As the Twig is Bent, so is the Tree Inclined: Missing Data in Generalized Linear Mixed-Model (GLMM) Trees

Weerman, Nino

As the Twig is Bent, so is the Tree Inclined:

Missing Data in Generalized Linear Mixed-Model (GLMM) Trees

Nino Arjan Weerman

Master's Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioral Sciences, Leiden University

Date: 24 March 2021

Student number: 2582686

Supervisor: Dr. Marjolein Fokkema and Prof. Dr. Elise Dusseldorp

Abstract

Objective: The Generalized linear mixed-model (GLMM) tree is a decision-tree method which allows for subgroup detection in a wide range of multilevel datasets. This thesis provides a first evaluation of how missing data can be handled in GLMM trees, by assessing the performance of listwise deletion (LD), mean or mode single imputation (SI), multiple imputation (MI) and missingness incorporated in attributes (MIA), in terms of predictive accuracy and tree size accuracy.

Method: Different missingness mechanisms, proportions of missing cases and missing data were artificially introduced into data retrieved from the Early Childhood Longitudinal Study Kindergarten class of 1998-1999.

Results: As expected, MI yielded the highest performance overall, closely followed by MIA, which exhibited an approximately similar performance. SI performed somewhat worse than MIA and MI, whereas LD showed a substantially inferior performance. Individually, MI and MIA performed very similar for lower amounts of missing data, MI slightly outperformed MIA for higher amounts of missing data, missing completely at random (MCAR) and missing at random (MAR) data and MIA slightly outperformed MI for MNAR (missing not at random) data. When comparing the size of fitted GLMM trees with those fitted on the complete data, MI tended to overfit and yield ensembles of more complex trees, whereas LD, SI and MIA tended to underfit and yield simpler decision trees. Furthermore, the performance of LD was lowest across all conditions and deteriorated even further as the number of cases with missing increased.

Conclusion: For handling missing data in GLMM trees, MI is recommended predominantly for prediction purposes, but lacks interpretability. Alternatively, MIA is recommended for interpretability and when a smaller tree size is preferred. Conversely, using either LD or SI is discouraged, even though SI is preferred over LD.

Table of contents

Introduction

Decision-tree (or recursive partitioning) methods are nonparametric supervised machine-learning techniques, applicable for both regression and classification problems (Witten, Hastie & Tibshirani, 2013). Decision-tree methods recursively partition the observations in a dataset into subgroups (nodes) based on the values of the covariates, aiming to maximize the homogeneity within, and the heterogeneity between subgroups (nodes) in terms of an outcome of interest. The resulting model provides a series of decision rules for determining the predicted value of an outcome of interest, which can be graphically depicted as a decision tree. As these decision trees are relatively easy to interpret and apply by human decision makers (Breiman et al., 1984; Quinlan, 1986), decision-tree methods are commonly used in many research fields, such as the social and behavioural sciences (Kopf, Augustin & Strobl 2010; Strobl, Malley, & Tutz, 2009). Since decision trees can automatically detect (higher-order) interactions, they can be used to identify higher-order interactions. For example, which treatment is more effective for which subgroup of patients, also known as treatment-subgroup interactions (e.g., Doove, Dusseldorp, Van Deun & Van Mechelen, 2014). A further advantage of decision-tree methods is their non-parametric character. As such, decision-tree methods are able to handle non-normality, extreme outliers, multicollinearity, non-linear relationships and a large number of potential covariates, which may exceed the number of observations.

A more recent branch of decision-tree methods is focused on assessing multilevel data, data structures particularly common in social and behavioural research. In multilevel datasets, individual observations can be nested in higher-level units: E.g., individual patients nested within treatment centers, or individual assessments are nested within the same subject in longitudinal studies, introducing a dependence between the lower- and higher-level units. Fitting decision-tree methods on such datasets while ignoring this dependence can result in identifying spurious subgroups and biased variable selection (e.g., Sela & Simonoff, 2012; Martin, 2015). To address this issue, Fokkema et al. (2018) introduced generalized linear mixed-effects model (GLMM) trees which allow for analysing multilevel data, following the estimation procedure similar to that proposed by Sela and Simonoff (2012) and Hajjem, Bellavance and Larocuqe (2011). The GLMM tree algorithm builds on generalized linear model trees, an extension of the unbiased model-based recursive partitioning algorithm by Zeileis et al. (2008), by accounting for the correlated residuals in multilevel data (for additional technical details on the algorithm, consult Fokkema et al., 2018). The GLMM tree is a promising method for assessing multilevel data, as it can accurately detect treatment-

subgroup interactions for both regression and classification problems and yields a predictive accuracy that is competitive with traditional GLMMs, and sometimes even with random forests, while requiring evaluation of less variables in order to make predictions (Fokkema et al., 2018; Fokkema, Edbrooke-Childs & Wolpert, 2020).

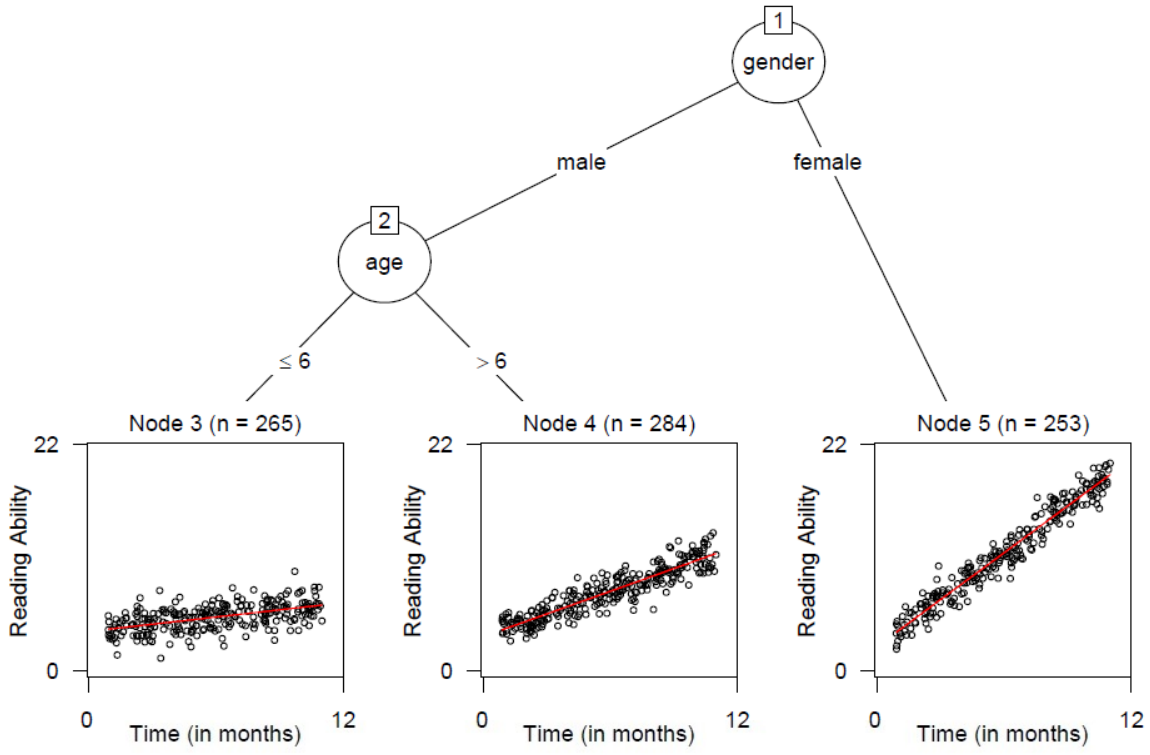

Figure 1. Example GLMM tree for children's development of reading ability as a result of a year-long reading training on the generated artificial motivating dataset (N = 802). The x-axes in the terminal nodes represent the time in months and the y-axes represent the reading ability. Age and gender were used as potential partitioning variables.

Figure 1 presents an example of a fitted GLMM tree. For this example, an artificial longitudinal motivating dataset was generated, containing repeatedly evaluated reading ability scores of N = 802 children who received a reading training. The GLMM tree was fitted using reading ability as a response variable, months of training received as a predictor variable and age and gender as potential partitioning variables. In addition, a random intercept was modelled with respect to children to account for the dependence between children's measures of reading ability. The resultant GLMM tree in Figure 1 partitioned the children into groups of observations who differed in their development of reading ability. A first split was made using the gender variable and a second split was made for the male subgroup using the age variable. The identified subgroups are defined by the terminal nodes (rectangles) which

contain subgroup-specific (generalized) linear models (red lines), consisting of the sub-group specific effect of months of treatment received (*x*-axis) on the reading ability (*y*-axis). These reveal a slight increase of reading ability for boys who are 6 or younger (node 3), a moderate increase in reading ability for boys who are older than 6 (node 4) and a steep increase in reading ability of boys who are older than 6 (node 5) as a result of a reading training.

As with all statistical techniques, one important and commonly encountered issue when fitting a GLMM tree is missing data (Little & Rubin, 2019). Missing data is an important issue for decision-tree methods since incomplete data may not only negatively impact interpretations based on a decision tree created from the data but may also negatively affect the prediction accuracy of a particular fitted decision tree, which could result in inaccurate predictions and misleading interpretations (e.g., Ding & Simonoff, 2010; He, 2006; Twala, 2005). Moreover, inappropriate handling of missing data for decision-tree methods could cause the introduction of bias, which may lead to inaccurate conclusions being drawn and weakens the generalizability of results (Enders, 2010; Little & Rubin, 2019; Schafer, 1997).

To handle missing data for decision-tree methods, two commonly used techniques are listwise deletion (LD) and mean or mode single imputation (SI). However, these ad-hoc methods are known for biasing inference, a loss of power, underestimating variability (Schafer & Graham, 2002) and generally show relatively lower predictive accuracy compared to other missing data techniques for decision-tree methods (e.g., Ding & Simonoff, 2010; He, 2006; Twala, 2005). On the other hand, multiple imputation (MI) is regarded as a state-of-the art technique (Schafer & Graham, 2002) and often recommended for handling missing data in decision-trees as it generally yields relatively high predictive accuracy, superior to SI and LD (e.g., He, 2006; Twala, 2005). However, the superior accuracy of MI comes at the expense of losing the interpretability and ease-of-use of a single decision tree. Using MI results in an ensemble of decision trees which are difficult for humans to visually grasp (Rokach, 2010; Witten, Hastie & Tibshirani, 2013). Alternatively, missingness incorporated in attributes (MIA; Twala, Jones & Hand, 2008) is a promising missing data technique for decision-tree methods which retains the benefits of a single decision tree, while yielding predictive accuracy similar to MI. Even though the performance of MIA for handling missing data in decision-tree methods has been compared to MI, it has not been compared to LD and SI yet. Furthermore, it should be noted that surrogate variable splitting (SVS; Breiman, Friedman, Olshen & Stone, 1984) is a popular approach for handling missing data in decision-tree methods. However, SVS generally shows a lower predictive accuracy as compared to other

missing data techniques, (e.g., Feelders 1999; Provost & Saar-Tsechansky, 2007), can cause a variable selection bias (Kim & Loh, 1999) and has an increased computation time (e.g., Twala, 2005). Therefore, it was decided to leave SVS outside the scope of this thesis.

While there are some studies available on handling missing data in decision-tree methods, thus far no studies have investigated how missing data should best be handled for the GLMM tree method, which currently employs LD as a default. In addition, previous studies on missing data in decision-tree methods predominantly focused on evaluating predictive accuracy. However, since the main strength of decision-tree methods lies in their interpretability (Breiman et al., 1984; Quinlan, 1986), it is also important to assess the degree to which missing data techniques can accurately resemble the tree size of GLMM trees fitted on complete data. Therefore, this thesis will evaluate the performance of LD, MIA, MI and SI for handling missing data in GLMM trees, in terms of predictive accuracy and tree size accuracy.

This thesis is structured as follows: In the remainder of this Introduction, the underlying mechanisms yielding missing values and techniques for dealing with missing data (MIA, MI, SI and LD) will be discussed in detail. In the Method and Results, the performance of MIA, MI, SI, and LD for handling missing data in GLMM trees will be evaluated through experiments with an existing longitudinal dataset of the Early Childhood Longitudinal Study Kindergarten Class of 1998-1999 (ECLS-K; National Center for Education Statistics, 2016). Lastly, this thesis will conclude with a Discussion, in which the experimental results are summarized and limitations and directions for future research are discussed.

Missing data mechanisms

When handling missing data, an important task is to investigate the underlying mechanisms which could have caused the missingness as this underlying mechanism greatly impacts the performance of missing data techniques (Enders, 2010; Little & Rubin, 2019; Schafer, 1997). Moreover, the underlying missing data mechanism has even been shown to have a bigger impact on the modeling results than the proportion of missing values (e.g., Little & Rubin, 2019; Molenberghs et al., 2014). Consequently, it is an essential issue to address when handling missing data. The underlying mechanism of missing data can be classified into three classes with the common and widely accepted taxonomy of Rubin (1976) and Rubin and Little (2019, p. 10): missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR).

Under a MCAR scheme, missingness is assumed to be completely random. Each value has an equal chance of being missing and the missingness depends neither on the observed nor unobserved values. As MCAR assumes complete random missingness, it is the strongest assumption that can be made about the underlying mechanism of missingness. MCAR data is often deemed as ignorable since MCAR data does not introduce bias into the results, regardless of which missing data technique. However, the effect of missingness can be strongly exacerbated through the use of LD, which will be discussed in the next section. An example of MCAR is a respondent accidentally skipping an item of a questionnaire.

Under a MAR scheme, missingness is assumed to depend on observed but not on the unobserved values in the model. Therefore, MAR is a weaker assumption compared to MCAR and can be seen as a conditional form of MCAR. Like MCAR data, MAR data can be seen as ignorable but will only produce unbiased results for missing data techniques that are appropriate for handling MAR data. A specific example of MAR would be when women are less likely to report their weight compared to men.

