. *Except. Child.* 73, 8–30. doi: 10.1177/001440290607300101

Fuchs, L. S., Fuchs, D., and Compton, D. L. (2004). Monitoring early reading development in first grade: word identification fluency versus nonsense word fluency. *Except. Child.* 71, 7–21. doi: 10.1177/001440290407100101

Gough, P. B., and Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial Spec. Educ.* 7, 6–10. doi: 10.1177/074193258600700104

Jednoróg, K., Gawron, N., Marchewka, A., Heim, S., and Grabowska, A. (2014). Cognitive subtypes of dyslexia are characterized by distinct patterns of grey matter volume. *Brain Struct. Funct.* 219, 1697–1707. doi: 10.1007/s00429-013-0595-6

King, B., Wood, C., and Faulkner, D. (2007). Sensitivity to auditory and visual stimuli during early reading development. *J. Res. Read.* 30, 443–453. doi: 10.1111/j.1467-9817.2007.00346.x

Korkman, M., and Häkkinen-Rihu, P. (1994). A new classification of developmental language disorders (DLD). *Brain Lang.* 47, 96–116. doi: 10.1006/brln.1994.1044

Kornilov, S. A., and Grigorenko, E. L. (2017). What reading disability? Evidence for multiple latent profiles of struggling readers in a large Russian sibpair sample with at least one sibling at risk for reading difficulties. *J. Learn. Disabil.* doi: 10.1177/0022219417718833 [Epub ahead of print].

Kuppen, S. E., and Goswami, U. (2016). Developmental trajectories for children with dyslexia and low IQ poor readers. *Dev. Psychol.* 52, 717–734. doi: 10.1037/a0040207

Lauterbach, A. A., Park, Y., and Lombardino, L. J. (2017). The roles of cognitive and language abilities in predicting decoding and reading comprehension: comparisons of dyslexia and specific language impairment. *Ann. Dyslexia* 67, 201–218. doi: 10.1007/s11881-016-0139-x

Leonard, C., Eckert, M., Given, B., Verninger, B., and Eden, G. (2006). Individual differences in anatomy predict reading and oral language impairments in children. *Brain* 129, 3329–3342. doi: 10.1093/brain/awl262

Leonard, C. M., and Eckert, M. A. (2008). Asymmetry and dyslexia. *Dev. Neuropsychol.* 33, 663–681. doi: 10.1080/87565640802418597

Little, R. J. A. (1988). Missing data adjustments in large surveys. *J. Bus. Econ. Stat.* 6, 287–301.

Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd Edn. Hoboken, NJ: Wiley. doi: 10.1002/9781119013563

Locascio, G., Mahone, E. M., Eason, S. H., and Cutting, L. E. (2010). Executive dysfunction among children with reading comprehension deficits. *J. Learn. Disabil.* 43, 441–454. doi: 10.1177/0022219409355476

Miciak, J., Taylor, W. P., Denton, C. A., and Fletcher, J. M. (2015). The effect of achievement test selection on identification of learning disabilities within a patterns of strengths and weaknesses framework. *Sch. Psychol. Q.* 30, 321–334. doi: 10.1037/spq0000091

Naples, A. J., Chang, J. T., Katz, L., and Grigorenko, E. L. (2009). Same or different? Insights into the etiology of phonological awareness and rapid naming. *Biol. Psychol.* 80, 226–239. doi: 10.1016/j.biopsycho.2008.10.002

Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop: bespoke creation of synthetic data in R. *J. Stat. Softw.* 74, 1–26. doi: 10.18637/jss.v074.i11

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York, NY: Wiley & Sons. doi: 10.1002/9780470316696

Samuelsson, S., Olson, R., Wadsworth, S., Corley, R., DeFries, J. C., Willcutt, E., et al. (2007). Genetic and environmental influences on prereading skills and early reading and spelling development in the United States, Australia, and Scandinavia. *Read. Writ.* 20, 51–75. doi: 10.1007/s11145-006-9018-x

Schafer, J. L., and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychol. Methods* 7, 147–177. doi: 10.1037/1082-989X.7.2.147

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., and Foorman, B. R. (2004). Kindergarten prediction of reading skills: a longitudinal comparative analysis. *J. Educ. Psychol.* 96:265. doi: 10.1037/0022-0663.96.2.265

Stanovich, K. E. (1988). Explaining the differences between the dyslexic and the garden-variety poor reader: the phonological-core variable-difference model. *J. Learn. Disabil.* 21, 590–604. doi: 10.1177/002221948802101003

Stekhoven, D. J., and Bühlmann, P. (2011). MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi: 10.1093/bioinformatics/btr597

Talcott, J. B., Witton, C., and Stein, J. F., (2013). Probing the neurocognitive trajectories of children's reading skills. *Neuropsychologia* 51, 472–481. doi: 10.1016/j.neuropsychologia.2012.11.016

Tamboer, P., Vorst, H. C., and Oort, F. J. (2016). Five describing factors of dyslexia. *J. Learn. Disabil.* 49, 466–483. doi: 10.1177/0022219414558123

Torppa, M., Tolvanen, A., Poikkeus, A. M., Eklund, K., Lerkkanen, M. K., Leskinen, E., et al. (2007). Reading development subtypes and their early characteristics. *Ann. Dyslexia* 57, 3–32. doi: 10.1007/s11881-007-0003-0

Vaden, K. I., Gebregziabher, M., Kuchinsky, S. E., and Eckert, M. A. (2012). Multiple imputation of missing fMRI data in whole brain analysis. *Neuroimage* 60, 1843–1855. doi: 10.1016/j.neuroimage.2012.01.123

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* 45, 1–67. doi: 10.18637/jss.v045.i03

Wagner, R. K., Torgesen, J. K., and Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing.* Austin, TX: Pro-Ed Inc.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence.* New York, NY: The Psychological Corporation.

Wechsler, D. (2004). *The Wechsler Intelligence Scale for Children-Revised*, 4th Edn. New York, NY: Pearson Assessment.

Wolf, M., and Denckla, M. B. (2005). *Rapid Automatized Naming and Rapid Alternating Stimulus Tests (RAN/RAS).* Austin, TX: Pro-Ed Inc.

Wolff, U. (2010). Subgrouping of readers based on performance measures: a latent profile analysis. *Read. Writ.* 23, 209–238. doi: 10.1007/s11145-008-9160-8

Woodcock, R. (1987). *Woodcock Reading Mastery Test: Revised.* Circle Pines, MN: American Guidance Service.

Woodcock, R. W., Mather, N., McGrew, K. S., and Shrank, F. A. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities.* Rolling Meadows, IL: Riverside Publishing.