Recursive partitioning of growth curve models with generalised linear mixed-effects regression trees

Maaike Jorink

Master's Thesis Psychology,
Methodology and Statistics Unit, Institute of Psychology,
Faculty of Social and Behavioral Sciences, Leiden University
Date: October 2018
Student number: 1907387
Supervisor: Dr. Marjolein Fokkema

Acknowledgments

In the first place, I would like to thank my supervisor, Dr. Marjolein Fokkema. At the start of my thesis I had only a slight idea of what recursive partitioning of growth curve models is. Marjolein assured me that at the end of my thesis I would understand (at least most of) what I have been working on. Without the help of Marjolein, I would not have dared to confidently say that I now do know what recursive partitioning of growth curve models is and how these can be very useful in psychological research. Thank you, Marjolein, for all your help and support during my thesis!

Secondly, I would like to thank the people close to me. I am grateful to my boyfriend for being supportive, patient, a listening-ear and in general for making me feel confident during this project. I would also like to thank my family for their support and helpfulness, and my friends for the, every now and then, needed distractions.

Abstract

In this study we investigated how to best fit a recursive partitioning growth curve model (RP-GCM) with the R package glmertree on longitudinal data. We used data on reading, math and science ability of children from kindergarten through eighth grade on five measurement occasions in the Early Childhood Longitudinal Study-Kindergarten class of 1998-99 (ECLS-K; National Center for Education Statistics, 2016). We used reading, math and science ability as the response variable and different child characteristics as partitioning variables (such as motor skills and gender). We investigated how time and clustering of observations can be accounted for in a RP-GCM. Specifically, we assessed the effect of two parameters in the recursive partitioning algorithm: I) Whether cluster- (C) or observation-level (O) parameter stability tests are employed for selecting partitioning variables and II) whether initializing model estimation with the random effects (R) or the tree structure (T) yields a more accurate and/or less complex model. The effects of the two parameters were assessed in both a random intercept (RI) and a random intercept and slope (RIS) model. In the RI models, CT and CR yielded higher predictive accuracy and lower complexity. In the RIS models OR performed best with highest predictive accuracy and lowest complexity. The best RI and RIS model yielded similar substantive results: The detected subgroups differed mostly in socioeconomic status, fine motor skills and race. Thus, these variables were the most important predictors of reading, math and science trajectories, but it should be noted that the subgroups differed more strongly in their baseline performance than in their growth rate over time. Although the results of the best RI and RIS models were comparable, the OR RIS model may be preferred over the CT/CR RI model, because of its lower *MSE* and tree size.

Table of contents

Introduction

Longitudinal data are commonly encountered within empirical research in psychology. This type of data is often gathered for assessing associations between covariates and the development of, for instance, problem behaviour (e.g. early-onset of conduct problems; Beauchaine, Webster-Stratton, & Reid, 2005), mental constructs (e.g. depression and self-esteem; Reddy, Rhodes, Mulhall, 2003) and quality of life (e.g. in Parkinson's disease patients; Jones, Marsiske, Okun, & Bowers, 2015). In these three examples, latent growth curve modelling (LGCM) was used to model the longitudinal nature of the data. LGCM is often used for assessing change or development of psychological constructs over time, because it allows for flexibly modelling the effect of time and covariates on psychological constructs (Duncan & Duncan, 2009). In LGCM it is possible to model individual growth trajectories by estimating person-specific intercept and slope values through a latent intercept and latent slope (Beaujean, 2014). In addition, within LGCMs it is possible to include complex relations such as mediation and moderation, multiple predictors and/or populations, and multilevel or hierarchical structures (Duncan & Duncan, 2009). As an example, in the study of Jones et al. (2015) LGCM analysis was used to examine the influence of different symptoms (predictors) on quality of life of Parkinson's disease patients over time. They found depression symptoms to be most strongly related to quality of life, and a smaller effect was found for motor symptoms, gender and age. Apathy was not found to be related to quality of life. Such findings can be used to gain insight into the determinants of quality of life of Parkinson's disease patients and/or improve care, through for example early targeting of relevant determinants of quality of life.

A disadvantage of LGCM is that the results may be hard to interpret and implement in practice, because LGCM yields model fit statistics and parameter values for different associations within the model, but these may be very abstract for someone who is not familiar with these models. Parameters that explain different aspects of change may be preferred as they are easier to interpret (Grimm, Ram, & Hamagami, 2011). Another disadvantage of LGCM is that no subgroups are identified. Taking the study of Jones et al. (2015) as an example, they found depression to be a predictor of quality of life, but for clinical decision making it may be more helpful to detect specific subgroups of patients which are at high risk of (strongly) deteriorating quality of life over time.