Under a NMAR scheme (also known as informatively missing), missingness is assumed to depend on unobserved data (including the missing value itself) and cannot be classified as MCAR nor MAR. NMAR data is deemed non-ignorable and challenging to deal with. While many missing data techniques exist for handling MCAR and MAR (i.e., ignorable) data, very few missing data techniques currently exist that can handle NMAR data (Little & Rubin, 2019). As such, researchers may be forced to use techniques which are less appropriate for handling NMAR data. An example of NMAR is when the most depressed people are more likely to drop out of a study on depression.

Missing data techniques

Listwise deletion

Listwise deletion (LD) entails discarding all cases containing missing values and using only complete cases for subsequent analysis. Due to its simplicity and ease of use, it is a common approach for handling missing data in decision-tree methods, even though it is an ad-hoc method with little theoretical justification and yields relatively less predictive accuracy compared to other missing data techniques (e.g., He, 2006; Twala, 2005).

Mean or mode single imputation

With mean or mode single imputation (SI), missing values are imputed with the variable mean for continuous variables and the variable mode for categorical variables.

Similar to LD, SI is an ad-hoc method which is relatively simple and easy to apply and commonly used for handling missing data in decision-tree methods. Compared to LD, SI generally yields higher predictive accuracy (e.g., He, 2006; Twala, 2005), which can be explained by SI preserving the full sample size through imputation, thereby preventing a loss of power. Compared to other missing data techniques for decision-tree methods, SI generally yields lower predictive accuracy, because it does not account for the uncertainty of imputation and assumes data to be MCAR (Little & Rubin, 2019).

Multiple imputation

Multiple imputation (MI) can be described as a three-phase process (Little & Rubin, 2019). In the first phase, the missing values are imputed by creating *m* sets of possible values for the missing values. This is done by utilizing an underlying distribution or model based on the complete observations to predict the missing values and adding a random error to each model-based prediction to reflect the uncertainty of the missing data. In the second phase, a statistical model (e.g., linear regression) is fitted on each of the *m* complete datasets containing unique estimates of the missing values for each dataset. In the third and final phase, the resultant *m* separate parameter estimates (e.g., coefficients and standard errors) are combined or pooled into a single final result using Rubin's Rule (1987). Since this process allows for incorporating imputation uncertainty, MI can represent missing observations more accurately compared to single imputation methods such as SI (Enders, 2010; Little & Rubin 2019). In addition, this process allows MI to generally yield valid results when data is MCAR or MAR.

MI is considered a state-of-the art missing data technique (Schafer & Graham, 2002), but decreases interpretability of decision trees. When using MI, multiple decision trees are fitted on *m* datasets, resulting in an ensemble of decision trees which will likely contain *m* different tree structures due to the inherent instability of decision-tree methods (Strobl, Malley & Tutz, 2009). As decision-tree methods have a non-parametric nature, it is not possible to combine these tree ensembles into a single tree. However, a single decision tree may be preferred because they are relatively easy to interpret and apply by human decision makers (Rokach, 2010; Witten, Hastie & Tibshirani, 2013).

Missingness incorporated in attributes

Missingness incorporated in attributes (MIA) is a missing data technique which treats "missing" as a separate variable value on its own (Twala, Jones & Hand, 2008). If a missing

value occurs within a categorical variable, that missing value is treated as a separate category . For continuous variables, variables with missing values are duplicated and missingness is filled in with values outside of the observed data range: two variables are created for each variable with missing values: one obtains a value below the minimum for each missing value, the other has a value above the maximum for each missing value.

Compared to MI, MIA yields similar predictive performance (e.g., Ding & Simonoff, 2010; Twala 2005) while retaining a single decision tree, which is easily interpreted and applied by human decision makers (Rokach, 2010; Witten, Hastie & Tibshirani, 2013). In addition, MIA is computationally cheaper and does not require choosing a statistical model for imputation of the data. Furthermore, MIA is not only appropriate for when missingness is MCAR or MAR but also when missingness is NMAR (e.g., Twala 2005). By treating missingness as a separate value, MIA can use the information in the missing values and might therefore be better suited to handle a NMAR scheme. Additionally, by considering missingness as an “observed” value, MIA can be effective when missingness is predictive of the outcome (e.g., Ding & Simonoff, 2010).

Method

Dataset

Data was retrieved from the Early Childhood Longitudinal Study-Kindergarten (ECLS-K; National Center for Education Statistics, 2016). The ECLS-K was aimed at measuring child development, early school experiences and school readiness from kindergarten through eighth grade of a cohort of children who were in kindergarten in 1998. This thesis focused exclusively on five time points in kindergarten, 1st, 3rd, 5th and 8th grade, resulting in the inclusion of data of 21,304 children from 1,018 different schools across the United states of America.

Following a design comparable to Stegmann et al., 2018, the trajectory of reading ability was chosen as the outcome of interest. Reading ability was measured using a 20-item direct cognitive assessment test where the difficulty of the items varied based on a prior routing test consisting of 72 items. The resultant reading scores were quantified into θ scores with $M = 0$ and $SD = 1$, reflecting the standardized latent reading ability scores (Kline, 2013).

Potential partitioning variables were child-level covariates measured at the baseline in kindergarten: gender, race, socioeconomic status, externalizing problem behaviour, internalizing problem behaviour, fine motor skills, gross motor skills, interpersonal skills, self-control and whether it was a child's first time in kindergarten. In addition, a variable measuring children's age in months in kindergarten was added as a potential partitioning variable because this was expected to yield a substantial influence on children's initial level of reading ability.

Besides the trajectory of reading ability, the trajectories of science and math ability were used as outcomes for a sensitivity analysis. This allowed a comparison whether the results for the reading outcome data would also hold for science and math outcome data, resulting in more robust inferences. As compared to reading ability, science and math ability were measured with a similar procedure. However, science ability was measured using three rather than five time points, which occurred during the spring of third, fifth and eighth grade.

There were missing values present in the reading, math and science ability outcome data due to attrition. To have complete control over simulating missingness and as the datasets were sufficiently large, only children without missing values were included. This resulted in a sample size of $N = 6{,}277$ for the reading, $N = 6{,}512$ for the math and $N = 6{,}625$ for the science ability outcome data. The summary statistics of the reading ability outcome data are presented in Table 1 and the summary statistics of the math and science ability outcome data are presented in Appendix A.

Table 1.

Descriptive statistics of the baseline Early Childhood Longitudinal Study-Kindergarten reading ability outcome data (*N* = 6277)

	M (*SD*) or %	Range
Gender		
Male	49.4%	
Female	50.6%	
Race		
White, non-Hispanic	66.0%	
Black or African American, non-Hispanic	9.7%	
Hispanic, race specified	7.5%	
Hispanic, race not specified	7.6%	
Asian	4.0%	
Native Hawaiian, other Pacific Islander	1.0%	
American Indian or Alaska native	1.8%	
More than one race, non-Hispanic	2.4%	
First time in kindergarten		
Yes	96.5%	
No	3.5%	
Socioeconomic status	0.19 (0.77)	-4.75 – 2.75
Gross motor skills	6.47 (1.78)	0.00 – 8.00
Fine motor skills	6.12 (1.94)	0.00 – 9.00
Interpersonal skills	3.06 (0.61)	1.00 – 4.00
Self-control	3.16 (0.59)	1.00 – 4.00
Internalizing problem behaviour	1.49 (0.49)	1.00 – 4.00
Externalizing problem behaviour	1.55 (0.59)	1.00 – 4.00
Age at baseline (in months)	73.75 (4.15)	64.00 – 88.00
Reading ability[1]		
Kindergarten year	-0.58 (0.43)	-2.18 – 0.92
First-grade year	0.14 (0.38)	-1.89 – 1.23
Third-grade year	0.79 (0.37)	-0.64 – 2.02
Fifth-grade year	1.18 (0.39)	-0.12 – 2.33
Eight-grade year	1.51 (0.43)	0.24 – 2.54

[1]Theta scores, reading ability was measured during the fall.

Experimental design

All subsequent simulations, application of missing data techniques and performance evaluation analyses were performed in the statistical programming environment R (version 4.0.0; R Core Team, 2020). Prior to the simulation of missingness, the data was partitioned into training and test data sets by employing cluster-level (i.e., child-level) subsampling. Subsequently, missing was simulated into the partitioning variables on cluster-level using the "ampute" function (Schouten, Lugtig and Vink, 2018) from the mice package (version 3.9.0; Van Buuren & Groothuis-Oudshoorn, 2011). To this end, the following four levels of data characteristics were varied:

1. **sample size**: $N = 100$, $N = 200$, $N = 400$ and $N = 800$, where 10 subsamples of each size were randomly drawn on cluster-level (i.e., child-level).
2. **proportion of cases with missing:** 25%, 50% and 75% of cases which could contain missingness were randomly selected.
3. **missingness mechanism**: MCAR, MAR and MNAR, where a random generator was employed for simulating MCAR data and a right-tailed logistic model was employed for simulating MAR and MNAR data.
4. **proportion of missing values**: 15%, 30%, 50%, where missing values were simulated according to a pattern with each variable having an equal probability of being missing.

After simulation of missingness, LD, MI, MIA and SI were applied to handle the missing data in the incomplete training datasets, using the following procedure:

- **LD**: all cases containing missing values were omitted.
- **SI**: all missing values were replaced with the variable mean for numerical variables and the variable mode for categorical variables, based on non-missing data only.
- **MI**: executed using the mice package (version 3.9.0; Van Buuren & Groothuis-Oudshoorn, 2011), where 10 imputations were generated using predictive mean matching.
- **MIA**: executed using a custom function created in R (see Appendix B).

GLMM trees

GLMM trees were fitted on the LD, MI, MIA and SI dataset and on the complete training datasets. For fitting GLMM trees, the "lmertree" function from the glmertree package

(version 0.2-0; Fokkema et al., 2018) was used, employing cluster-level (i.e., child-level) covariances for the parameter stability tests and a maximum node depth of five, thus limiting the number of terminal nodes to 32. To capture development of children's reading ability over time, months passed since the baseline in kindergarten was modelled as a timing metric. As the timing metric did not follow a linear trend and the GLMM tree employs model-based recursive partitioning based on a (generalized) linear mixed model (Zeileis, 2008), a square root transformation was applied to months passed since kindergarten, resulting in an approximately linear trend. In addition, a random intercept with regard to children was modelled to capture children's individual differences in reading ability at the baseline in kindergarten. As potential partitioning variables, the previously discussed variables were included. Furthermore, note that while a similar procedure was followed for fitting GLMM trees on the science and ability outcomes (i.e., sensitivity analysis) the timing metric for the change in science ability was transformed to the power ⅔ as opposed to a square root transformation for the change in reading and math ability.

Performance evaluation

The performance of LD, MI, MIA and SI was assessed by means of prediction accuracy and tree size accuracy, using GLMM trees fitted on the baseline training data as a benchmark. The performance measures were operationalized in the following way:

- **predictive accuracy:** operationalized with the relative excess error. To estimate the relative excess error, first the test mean squared error (MSE) between the predicted outcomes and observed outcomes was computed for each GLMM trees trained on the LD, MI, MIA and SI datasets. Second, the MSE was subtracted from and divided by the MSE of the corresponding benchmark subsample without missing values.
- **tree size accuracy:** operationalized with the tree size deviation, which was estimated by subtracting the tree size (i.e., number of nodes) of each GLMM tree fitted on LD, MI, MIA, and SI training datasets from the tree size of the corresponding benchmark training data. Note that for MI, the tree size deviation was computed by averaging the tree size of GLMM trees fitted onto 10 imputed datasets into a single tree size.

In addition, a mixed design ANOVA was used to examine the effects of the design on the predictive accuracy and tree size accuracy. The ANOVA included main effects of missing data technique, proportion of cases with missing and proportions of missing values as within-

subject factors and the main effect of sample size as a between-subject factor. In addition, the ANOVA included first-, second- and third-order interactions between missing data technique and each of the data characteristics. Due to the large number of effects, an $\alpha = 0.01$ level of significance was employed. To interpret the effect sizes, the $_{\text{partial}}\ \eta^2$ is provided and its magnitude interpreted according to guidelines provided by Kirk (1996). Variables were concluded to have a leading effect on the predictive accuracy or tree size accuracy if they yielded a $_{\text{partial}}\ \eta^2 > .06$ (medium effect).

Results

Below, the performance of LD, MI, MIA and SI for handling missing data in GLMM trees are described in terms of predictive accuracy and tree size accuracy under varying levels of data characteristics for the ECLS-K reading ability outcome data. The results of the sensitivity analysis for the math and science ability (Appendix C) yielded approximately similar results and are therefore not discussed any further.

Predictive accuracy

The results of the mixed design ANOVA on the excess error are depicted in Table 2. These indicated, as expected, significant and large main effects of missing data technique and all data characteristics. Comparatively, the main effect of sample size was smaller. In addition, the ANOVA showed significant large two-way interaction effects between missing data technique and each of the missing data generating parameters. In contrast, the two-way interaction effect between missing data technique and sample size was insignificant. Furthermore, no significant higher order interaction effects were found.

Table 2.

Mixed design ANOVA on the excess error for the 40 subsamples of the Early Childhood Longitudinal Study-Kindergarten reading ability outcome data

Effect	*df*	*F*	*p*	$_{\text{partial}}\ \eta^2$
Technique	3	323.55	<.001	.90
Missingness mechanism	2	260.44	<.001	.89
Cases with missing (%)	2	246.61	<.001	.87
Missing values (%)	2	198.92	<.001	.85
Sample size	3	6.30	.002	.34
Technique * missingness mechanism	6	21.96	<.001	.38
Technique * cases with missing (%)	6	31.04	<.001	.46
Technique * missing values (%)	6	6.05	<.001	.14
Technique * sample size	9	0.89	.530	.07

The main effect of missing data technique is visible in Figure 2, which depicts the averaged excess error of LD, MI, MIA and SI over the different data characteristics. Overall, MI yielded the highest predictive accuracy, closely followed by MIA whose predictive accuracy was not significantly different from MI. Thereafter, SI yielded a somewhat worse

predictive accuracy than MI and MIA, whereas LD yielded a substantially lower predictive accuracy than the other techniques.

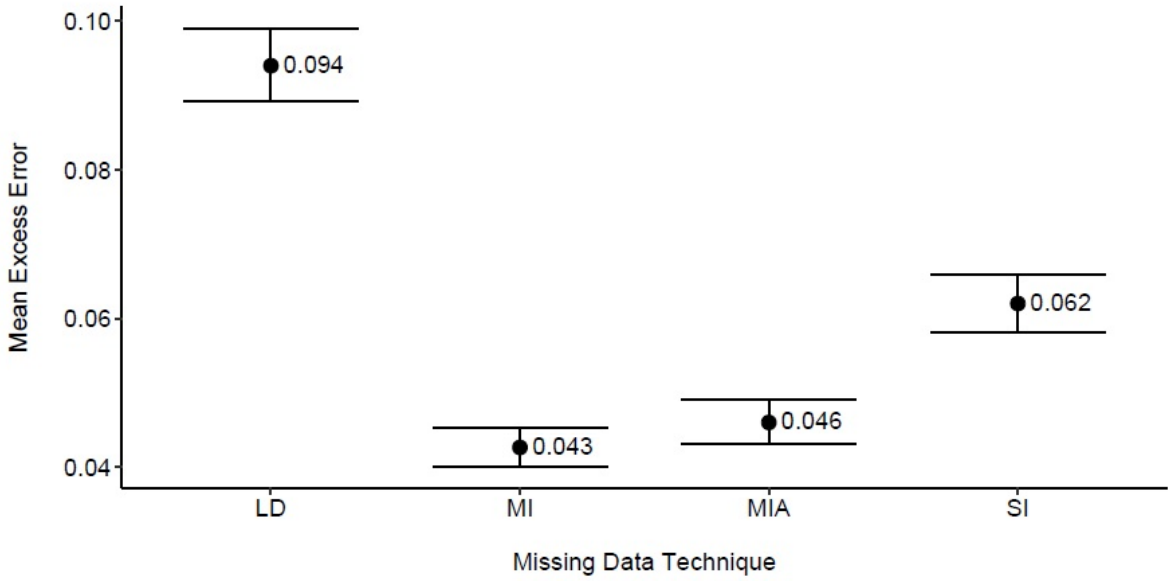

Figure 2. Mean excess error rate and 95% confidence intervals of LD, MI, MIA and SI for the Early Childhood Longitudinal Study-Kindergarten reading ability outcome data reading ability outcome data, averaged over the different missingness scenarios and subsamples of the design. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes.

As only the main effect of sample size was significant, it is depicted in Figure 3. The figure suggests that the predictive accuracy of all missing data technique increased as sample size increased. This increase in predictive accuracy flattened with increasing sample sizes.

Figure 4 visualizes the significant main effects of the missing data generating parameters and the significant two-way interaction effects with missing data technique. Across all conditions, LD again showed the lowest predictive accuracy, followed by SI, whereas MI and MIA showed the highest predictive accuracy. As expected, the main effects suggest that all missing data techniques achieved the lowest predictive accuracy for MNAR data, followed by MAR and then MCAR (panel A). Also, predictive accuracy of all techniques deteriorated for increasing proportions of cases with missing (panel B) and proportion of missing values (panel C).

Looking at the specific interactions, Figure 4A indicates that strikingly, MI showed higher predictive accuracy than MIA for MCAR and MAR data, whereas MIA showed a

higher predictive accuracy than MI for MNAR data. In addition, the predictive accuracy of LD, MI and SI was more strongly affected by data being MNAR, compared to the predictive accuracy of MIA.

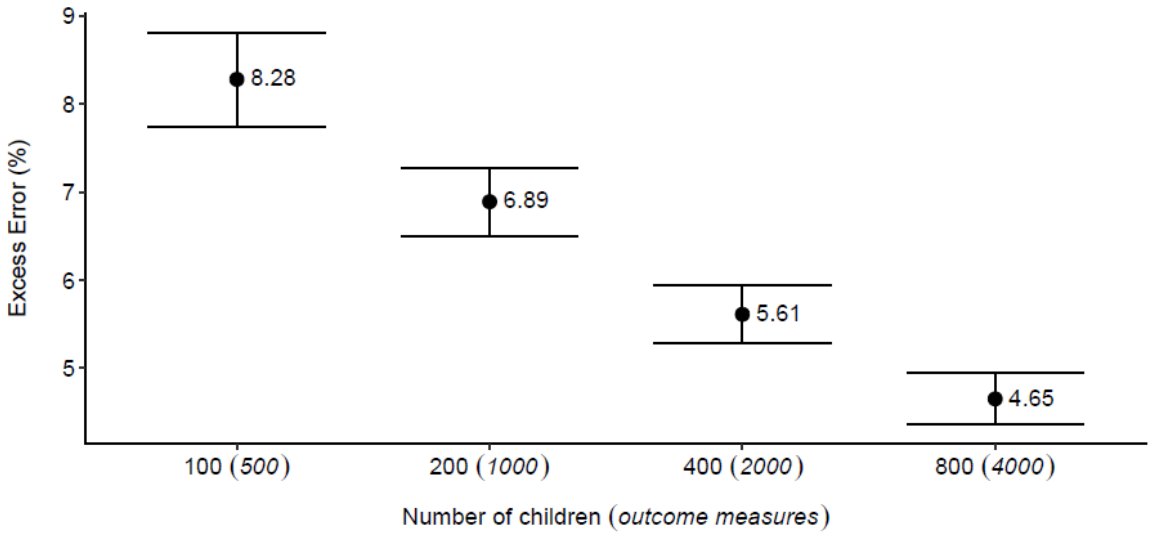

Figure 3. Mean excess error rate and 95% confidence intervals of sample size for the Early Childhood Longitudinal Study-Kindergarten reading ability outcome data, averaged over the missingness data techniques and missingness scenarios of the design.

Figure 4B and 4C indicate that MI and MIA yielded a similar predictive accuracy for lower amounts of missing data (i.e., 15% and 30% of missing values and 50% and 75% of cases with missing), whereas MI yielded a higher predictive accuracy than MIA for higher amounts of missing data (i.e., 75% of cases with missing and 50% of missing values). In addition, LD showed a more serious deterioration in predictive accuracy with increasing amounts of cases as compared MI, MIA and SI. Furthermore, for LD, the decrease in predictive accuracy with increasing proportions of missing values was steeper from 15% to 30% missing values but flatter from 30% to 50% missing values. Note that this pattern can be attributed to the simulation design used for this thesis: the effective number of observations retained through LD was likely very similar in datasets with 30% and 50% missing values.

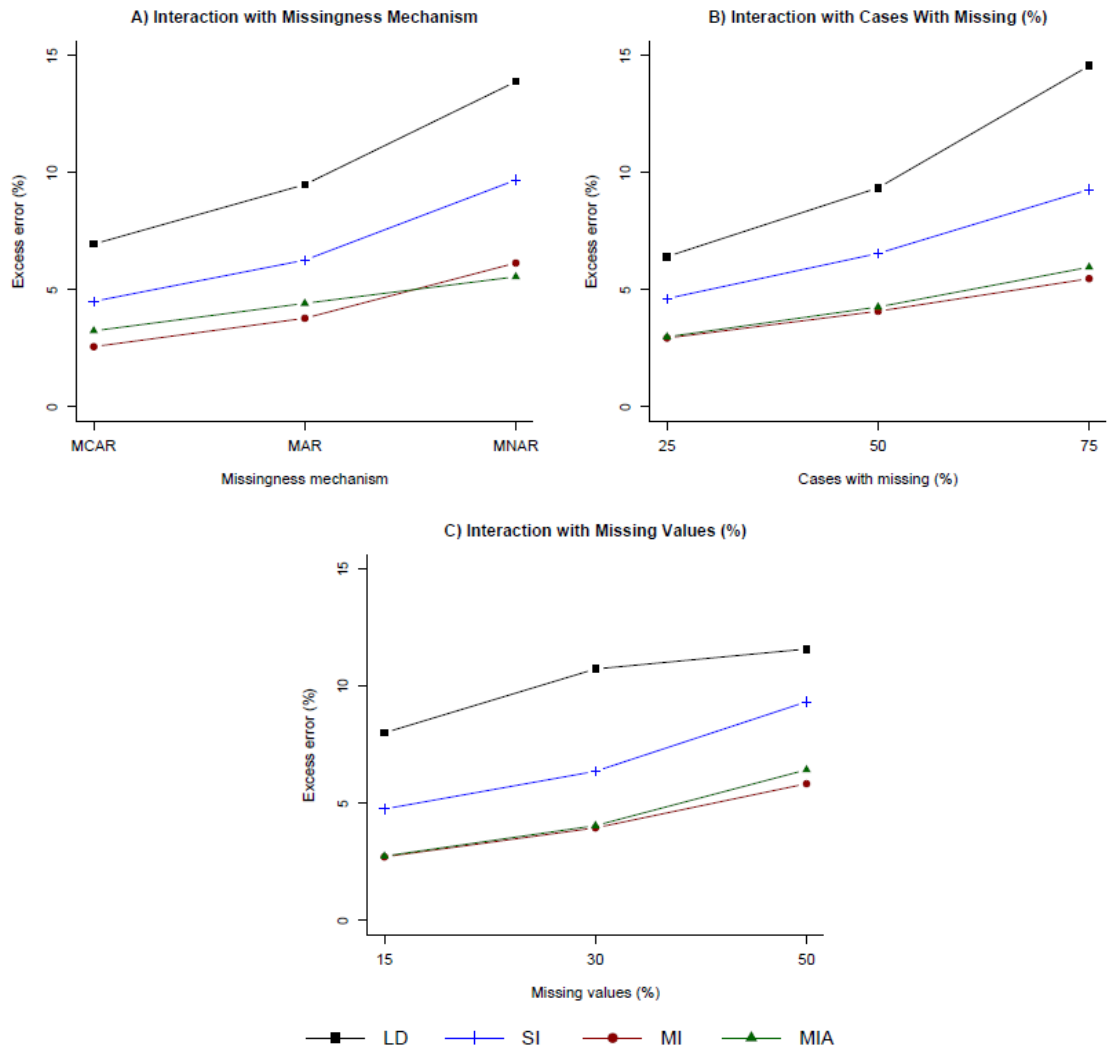

Figure 4. Interaction effects between missing data technique and A) missingness mechanism, B) proportion of cases with missing and C) proportion of missing values on the excess error for the Early Childhood Longitudinal Study-Kindergarten reading ability outcome data. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.

Tree size accuracy

Recall that for estimating the tree size accuracy of MI, the average tree size of GLMM trees fitted on 10 imputed datasets was computed. Thus, the mean tree size deviation should not be taken as a straightforward indicator of model complexity for MI, as it can be for MIA, LD and SI. The results of the 5-way mixed design ANOVA on the tree size deviation are depicted in Table 3. These indicated significant large main effects of missing data technique and each of the missing data generating parameters. In contrast, the main effect of sample size

was insignificant. In addition, the ANOVA showed significant large two-way interaction effects between missing data technique and each of the data characteristics. Comparatively, the two-way interaction effect between missing data technique and sample size was smaller. Furthermore, the ANOVA results indicated no significant higher order interaction effects.

Table 3.

Mixed design ANOVA on the tree size deviation for the 40 subsamples of the Early Childhood Longitudinal Study-Kindergarten reading ability outcome data

Effect	*df*	*F*	*p*	$_{\text{partial}}\ \eta^2$
Technique	3	399.26	<.001	.92
Cases with missing (%)	2	207.80	<.001	.85
Missing values (%)	2	118.01	<.001	.77
Missingness mechanism	2	93.41	<.001	.72
Sample size	3	2.24	.126	.14
Technique * cases with missing (%)	6	106.40	<.001	.75
Technique * missing values (%)	6	74.18	<.001	.67
Technique * missingness mechanism	6	128.29	<.001	.78
Technique * sample size	9	5.99	<.001	.33

The main effect of missing data technique is visible in Figure 5. Here, a similar pattern of performance as compared to the predictive accuracy is observed: overall, MI yielded the highest tree size accuracy, closely followed by MIA. In terms of absolute tree size deviation, MI and MIA did not differ significantly. In addition, MI tended to overfit and yield bigger trees whereas MIA tended to underfit and yield smaller trees. Compared to MI and MIA, SI showed a somewhat worse tree size accuracy, while LD showed a substantially worse tree size accuracy. Both SI and LD tended to underfit and yield smaller trees. Furthermore, note that MI achieved a relatively small 95% confidence interval, which can be attributed to the stabilization achieved by the aforementioned averaging of the tree size of MI over 10 imputations, resulting in a lower standard error and therefore smaller confidence interval.