Tree-based methods (or recursive partitioning) specifically aim to detect such subgroups and in addition are easier to interpret (King & Resick, 2014), can cope with large numbers of covariates and can handle (complex) interactions automatically (Hajjem, Bellavance, & Larocque, 2014). Tree-based methods build a decision tree that consists of inner and terminal nodes by separating the observations into subgroups with more homogeneity within the subgroup and more heterogeneity between the subgroups with respect to the outcome value (James, Witten, Hastie, & Tibshirani, 2017). This is done by exhaustively searching the partitioning variables for a splitting value that minimizes heterogeneity of the outcome variable within each resulting subgroup (Zeileis, Hothorn, & Hornik, 2008). The data is

then split into two groups/nodes on that specific value of the partitioning variable. This process is repeated until a prespecified criterion is met (e.g. minimum node size; Stegmann, Jacobucci, Serang, & Grimm, 2018).

Like most statistical methods, tree-based methods often assume independence between measurements. This assumption is violated when data is clustered, for example when multiple observations are nested within the same person (as in longitudinal datasets), or when observations on students are nested within classes and/or schools. This nested structure introduces dependence between the lower-level observations within a higher-level unit, which can be taken into account through mixed-effects models (e.g., Sela & Simonoff, 2012). A new algorithm that builds on tree-based methods and is able to account for nested data structures is the generalized linear mixed models trees (GLMM trees) algorithm developed by Fokkema, Smits, Zeileis, Hothorn and Kelderman (in press). GLMM tree is a promising method as it has shown better performance than other tree-based methods and linear mixed-effects models in simulation studies (Fokkema et al., in press). The GLMM tree algorithm is implemented in the R (R Core Team, 2017) package glmertree (version 0.1-2; Fokkema & Zeileis, 2016).

The current study focusses on the use of GLMM tree for detecting subgroups in LGCMs. As it is currently unclear how to best specify and fit a LGCM-based recursive partition, further referred to as a recursive partitioning growth curve model (RP-GCM), we will evaluate the effects of several model fitting procedures and settings on an existing dataset of the Early Childhood Longitudinal Study-Kindergarten class of 1998-99 (ECLS-K; National Center for Education Statistics, 2016). It is currently unknown how to best fit such a RP-GCM with glmertree, which leads to the following research question: 'How can we fit a RP-GCM using glmertree in order to detect subgroups with different growth trajectories?'. Subsequently, when we know how to define a RP-GCM with glmertree, we are also interested in different settings within the GLMM tree algorithm, namely; whether observation- or cluster-level parameter stability tests, and whether initialization with the random effects or tree structure yields more accurate results.

The remainder of the Introduction is structured as follows: First, the GLMM tree algorithm will be explained, followed by an explanation of how dependence between observations can be taken into account within the algorithm. Next, we discuss the different parameter stability tests and initialization approaches, and the corresponding hypotheses in the current study. In the Methods and Results section, we will present an empirical evaluation of the performance of the GLMM-tree algorithm with different parameter settings in the aforementioned ECLS-K dataset.

The GLMM tree algorithm

The GLMM algorithm uses the generalized linear model (GLM) tree algorithm (Zeileis et al., 2008), which estimates a tree through the following steps:

Step 1: Fit a GLM to the observations in the current node, for example a model with a single intercept and a slope for time.

Step 2: Parameter stability tests are used to decide which partitioning variable (covariate) is most strongly associated with instability of the parameters of the global GLM fitted in step 1.

Step 3: A split is made using the variable most strongly associated with instability, using the splitting value that minimizes the loss function in both of the resulting subgroups (e.g., residual sum of squares or the negative log likelihood).

Step 4: Repeat steps 1-3 in the resulting subgroups until none of the partitioning variables yield a significant result of the parameter stability test or the subgroups become too small.

Note that the above steps fit a fixed-effects GLM in every node of the tree and thus do not take clustering into account. GLMM tree adds a mixed-effects model in order to account for the dependence between observations. In the GLMM-tree model, a random intercept and/or slope is estimated per cluster. As it is not possible to estimate the fixed- (GLM tree) and random-effects parts simultaneously, an estimation-minimization (EM)-type approach that iterates between estimating the random effects and the partition (tree structure) is used. In the default settings, GLMM tree starts with assuming the random effects to be 0, as the random effects are initially unknown. The algorithm then iterates between:

Step A: Given the current random effects, estimate the partition (tree).