Figure 6 visualizes the significant main effects of the data characteristics and the significant two-way interaction effects with missing data technique. Across all conditions, as in Figure 5, LD showed the lowest tree size accuracy followed by SI, whereas MI and MIA showed the highest tree size accuracy. As observed in Figure 5, MI tended to overfit and yield bigger trees whereas LD, SI and MIA tended to underfit and yield smaller trees. As expected,

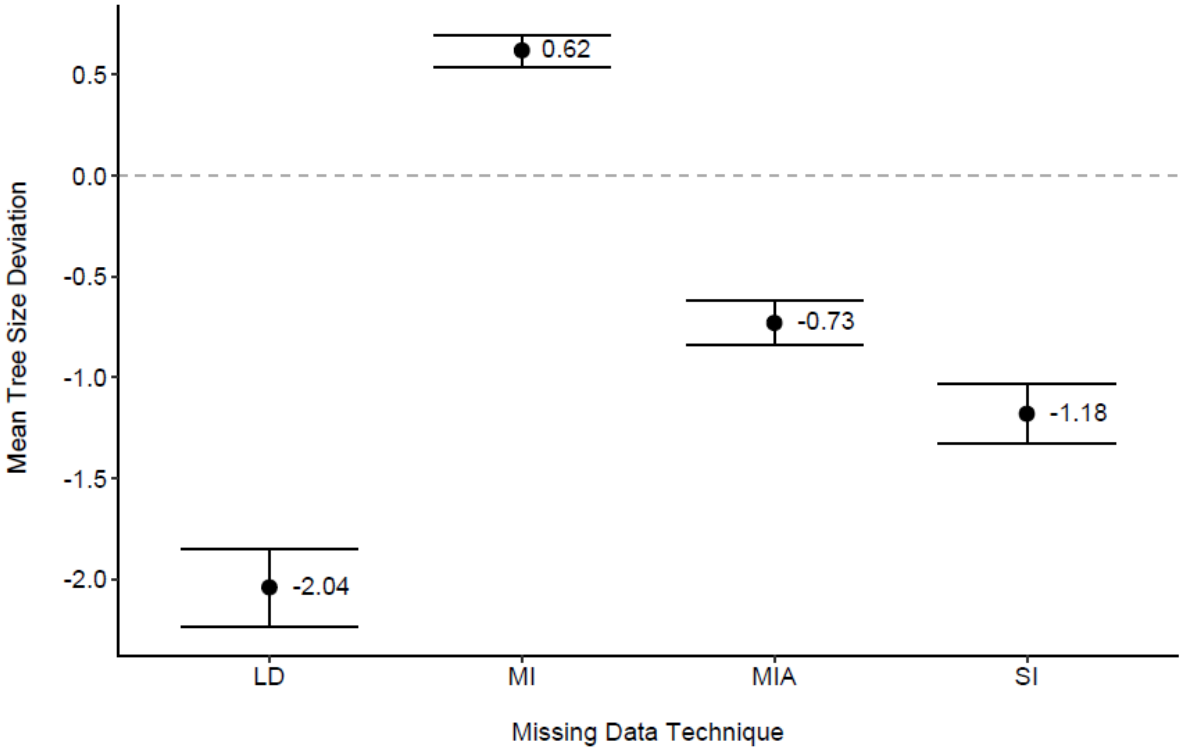

Figure 5. Mean tree size deviation and 95% confidence intervals of LD, MI, MIA and SI for the Early Childhood Longitudinal Study-Kindergarten reading ability outcome data, averaged over the different missingness scenarios and subsamples of the of the design. The dotted grey line represents the "true" tree size for the benchmark data without missingness. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes

the main effects indicate that the tree size accuracy of the missing data techniques deteriorated for increasing proportions of missing cases (panel A) and missing values (panel B). Also, all missing data techniques achieved the lowest tree size accuracy for MNAR data, followed by MAR and MCAR, respectively (panel C). These figures show a similar pattern for tree size accuracy as compared to predictive accuracy.

Looking at the specific interactions, Figure 6A and 6B suggest that MI and MIA yielded a similar tree size accuracy for lower amounts of missing data (i.e., 15% and 30% of missing values and 50% and 75% of cases with missing), whereas MI yielded a higher tree size accuracy than MIA for higher amounts of missing data (i.e., 75% of cases with missing and 50% of missing values). In addition, LD yielded a more serious deterioration in tree size accuracy with increasing number of cases as compared to MI, MIA and SI. Also, for LD, the decrease in tree size accuracy with increasing proportions of missing values was steeper from

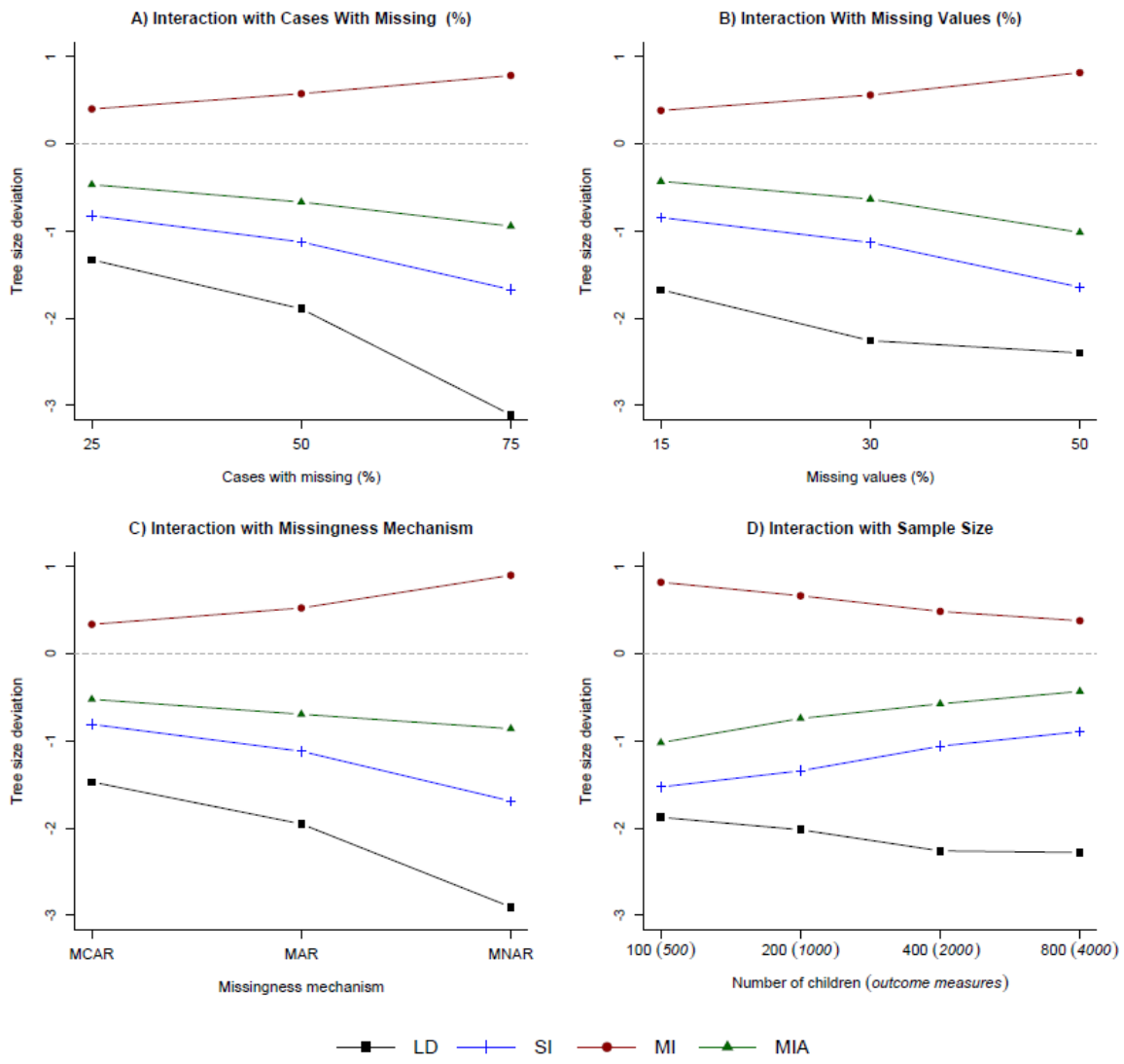

Figure 6. Interaction effects between missing data techniques and A) cases with missing mechanism, B) proportion of missing values, C) missingness mechanism and D) sample size on the tree size deviation for the Early Childhood Longitudinal Study-Kindergarten reading ability outcome data. The dotted grey line represents the "true" tree size for the benchmark data without missingness. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.

15% to 30% of missing values, but flatter from 30% to 50% of missing values. Again, note that this pattern can be attributed to the simulation design used by this thesis.

Similar to the results on the predictive accuracy, Figure 6C indicates that MI achieved a higher tree size accuracy than MIA for MCAR and MAR data, whereas MIA showed a

higher tree size accuracy for MNAR data. In addition, the tree size accuracy of LD, MI and SI was more strongly affected by MNAR data as compared to the tree size accuracy of MIA.

Figure 6D visualizes the significant two-way interaction effect between sample size and missing data technique. Across sample sizes, MI showed the highest tree size accuracy, slightly higher than MIA, followed by SI and thereafter LD. Compared to the other techniques, the tree size accuracy of LD deteriorated as sample size increased. Remarkably, as sample size increased, the tree size accuracy of LD decreased while the tree size accuracy of LD, SI and MI improved. No plausible explanation was found for this pattern.

Discussion

Summary of findings

This thesis evaluated the performance of LD, MIA, MI and SI for handling missing data in GLMM trees, in terms of predictive accuracy and tree size accuracy. Overall, MI yielded the best performance, closely followed by MIA, which exhibited approximately similar performance. SI performed somewhat worse than MI and MIA, whereas LD showed a substantially inferior performance compared to the other missing data techniques. When comparing the size of fitted GLMM trees with those fitted on the complete data, MI tended to overfit and yield bigger trees, whereas LD, SI and MIA tended to underfit and yield smaller trees. For all missing data techniques, the tree size accuracy and predictive accuracy showed an approximately pattern of performance. As expected, all missing data techniques yielded the highest performance for MCAR data followed by MAR and then MNAR data. Also, the performance of all missing data techniques decreased for increasing proportions of missing values and cases with missing and improved for larger sample sizes. Individually, MI and MIA performed very similar for lower amounts of missing data and missing cases (i.e., 15% and 30% of missing values and 25% and 50% of cases with missing). However, MI outperformed MIA when data was MCAR and MAR and for larger amounts of missing data (i.e., 50% of missing values and 75% of cases with missing), whereas MIA outperformed MI when data was MNAR. Furthermore, the performance of LD was lowest across all conditions and its performance deteriorated even further as the number of cases with missing increased. Also, the tree size accuracy of LD deteriorated for increasing sample sizes.

Comparison with previous studies

MI showing the highest performance is in line with prior research (e.g., He, 2006; Twala, 2005), supporting the belief that it can be regarded as a state-of-the art technique for handling missing data (Schafer & Graham, 2002). Even though these results indicate that MI is the best missing data technique in terms of predictions for handling missing data in GLMM trees, MI lacks interpretability: in contrast to the other missing data techniques, using MI resulted in ensembles of trees which varied individually. Also, recall that the tree size accuracy of MI was computed through averaging the tree size over an ensemble of 10 trees. Thus, the tree size accuracy of MI should not be taken as a straightforward indicator of model complexity and the relatively higher tree size observed for MI should perhaps be multiplied by the size of the ensemble

Finding MIA to outperform SI and LD and closely following the performance of MI is in line with research from Twala (2005). MIA yields decision trees whose accuracy in terms of predictions and tree size is approximately similar to MI, and higher for MNAR data. In addition, MIA tended to underfit and yield smaller and less complex trees than MI, which might indicate less power to detect splits but also a lower type I error rate. As MIA achieved an approximately similar accuracy to MI while resulting in highly interpretable decision trees rather than tree ensembles, it yields a better accuracy-complexity trade-off. Therefore, MIA can be regarded as a good technique for handling missing data in GLMM trees for both interpretational and predictive purposes, which is particularly advantageous for MNAR data.

LD showing the worst performance overall is in accordance with prior research (e.g., Ding & Simonoff, 2010; He, 2006; Twala, 2005). Moreover, as LD showed an inferior tree size accuracy, its underfitting does not only indicate a severe lack of power to detect splits, but also that it yields overly simplistic decision trees which do not accurately capture the relationship between the predictors and the outcomes (Freitas, 2014). Consequently, LD can be regarded as an inferior missing data technique for handling missing data in GLMM trees.

Finding SI to outperform LD while being outperformed by MI and MIA corroborates previous research (Twala, 2005). SI yields trees whose accuracy in terms of predictions and tree size is low compared to MI and MIA, but higher than LD. In addition, SI tended to underfit, which, based on its low tree size accuracy, suggests both a lack of power to detect splits and that it yields simplistic decision trees which underrepresent the relationship between the predictors and outcomes. Consequently, SI can be regarded as a poor technique for handling missing data in GLMM trees.

Regarding the performance of LD, MI, MIA and SI, an interesting finding is that the predictive accuracy and size accuracy showed an approximately similar pattern of performance. This finding may indicate that for handling missing data in GLMM trees, the predictive accuracy of missing data techniques could at least partly be attributed to the accuracy of the fitted tree. However, it should be noted that this is a preliminary conclusion which merely scratches the surface of this area of research as tree accuracy entails the accuracy of splitting variables and values, in addition to tree size (Breiman, Friedman, Stone & Olshen, 1984; Strobl, Malley & Tutz, 2009).

Implications

Based on the results of this thesis, LD is discouraged for handling missing data in GLMM trees as its accuracy of predictions and interpretations is substantially worse than the

other techniques. The use of SI is preferred over LD as it is more accurate but is still advised against due to being less accurate than MI and MIA. Alternatively, using MI is recommended predominantly for prediction purposes as it yields the most accurate predictions, but lacks in interpretability. However, using MI might not be recommended as it is computationally more demanding (e.g., Twala, 2005) and more difficult to correctly implement due to the challenge of having to specify a correct imputation model (Van Buuren; Brand; Groothuis-Oudshoornb & Rubin, 2006**)** compared to the other techniques. For interpretability, and when a lower complexity of the fitted model is preferred, it is recommended to use MIA for handling missing data in GLMM trees as it yields highly interpretable trees with an accuracy similar to MI. A particularly attractive property of MIA is that it is the best missing data technique for handling MNAR data, which is deemed challenging to deal with (Little & Rubin, 2019) and because very few techniques currently exist that can handle MNAR data (Little & Rubin, 2019). Moreover, using MIA may be convenient in practice as it is easy to implement and does not require settings or tuning of parameters (Twala, 2005). Therefore, for fitting GLMM trees it is suggested to add MIA as an option to the "glmertree" package, providing users a substantially better alternative for handling missing data than LD, which is the current default (Fokkema, 2018).

Strengths, limitations and future research

There are two strengths worth mentioning that enrich the relevance and utility of this thesis as an element of reference. First, the performance of missing data techniques was assessed in terms of both the predictive accuracy and the tree size accuracy, whereas prior research focuses primarily on evaluating the predictive performance. By also assessing the tree size, this thesis could not only illuminate the predictive performance but also the effects on interpretability of the different missing data techniques for GLMM trees. Second, the missing data design of this thesis covered a broad range of different sample sizes, missingness mechanisms and proportions of missing values and incomplete cases. Prior studies predominantly focused on a smaller subset of design factors, such as only considering a MCAR scheme and not varying sample sizes (Lin & Tsai, 2020). As the simulation design included a large combination of different missingness scenarios that are encountered in practice, the robustness of the results is supported. Notably, a particular advantage of the design was the evaluation of the effect introduced by the proportion of cases with missing, which is considered to be more representative to real world data as an equal spread of

missingness across all cases is rarely encountered in practice and can be considered as rather artificial (Hapfelmeier, Hothorn & Ulm, 2012).

Although this thesis covered a wide range of missingness scenarios, a limitation is that the missingness pattern was limited to varying different proportions of cases with missingness, whereas other common patterns of missingness were not considered. The three most common missingness patterns entail missingness being confined to one variable (univariate), missingness occurring in any set of variables for any case (arbitrary) or a monotone pattern where all prior variables are observed if the later variable is observed (Little & Rubin, 2019). As the relative performance of missing data techniques might depend on the missingness patterns, future studies should systematically vary the missingness pattern. Particularly, a monotone pattern should be investigated as these frequently occur due to attrition in longitudinal datasets, datasets for which GLMM trees are designed to assess (Fokkema, 2018).

Another limitation of this thesis might be the difficulty to make broad generalizations to real-life settings. First, a simulation design tailored specifically for this thesis was used. Therefore, it could be that the relative performance of the missing data techniques is specific for the missingness scenarios used. Second, a single dataset (i.e., ECLS-K) was used to draw subsamples from. Even though a sensitivity analysis with different outcomes was carried out, the structure of all the subsamples was quite similar as they were comprised of the same set of predictors. Therefore, the observed performance of the discussed missing data techniques may be specific for data structures similar to the ECLS. To validate the findings of this thesis, future studies are advised to use a wide range of real-life data, with a broad combination of varying data structures and of missingness scenarios.

Furthermore, the assessment was limited to LD, MI, MIA and SI while other techniques for handling missing data in decision-tree methods exist, such as surrogate variable splitting (Breiman, Friedman, Olshen & Stone, 1984). Moreover, techniques that can average the ensembles of MI into a single structure, such as case weighting the imputed datasets (Wood, White & Royston, 2008) and using a stochastic multiple imputation algorithm (Wallace, 2010), were not included. It is plausible that one of the excluded techniques (for averaging MI into a single tree) yields an even better solution for handling missing data in GLMM tree. Thus, future studies are advised to include other existing missing data techniques and techniques which allow MI to be combined into a single resultant decision tree.

For future research, there are two possible improvements for the performance of MIA that are worth mentioning. First, perhaps a more liberal Bonferroni correction could be chosen

to correct for the MIA duplicating numerical variables with missing, as the number of potential predictor variables directly affects the power of the variable selection tests for potential splits of the GLMM tree algorithm (Fokkema et al., 2018). Since these tests are Bonferroni corrected by default, the likelihood that a potential predictor variable yields a *p* value lower than the prespecified *a* level decreases as the number of potential predictor variables increases. If a less conservative Bonferroni correction is chosen, this would likely result in MIA yielding trees which underfit less, resulting in more tree size accuracy and an increased performance of MIA. Second, experimentation during the thesis process through coding a numerical variable as a categorical variable suggested that MIA performs better when there are more categorical variables present in a dataset, which is a similar result found by Twala (2005). In a similar vein, note that MIA is easier to implement for categorical than for numerical variables as it does not require duplicating variables with missing.

Main conclusions

This thesis provides a first evaluation of how missing data should be handled in GLMM trees by assessing the relative strengths and weaknesses of LD, MI, MIA and SI. In sum, the results outline the following recommendations:

- MIA is recommended for handling missing data in GLMM trees as it is a relatively easy-to-apply missing data technique which yields highly interpretable trees with a predictive accuracy that is approximately similar to MI.
- MI is recommended predominantly for prediction purposes as it yields the most accurate predictions but lacks in interpretability.
- Using LD is discouraged as it always performs substantially worse than the other techniques.
- The use of SI should be preferred over LD as it performs better but is still discouraged due to being less accurate than MI and MIA.

Literature

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, California: Wadsworth International Group.

Ding, Y., & Simonoff, J. S. (2010). An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, *11*(1). https://www.jmlr.org/papers/volume11/ding10a/ding10a.pdf

Doove, L. L., Dusseldorp, E., Van Deun, K., & Van Mechelen, I. (2014). A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Advances in Data Analysis and Classification*, *8*(4), 403-425. https://doi.org/10.1007/s11634-013-0159-x

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford press. http://hsta559s12.pbworks.com/w/file/fetch/52112520/enders.applied

Feelders, A. (1999, September). Handling missing data in trees: surrogate splits or statistical imputation? *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 329-334). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-48247-5_38

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, *15*(1), 1-10. https://doi.org/10.1145/2594473.2594475

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior research methods*, *50*(5), 2016-2034 https://doi.org/10.3758/13428-017-0971-x

Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. (2020). Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research*, 1-13. https://doi.org/10.1080/10503307.2020.1785037

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & probability letters*, *81*(4), 451-459. https://doi.org/10.1016/j.spl.2010.12.003

Hapfelmeier, A., Hothorn, T., & Ulm, K. (2012). Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. *Computational Statistics & Data Analysis*, *56*(6), 1552-1565. https://doi.org/10.1016/j.csda.2011.09.024

He, Y. (2006) *Missing data imputation for tree-based models* (Publication No. 3226020) [Doctoral dissertation, University of California, LA]. ProQuest Dissertations Publishing.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674. https://doi.org/10.1198/106186006x133933

Kim, H., & Loh, W. Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, *96*(454), 589-604. https://doi.org/10.1198/016214501753168271

Kline, P. (2013). *The handbook of psychological testing* (2nd ed.). London, UK: Routledge. https://doi.org/10.4324/9781315812274

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological measurement*, *56*(5), 746-759. https://doi.org/10.1177/00131644960560

Kopf, J., Augustin, T., & Strobl, C. (2010). The potential of model-based recursive partitioning in the social sciences-Revisiting Ockham's Razor. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary Issues in Exploratory Data Mining in the Behavioural sciences* (pp. 97-117). London, UK: Routledge. https://doi.org/10.4324/9780203403020-12

Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*(2), 1487-1509. https://doi.org/10.1007/s10462-019-09709-4

Little, R.J.A and Rubin, D.B. (1987). *Statistical Analysis with missing data.* New York, NY: Wiley.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287-296. https://doi.org/10.2307/1391878

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed., Vol. 793). New York, NY: John Wiley & Sons. https://doi.org/10.1002/9781119482260

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC medical research methodology*, *10*(1), 7. https://doi.org/10.1186/1471-2288-10-7

Martin, D. P. (2015). *Efficiently Exploring Multilevel Data with Recursive Partitioning.* [PhD Dissertation, University of Virginia]. https://dpmartin42.github.io/extras/dissertation.pdf

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (Eds.). (2014). *Handbook of missing data methodology*. Boca Raton: CRC Press. https://doi.org/10.1201/b17622

Provost, F., & Saar-Tsechanski, M. (2007). Handling Missing Values when Applying Classification Models. *Journal of Machine Learning Research*, *8*. https://www.jmlr.org/papers/volume8/saar-tsechansky07a/saar-tsechansky07a.pdf

Quinlan, J.R., 1986. Induction of decision trees. *Mach. Learn. 1*(1), 81−106. https://doi.org/10.1007/bf00116251

R Core Team. (2020). *R language definition*. Vienna, Austria: R foundation for statistical computing.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial intelligence review*, *33*(1-2), 1-39. https://doi.org/10.1007/s10462-009-9124-7

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, NY: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, *91*(434), 473-489. https://doi.org/10.1080/01621459.1996.10476908

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton: CRC Press. https://doi.org/10.1201/9781439821862

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147. https://doi.org/10.1037/1082-989x.7.2.147

Schenker, N., & Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational statistics & data analysis*, *22*(4), 425-446. https://doi.org/10.1016/0167-9473(95)00057-7

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine learning*, *86*(2), 169-207. https://doi.org/10.1007/s10994-011-5258-3

Stegmann, G., Jacobucci, R., Serang, S., & Grimm, K. J. (2018). Recursive Partitioning with Nonlinear Models of Change. *Multivariate Behavioral Research 53*(4), 559–570. https://doi.org/10.1080/00273171.2018.1461602

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*(4), 323-348. https://doi.org/10.1037/a0016973

Twala, B. E. (2005). *Effective techniques for handling incomplete data using decision*

trees [Doctoral dissertation, The Open University]. http://oro.open.ac.uk/59753/1/418465.pdf

Twala, B. E. T. H., Jones, M. C., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, *29*(7), 950-956. https://doi.org/10.1016/j.patrec.2008.01.010

Valdiviezo, H. C., & Van Aelst, S. (2015). Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*, *311*, 163-181. https://doi.org/10.1016/j.ins.2015.03.018

Van Buuren, S. (2018). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/9780429492259

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, *76*(12), 1049-1064. https://doi.org/10.1093/aje/kwp42

Van Buuren, S., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45(3). https://doi.org/10.18637/jss.v045.i03

Wallace, M. L., Anderson, S. J., & Mazumdar, S. (2010). A stochastic multiple imputation algorithm for missing covariate data in tree-structured survival analysis. *Statistics in medicine*, *29*(29), 3004-3016. https://doi.org/10.1002/sim.4079

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in medicine*, *27*(17), 3227-3246. https://doi.org/10.1002/sim.3177

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492-514. https://doi.org/10.1198/106186008x319331

Appendix A

Descriptive statistics for the math and science ability outcome data

Table A1.

Descriptive statistics of the baseline Early Childhood Longitudinal Study-Kindergarten science outcome data (*N* = 6625)

	M (*SD*) or %	Range
Gender		
Male	49.5%	
Female	50.5%	
Race		
White, non-Hispanic	65.7%	
Black or African American, non-Hispanic	9.7%	
Hispanic, race specified	7.6%	
Hispanic, race not specified	7.7%	
Asian	4.0%	
Native Hawaiian, other Pacific Islander	1.0%	
American Indian or Alaska native	1.9%	
More than one race, non-Hispanic	2.5%	
First time in kindergarten		
Yes	96.5%	
No	3.5%	
Socioeconomic status	0.15 (0.78)	-4.75 – 2.75
Gross motor skills	6.44 (1.79)	0.00 – 8.00
Fine motor skills	6.08 (1.95)	0.00 – 9.00
Interpersonal skills	3.05 (0.61)	1.00 – 4.00
Self-control	3.15 (0.59)	1.00 – 4.00
Internalizing problem behaviour	1.49 (0.49)	1.00 – 4.00
Externalizing problem behaviour	1.56 (0.60)	1.00 – 4.00
Age at baseline (in months)	73.72 (4.16)	64.00 – 88.00
Reading ability[1]		
Third-grade year	-0.49 (0.64)	-2.66 – 1.45
Fifth-grade year	0.11 (0.63)	-2.10 – 2.15
Eight-grade year	1.05 (0.82)	-1.63 – 3.00

[1]Theta scores, reading ability was measured during the fall.

Table A2.

Descriptive statistics of the baseline Early Childhood Longitudinal Study-Kindergarten math ability outcome data (*N* = 6512)

	M (*SD*) or %	Range
Gender		
Male	49.2%	
Female	50.8%	
Race		
White, non-Hispanic	68.0%	
Black or African American, non-Hispanic	9.7%	
Hispanic, race specified	6.5%	
Hispanic, race not specified	6.4%	
Asian	4.2%	
Native Hawaiian, other Pacific Islander	1.0%	
American Indian or Alaska native	1.75%	
More than one race, non-Hispanic	2.5%	
First time in kindergarten		
Yes	96.7%	
No	3.3%	
Socioeconomic status	0.16 (0.78)	-4.75 – 2.75
Gross motor skills	6.45 (1.79)	0.00 – 8.00
Fine motor skills	6.08 (1.95)	0.00 – 9.00
Interpersonal skills	3.05 (0.61)	1.00 – 4.00
Self-control	3.15 (0.59)	1.00 – 4.00
Internalizing problem behaviour	1.49 (0.49)	1.00 – 4.00
Externalizing problem behaviour	1.56 (0.59)	1.00 – 4.00
Age at baseline (in months)	73.75 (4.15)	64.00 – 88.00
Reading ability[1]		
Kindergarten year	-0.64 (0.46)	-2.31 – 1.10
First-grade year	0.21 (0.39)	-1.73 – 1.44
Third-grade year	0.86 (0.28)	-0.40 – 1.83
Fifth-grade year	1.11 (0.27)	0.07 – 1.94
Eight-grade year	1.38 (0.36)	0.37 – 2.40

[1]Theta scores, reading ability was measured during the fall.