Step B: Given the partition (tree), estimate the node-specific GLMs and the random effects.

The algorithm reaches convergence when the random effects no longer change between consecutive iterations. The predicted values for the observations in the terminal nodes are determined by the node-specific parameter estimates of the GLM, while adjusting for the (globally estimated) random effects (Fokkema et al., in press).

GLMM tree settings

Within GLMM tree it is possible to adjust the initialization approach of the algorithm. Instead of fixing the random-effects parameters to 0 and initializing model estimation with the tree structure, we can assume that there are no subgroups (or partitions in the tree) and start with step B, first estimating the random effects instead. The difference between the two methods is that when you start model estimation with the tree structure, the information that is accounted for by the tree cannot be accounted for by the random effects and vice versa. In the current study, we have no general expectation for which method will yield better results, as we will be using a real dataset. If there are substantial cluster-specific effects in the dataset, initializing the model estimation with the random effects (step B) may likely yield more

accurate results than initializing with estimating the tree structure (step A). The hypothesis for this setting is:

Hypothesis 1: Initializing model estimation with the random effects yields more accurate results than initializing with the tree structure.

The clustering structure can also be taken into account in the parameter stability tests, employed in Step 2 of the GLM tree algorithm, to select splitting variables. The null hypothesis of the parameter stability test is that the parameters of the node specific GLM are stable with respect to a given partitioning variable. This hypothesis is rejected when the observed values of the outcome variable deviate systematically from the predicted values of the node-specific GLM, with respect to the partitioning variable. The extent to which the predicted values systematically deviate from the observed values with respect to the partitioning variables is quantified by the *p*-value of the parameter stability test. A more in-depth explanation of the parameter stability test is beyond the scope of this thesis, we refer to Zeileis et al. (2008) for more in-depth information. Our main interest here is in two types of parameter stability tests; observation- and cluster-level parameter stability tests. The observation-level parameter stability test tests for systematic deviations based on all observations, without taking clustering into account. In contrast, the cluster-level parameter stability test accounts for the clustering. Not taking clustering into account is likely to artificially increase the Type-I error: the probability that the null hypothesis of the parameter stability test is falsely rejected. Using cluster-level parameter stability tests likely yields lower Type-I errors. We therefore expect that cluster-level stability tests will yield more accurate results in RP-GCM. The second hypothesis is with regard to the initialization approach:

Hypothesis 2: Cluster-level parameter stability tests yield more accurate results than observation-level parameter stability tests.

In the Method section we will explain how we will answer the main research question 'How can we fit a RP-GCM using glmertree in order to detect subgroups with different growth trajectories?' and how we will investigate the two hypotheses. In the Result section the results will be discussed and in the Discussion section we will summarize and interpret the results, discuss limitations and suggestions for future research.

Methods

Dataset

In this study, we used empirical data from the Early Childhood Longitudinal Study-Kindergarten class of 1998-99 (ECLS-K; National Center for Education Statistics, 2016). Data was collected from fall kindergarten 1998 through eighth grade 2007 on seven time points. The goal of the study was to examine child development, school readiness and early school experiences. In this study, we focussed exclusively on the measurements that took place during spring of kindergarten, first, third, fifth and eighth grade, resulting in five measurement occasions. Data of 21,304 children were collected from schools all across the United States of America. A total of 1,018 different schools were included in the study. On average the children were 6 years and 5 months old at baseline (ranging from 4 years and 5 months to 8 years old). Little over half of the children are males (51.1%). Most of the children are white, non-Hispanic (55.2%), 15.1% is black or African American non-Hispanic, 17.9% is Hispanic, 6.4% is Asian and 5.4% has a different background. For 95.3% of the children it was the first time they attended kindergarten.

Stegmann et al. (2018) used the same data for their research on recursive partitioning of longitudinal data. We used their model as a starting point for our RP-GCM. Stegmann et al. (2018) studied the change in reading ability (both language and literacy) of the children over time from spring of kindergarten to fall of eight grade. Reading ability was measured using direct cognitive assessment. The assessment consisted of a total of 72 items, but the children only received 20 items. Items were selected based on their reading ability indicated by a routing test. In our study, theta (ability) scores of reading were used. These scores have a mean of 0 and standard deviation of 1 (National Center for Education Statistics, 2016).