Appendix B

R-code for executing missingness in attributes (MIA)

```
#' @param formula: model formula, used to identify predictors and response.
#' @param data: dataset with missing values.
#' @param numerical_missing_levels: list of function(s) to use for imputing missing
#` values for numerical variables. The length of the argument determines
#` the number of variables created: The default value yields two new variables
#` for every numeric variable with missings: One which replaces missing values
#` with \code{min(x)-1}; one which replaces missing values with \code{max(x)+1}.
#` Alternatively, a numeric vector may be specified. E.g., \code{c(-Inf, Inf)}
#` or \code{c(-99, 99)} will impute these values for all numerical variables
#` with missings. If a single value (numeric value or function) is specified,
#` the value or function will be applied to fill in missing values in numerical
#` variables, and no additional variable will be created. Note that for CART and MOB
#` trees, the specific values specified (e.g., min(x)-1 or -Inf) are likely inconsequential.
#` For ctree, which employs linear asSociation tests for split selection, it will
#` likely make a difference, and perhaps the default should be preferred.
#' @param categorical_missing_level character vector of length one. Specifies
#' the label of the level used to code missing values for categorical predictors.
#' @param ordered_missing_level character vector of length one or two, or list with single
#` function. If of class character, specifies the label(s) of the levels used to code missing
#` values for ordered categorical predictors. If of class function, the specified function
#` will be employed to fill in missing values.
#' @param num_ord_labels: Character vector of length two. Specifies the labels
#` appended to the variable names of numerical and ordered variables with
#`` missings. Only used when argument numerical_missing_levels
#' and/or ordered_missing_levels have length 2.
MIA <- function(formula, data,
                numerical_missing_levels = list(
                  function(x) min(x, na.rm = TRUE)-1,
                  function(x) max(x, na.rm = TRUE)+1),
                categorical_missing_level = "NA",
                ordered_missing_levels = c("NA", "NA"),
                num_ord_labels = c("_m1", "_m2")) {
```

```
## Check arguments:
if (!inherits(data, "data.frame")) {
  warning("Argument 'data' should specify a data.frame.")
}
if (inherits(numerical_missing_levels, "list")) {
  if (!all(sapply(numerical_missing_levels,
                  function(x) inherits(x, c("function", "numeric", "integer"))))) {
    warning("Argument numerical_missing_levels should specify a function, or
            one or two numeric values.")
  }
} else if (!inherits(numerical_missing_levels, c("function", "numeric", "integer"))) {
  warning("Argument numerical_missing_levels should specify a function, or
          one or two numeric values.")
}
if (inherits(ordered_missing_levels, "list")) {
  if (!all(sapply(ordered_missing_levels,
                  function(x) inherits(x, c("function", "character"))))) {
    warning("Argument ordered_missing_levels should specify a function, or a
            character vector of length one or two.")
  }
} else if (!inherits(ordered_missing_levels, c("function", "character"))) {
  warning("Argument ordered_missing_levels should specify a function, or a
          character vector of length one or two.")
}
if (!class(ordered_missing_levels) %in% c("function", "character")) {
  warning("Argument ordered_missing_levels should specify a function, or a
          character vector of length one or two.")
}
if (!inherits(categorical_missing_level, "character")) {
  warning("Argument categorical_missing_level should provide a character vector
          of length one.")
}
```

```
## Keep only observations with non-missing values for response variables:
response_names <- all.vars(formula(Formula(formula), rhs = 0, lhs = NULL)[[2]])
if (any(rowSums(sapply(data[response_names], is.na)) > 0)) {
  warning("Observations with missing response variable values are present and will
        be removed from the dataset.")
}
data <- data[rowSums(sapply(data[response_names], is.na)) == 0, ]

## Identify ordinal, categorical and numerical variables with missings:
vars_with_missings <- colSums(is.na(data)) > 0
ord_missing_names <- names(data)[vars_with_missings &
                          sapply(data, function(x) inherits(x, "ordered"))]
cat_missing_names <- names(data)[vars_with_missings &
                          sapply(data, function(x) inherits(x, c("logical", "factor",
                                                "character"))) &
                          !sapply(data, function(x) inherits(x, "ordered"))]
num_missing_names <- names(data)[vars_with_missings &
                          sapply(data, function(x) inherits(x, c("numeric", "integer")))]


## Impute ordinal variables with missing:
for (i in ord_missing_names) {
  if (length(ordered_missing_levels) == 2L) {
    name1 <- paste0(i, num_ord_labels[1])
    data[ , name1] <- data[ , i]
    levels1 <- c(ordered_missing_levels[1], levels(data[ , name1]))
    data[, name1] <- as.character(data[ , name1])
    data[is.na(data[ , i]), name1] <- ordered_missing_levels[1]
    data[, name1] <- ordered(data[, name1], levels = levels1)
    name2 <- paste0(i, num_ord_labels[2])
    data[ , name2] <- data[ , i]
    levels2 <- c(levels(data[ , name2]), ordered_missing_levels[2])
    data[, name2] <- as.character(data[ , name2])
    data[is.na(data[ , i]), name2] <- ordered_missing_levels[2]
```

```
    data[, name2] <- ordered(data[, name2], levels = levels2)
    data <- data[ , -which(names(data) == i)]
  } else if (length(ordered_missing_levels) == 1L) {
    data[is.na(data[ , i]), i] <- ifelse(
      is.function(ordered_missing_levels),
      ordered_missing_levels(data[ , i]),
      ordered_missing_levels)
  }
}


## Impute categorical variables with missings:
for (i in cat_missing_names) {
  if (is.logical(data[ , i])) data[ , i] <- factor(data[ , i])
  if (is.character(data[ , i])) data[ , i] <- factor(data[ , i])
  levels(data[ , i]) <- c(levels(data[ , i]), categorical_missing_level)
  data[ , i][is.na(data[ , i])] <- categorical_missing_level
}


## Impute numerical variables with missings:
for (i in num_missing_names) {
  if (length(numerical_missing_levels) == 2L) {
    name1 <- paste0(i, num_ord_labels[1])
    data[ , name1] <- data[ , i]
    data[is.na(data[ , i]), name1] <- ifelse(
      is.function(numerical_missing_levels[[1]]),
      numerical_missing_levels[[1]](data[ , i]),
      numerical_missing_levels[[1]])
    name2 <- paste0(i, num_ord_labels[2])
    data[ , name2] <- data[ , i]
    data[is.na(data[ , i]), name2] <- ifelse(
      is.function(numerical_missing_levels[[2]]),
      numerical_missing_levels[[2]](data[ , i]),
      numerical_missing_levels[[2]])
```

```
      data <- data[ , -which(names(data) == i)]
    } else if (length(numerical_missing_levels) == 1L) {
      data[is.na(data[ , i]), i] <- ifelse(
        is.function(numerical_missing_levels),
        numerical_missing_levels(data[ , i]),
        numerical_missing_levels)
    }
  }
  return(data)
}
```

Appendix C

Results for the math and science ability outcome data

Results for the math ability outcome data

Predictive accuracy

Table C1.

Mixed design ANOVA on the excess error for the 40 subsamples of the Early Childhood Longitudinal Study-Kindergarten math ability outcome data

Effect	*df*	*F*	*p*	$_{\text{partial}}\ \eta^2$
Technique	3	267.68	<.001	.88
Cases with missing (%)	2	262.66	<.001	.87
Missingness mechanism	2	192.00	<.001	.85
Missing values (%)	2	148.65	<.001	.79
Sample size	3	8.07	<.001	.40
Technique * cases with missing (%)	6	33.73	<.001	.46
Technique * missingness mechanism	6	29.40	<.001	.43
Technique * missing values (%)	6	6.95	<.001	.15
Technique * sample size	9	0.45	.90	.04

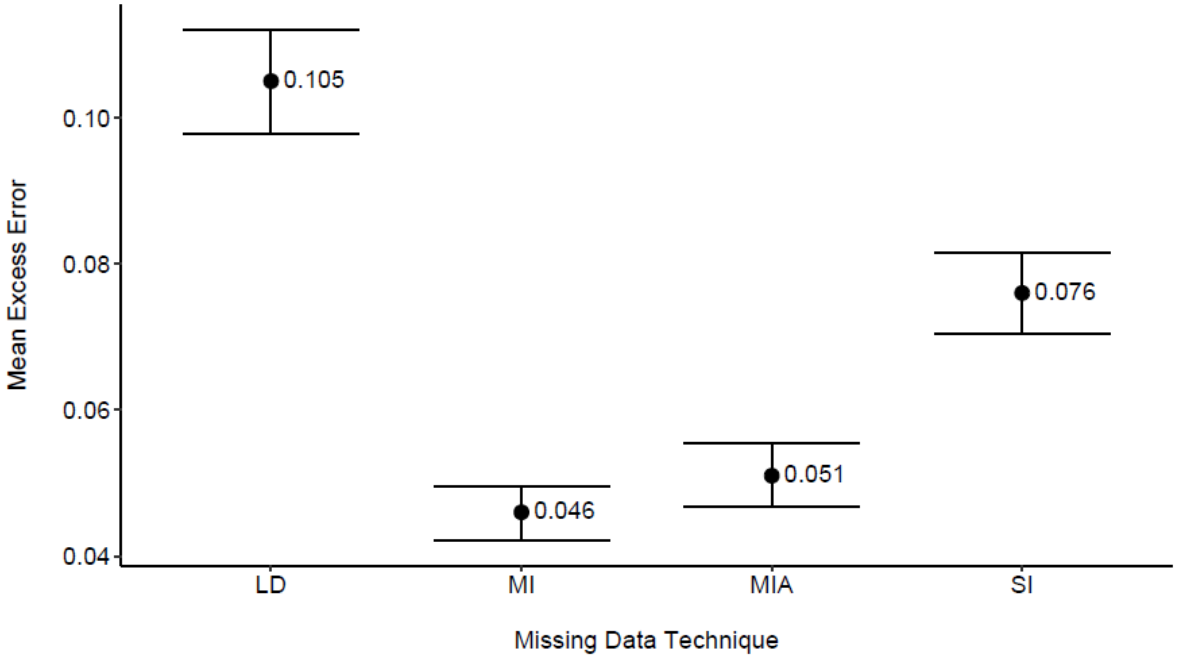

Figure C1. Mean excess error rate and 95% confidence intervals of LD, MI, MIA and SI for the Early Childhood Longitudinal Study-Kindergarten math ability outcome data reading ability outcome data, averaged over the different missingness scenarios and subsamples of the design. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes.

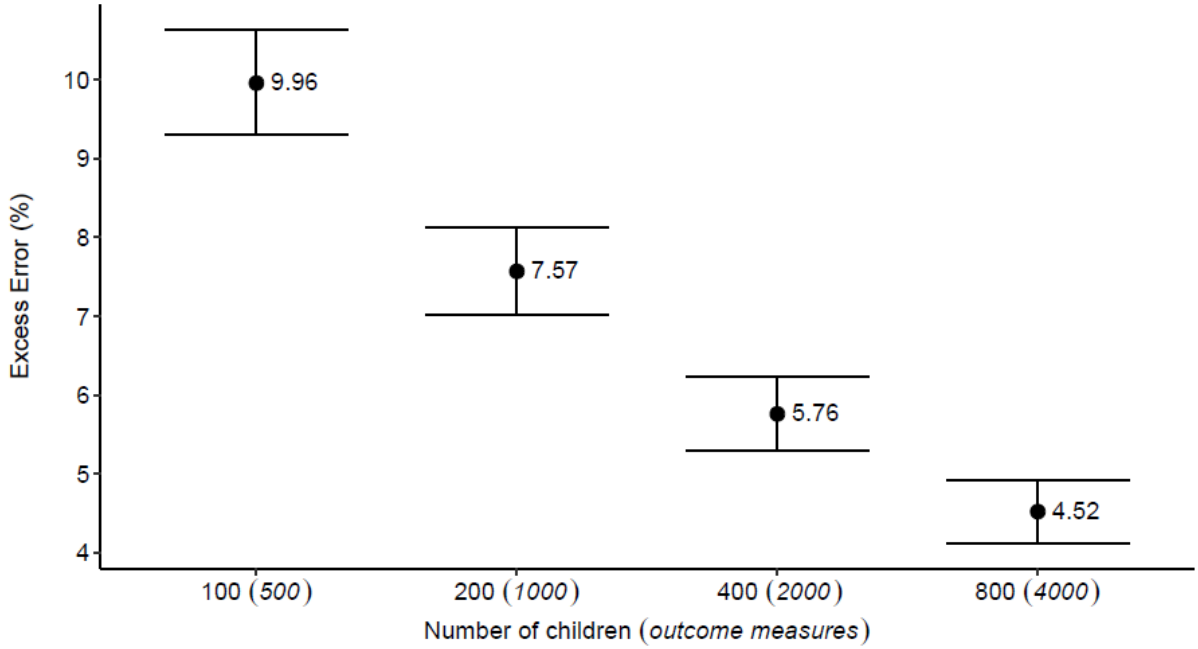

Figure C2. Mean excess error rate and 95% confidence intervals of sample size for the Early Childhood Longitudinal Study-Kindergarten math ability outcome data, averaged over the missingness data techniques and missingness scenarios of the design.

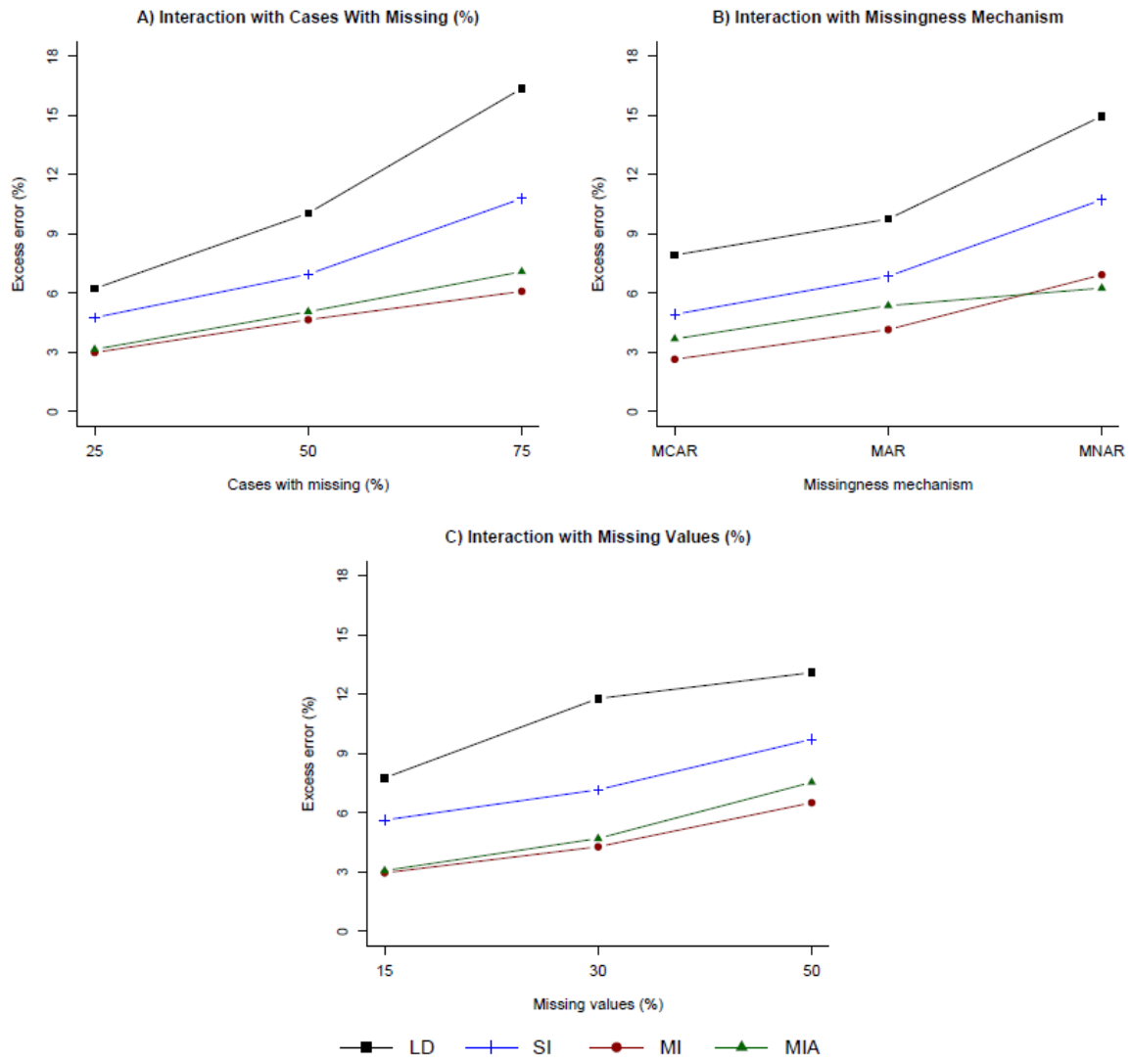

Figure C3. Interaction effects between missing data technique and A) missingness mechanism, B) proportion of cases with missing and C) proportion of missing values on the excess error for the Early Childhood Longitudinal Study-Kindergarten math ability outcome data. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.

Tree size accuracy

Table C2.

Mixed design ANOVA on the tree size deviation for the 40 subsamples of the Early Childhood Longitudinal Study-Kindergarten math ability outcome data

Effect	*df*	*F*	*p*	partial η^2
Technique	3	299.48	<.001	.89
Cases with missing (%)	2	167.75	<.001	.81
Missingness mechanism	2	148.65	<.001	.79
Missing values (%)	2	102.14	<.001	.72
Sample size	3	6.76	<.001	.35
Technique * cases with missing (%)	6	117.43	<.001	.75
Technique * missingness mechanism	6	99.75	<.001	.72
Technique * missing values (%)	6	42.35	<.001	.52
Technique * sample size	9	23.43	<.001	.66

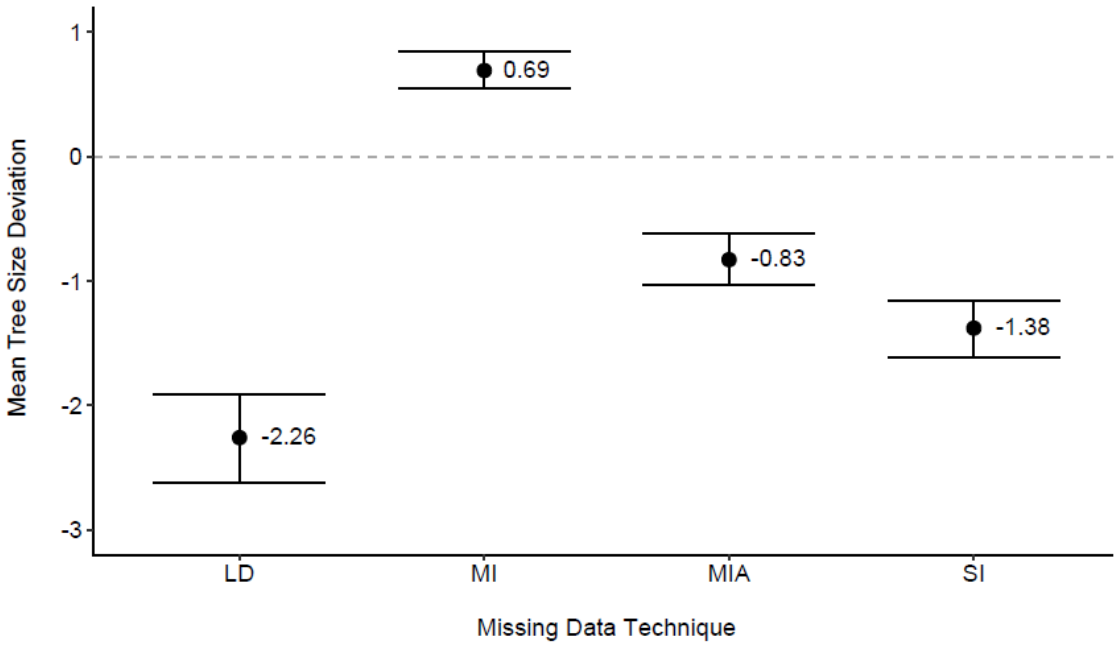

Figure C4. Mean tree size deviation and 95% confidence intervals of LD, MI, MIA and SI for the Early Childhood Longitudinal Study-Kindergarten math ability outcome data averaged over the different missingness scenarios and subsamples of the of the design. The dotted grey line represents the "true" tree size for the benchmark data without missingness. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes

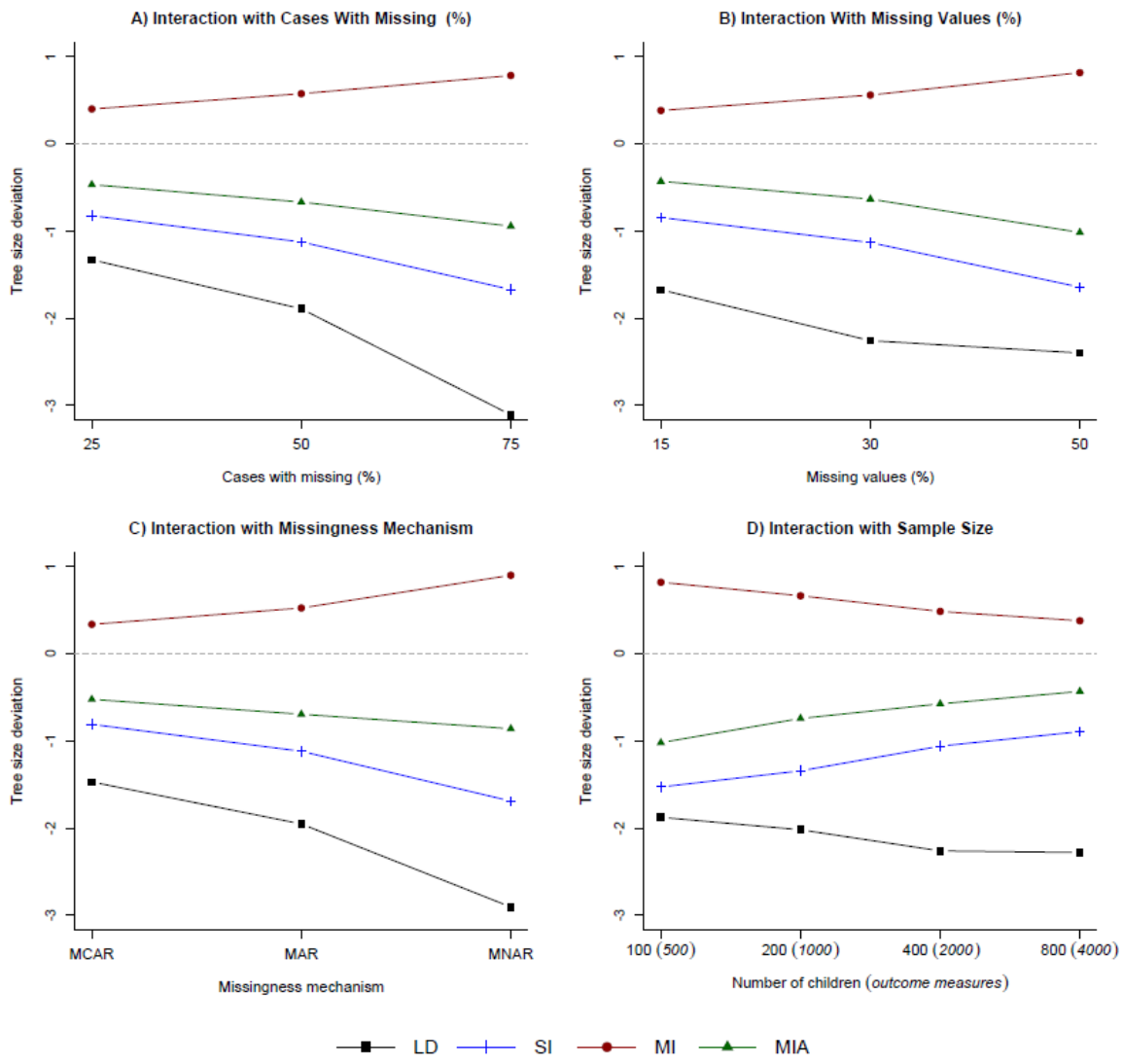

Figure C5. Interaction effects between missing data techniques and A) cases with missing mechanism, B) proportion of missing values, C) missingness mechanism and D) sample size on the tree size deviation for the Early Childhood Longitudinal Study-Kindergarten math ability outcome data. The dotted grey line represents the "true" tree size for the benchmark data without missingness. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.

Results for the science ability outcome data

Predictive accuracy

Table C3.

Mixed design ANOVA on the excess error for the 40 subsamples of the Early Childhood Longitudinal Study-Kindergarten science ability outcome data

Effect	*df*	*F*	*p*	$_{\text{partial}}\ \eta^2$
Technique	3	174.51	<.001	.83
Cases with missing (%)	2	161.00	<.001	.81
Missingness mechanism	2	152.51	<.001	.79
Missing values (%)	2	102.14	<.001	.72
Sample size	3	5.96	.002	.33
Technique * cases with missing (%)	6	23.55	<.001	.38
Technique * missingness mechanism	6	18.51	<.001	.32
Technique * missing values (%)	6	7.44	<.001	.16
Technique * sample size	9	1.08	.413	.08

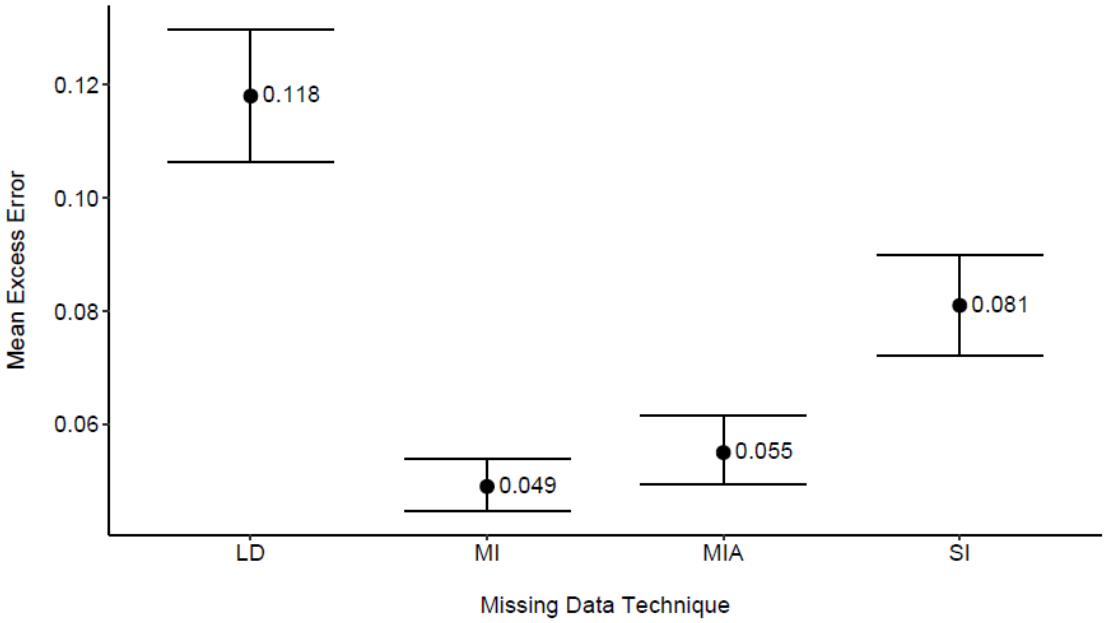

Figure C6. Mean excess error rate and 95% confidence intervals of LD, MI, MIA and SI for the Early Childhood Longitudinal Study-Kindergarten science ability outcome data reading ability outcome data, averaged over the different missingness scenarios and subsamples of the design. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes.

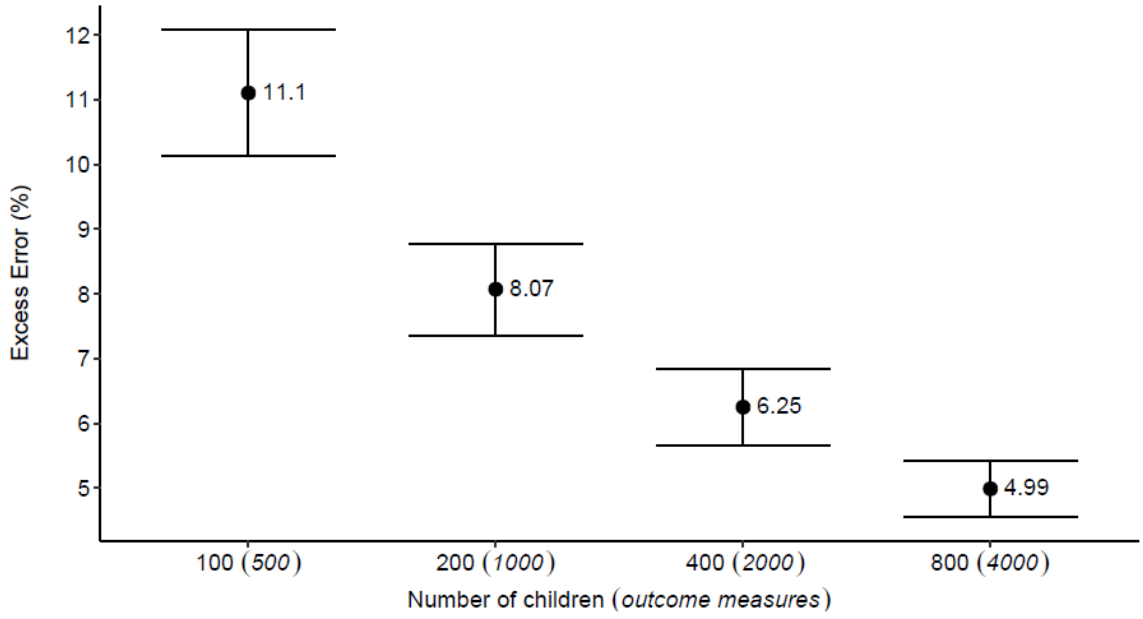

Figure C7. Mean excess error rate and 95% confidence intervals of sample size for the Early Childhood Longitudinal Study-Kindergarten science ability outcome data, averaged over the missingness data techniques and missingness scenarios of the design.

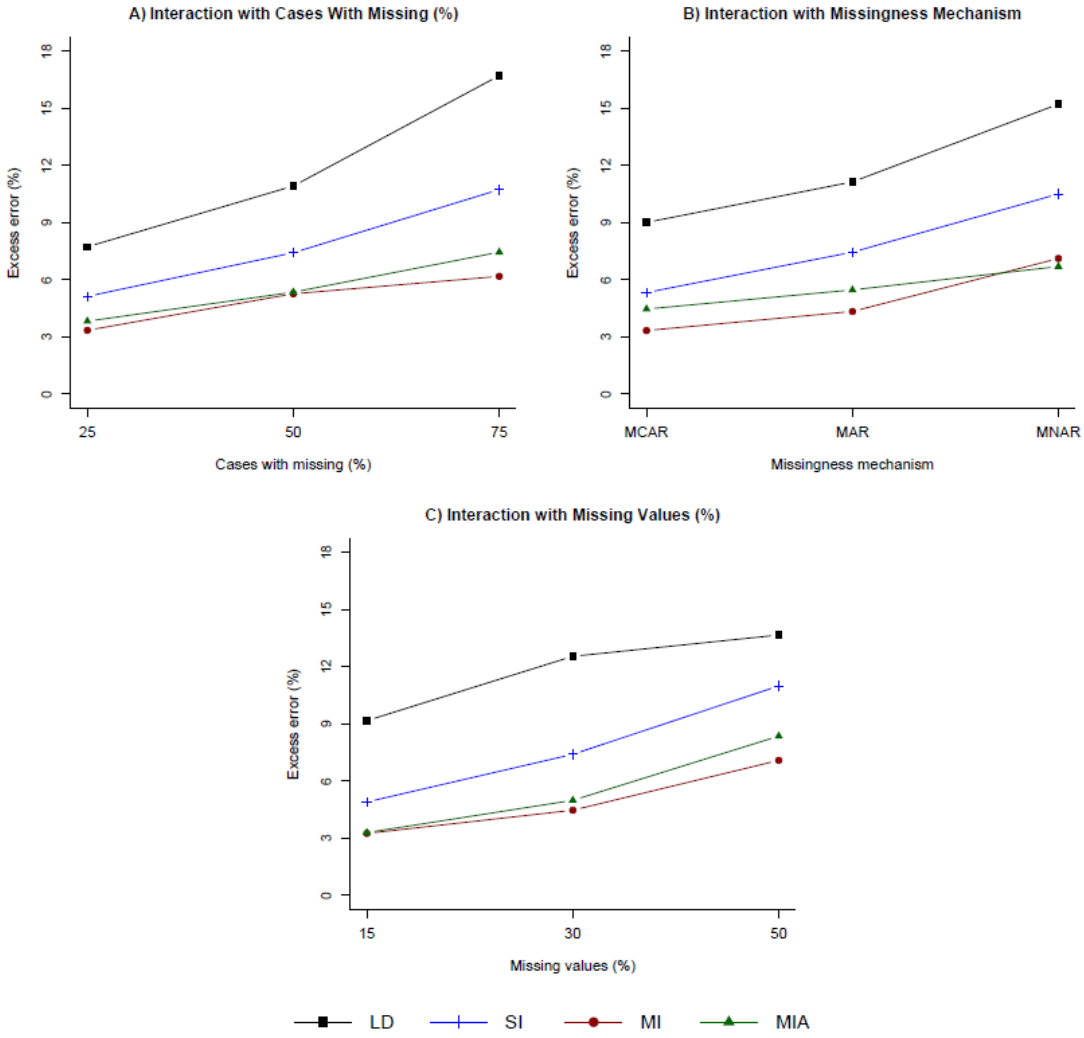

Figure C8. Interaction effects between missing data technique and A) missingness mechanism, B) proportion of cases with missing and C) proportion of missing values on the excess error for the Early Childhood Longitudinal Study-Kindergarten science ability outcome data. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.

Tree size accuracy

Table C4.

Mixed design ANOVA on the tree size deviation for the 40 subsamples of the Early Childhood Longitudinal Study-Kindergarten science ability outcome data

Effect	*df*	*F*	*p*	partial η^2
Cases with missing (%)	2	154.29	<.001	.79
Technique	3	124.60	<.001	.78
Missingness mechanism	2	98.17	<.001	.72
Missing values (%)	2	80.13	<.001	.67
Sample size	3	4.35	<.001	.27
Technique * cases with missing (%)	6	86.01	<.001	.69
Technique * missingness mechanism	6	52.85	<.001	.58
Technique * missing values (%)	6	42.33	<.001	.52
Technique * sample size	9	14.49	<.001	.63

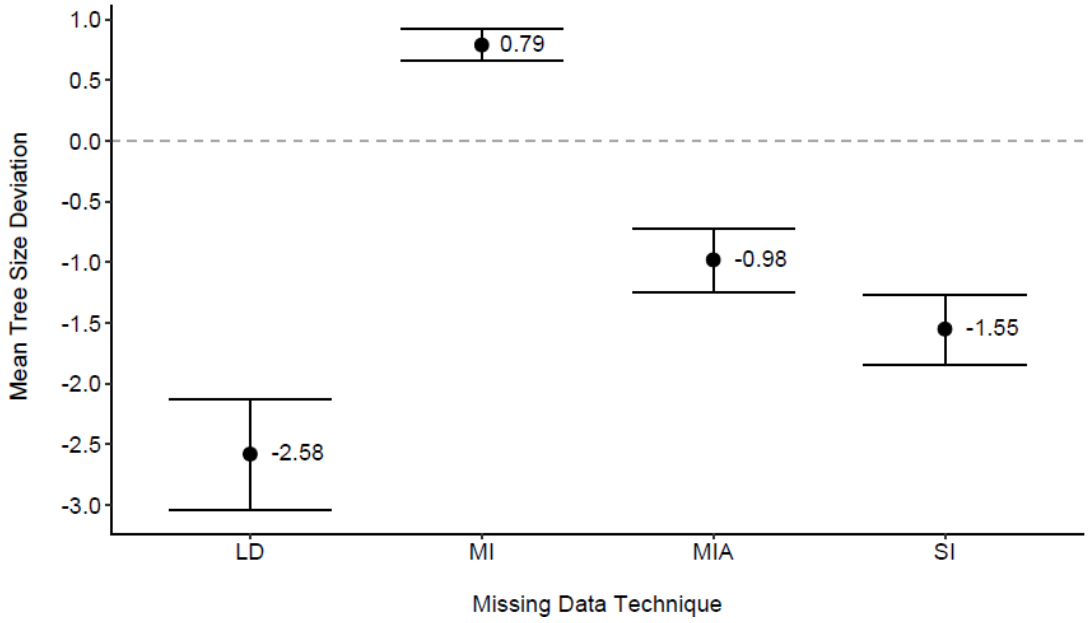

Figure D4. Mean tree size deviation and 95% confidence intervals of LD, MI, MIA and SI for the Early Childhood Longitudinal Study-Kindergarten science ability outcome data averaged over the different missingness scenarios and subsamples of the of the design. The dotted grey line represents the "true" tree size for the benchmark data without missingness. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes

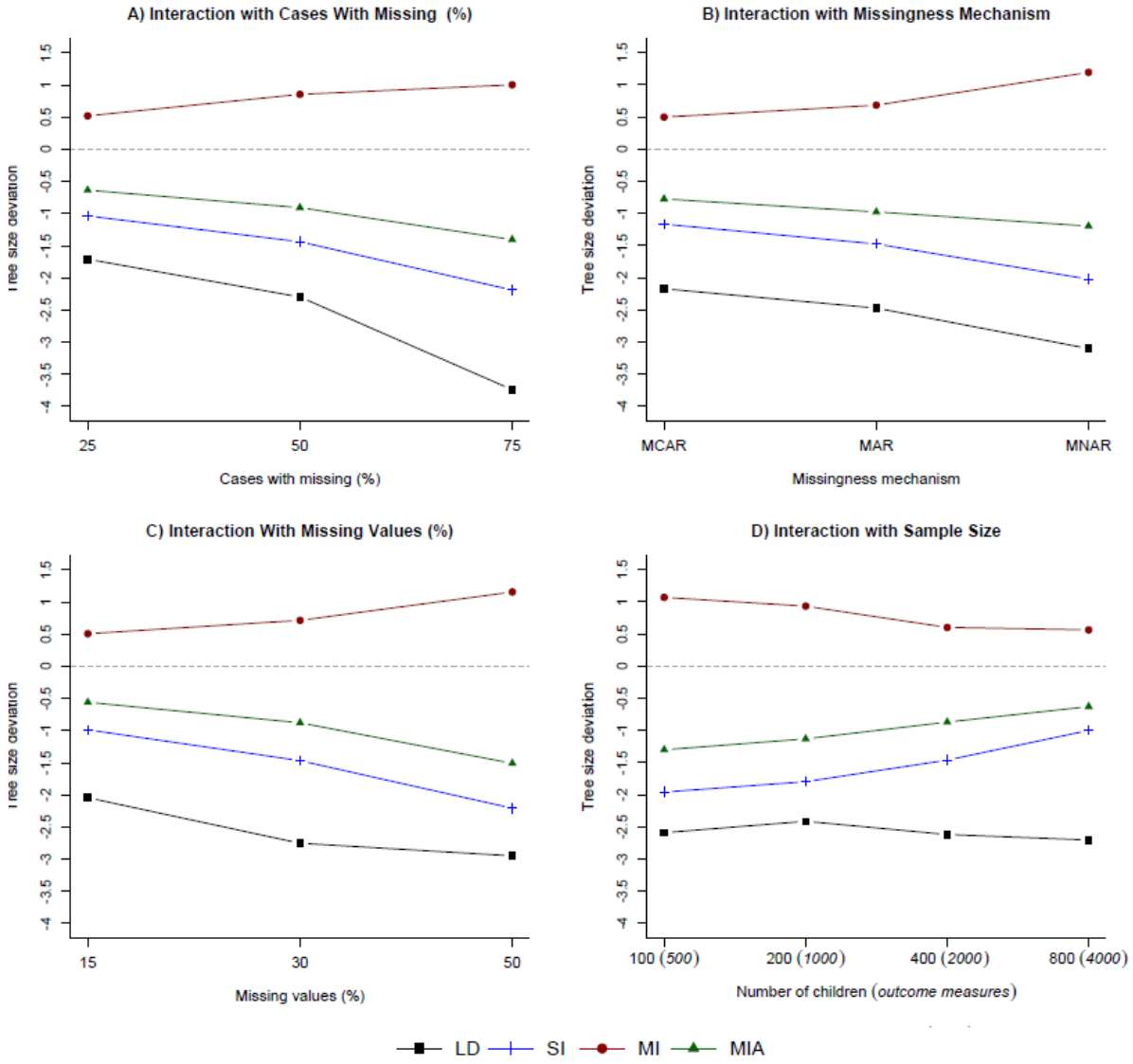

*Figure D5.*Interaction effects between missing data techniques and A) cases with missing mechanism, B) proportion of missing values, C) missingness mechanism and D) sample size on the tree size deviation for the Early Childhood Longitudinal Study-Kindergarten science ability outcome data. The dotted grey line represents the "true" tree size for the benchmark data without missingness. LD, Listwise deletion; SI, mean or mode single imputation; MI, multiple imputation; MIA, missingness incorporated in attributes; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.