Besides reading ability, math and science ability were also used as the response variable. Fitting RP-GCMs on multiple response variables enabled us to compare the results and give stronger conclusions on the research questions.

In the Stegmann et al. (2018) model, possible partitioning variables for the reading trajectories were gender, race, socioeconomic status, gross motor skills, fine motor skills, interpersonal skills, self-control, whether it was the first time in kindergarten, internalizing and externalizing problem behaviour. We decided to add age at baseline as a predictor, because we expected this variable to be of relevance. Descriptive statistics of the partitioning variables are given in Table 1 and coding for all partitioning variables can be found in Appendix A.

Fitting the RP-GCM requires that there are no missing values in the data. We used listwise deletion to remove children with missing values on the previously mentioned variables. Reading and math were measured on all five occasions, but science was only measured three times (third, fifth and eighth grade). There is attrition present in the data, so not all children were measured on all occasions. The advantage of using only the children with all measurements is that each child has a similar kind of

curve and the RP-GCM tree will not include splits based on attrition (e.g. because they moved away). A disadvantage is advantaged subgroup selection, because the children who were measured at all occasions might differ from children who did stay in the study, but our main goal in this study was not to find a model that is generalizable to the whole population, but to assess the performance of the GLMM tree in fitting RP-GCM models. Because the datasets are sufficiently large, we decided to include only those children with measurements on all measurement occasions. Data of N = 6,277 children were used in the reading dataset, N = 6,512 children in the math dataset and N = 6,625 children in the science dataset.

Table 1. *Descriptive statistics for partitioning variables (covariates) used in our model.*

Categorical variables	*Variable name*	*Category*	*%*
Gender	GENDER	Male	51.1
		Female	48.9
Race	RACE	White, non-Hispanic	55.2
		Black or African American, non-Hispanic	15.1
		Hispanic, race specified	8.6
		Hispanic, race not specified	9.3
		Asian	6.4
		Native Hawaiian, other pacific islander	1.0
		American Indian or Alaska native	1.8
		More than one race, non-Hispanic	2.6
First time in kindergarten	P1FIRKDG	Yes	95.3
		No	4.7

Continuous variables	*Variable name*	*M*	*SD*	Range
Socioeconomic status	WKSESL	.09	.79	-5 – 3
Gross motor skills	C1GMOTOR	6.41	1.81	0 – 8
Fine motor skills	C1FMOTOR	5.98	1.99	0 – 9
Interpersonal skills	T1INTERP	3.02	.62	1 – 4
Self-control	T1CONTRO	3.12	.60	1 – 4
Internalizing problem behaviour	T1INTERN	1.51	.51	1 – 4
Externalizing problem behaviour	T1EXTERN	1.59	.62	1 – 4
Age at baseline (in months)	AGEBASELINE	73.62	4.33	53 – 96

Model specification procedures

In the first stage of this research, we explored how to best fit a RP-GCM using the GLMM tree algorithm as implemented in the R package glmertree (Fokkema & Zeileis, 2016). As a starting point, we aimed to replicate the model from Stegmann et al. (2018). Glmertree requires the model to be specified through the formula argument. The first step in the current project was to find the right model specification formula; that is, how to specify the growth curves, how to account for the longitudinal nature of the data, and how to specify potential partitioning (covariates) variables in the formula. The formula consists of

the response variable and a three-part right-hand side describing the regressors, random effects and partitioning variables (Fokkema & Zeileis, 2016). For example, in a longitudinal model, the formula may read:

$$y \sim t \mid (1 \mid z) \mid x_1 + \cdots + x_p$$

where y represents the response variable; t represents the regressor (e.g. time); $y \sim t$ represents the node-specific regression model; $(1 \mid z)$ represents the random effects of, in this case, a random intercept regressed on cluster indicator z; and $x_1 + \cdots + x_p$ represents the partitioning variables. We explored how to take into account the longitudinal nature of the data, how the random effects needed to be specified and how partitioning variables could be added in our RP-GCM.

Hypotheses

RP-GCMs were fit with the glmertree package (version 0.1-2; Fokkema & Zeileis, 2016) using R (R Core Team, 2017). A two-factor design (see Table 2) was used to assess the hypotheses on the two initialization approaches and two types of parameter stability tests. The default settings of glmertree are to estimate the tree structure first and to employ observation-level parameter stability tests. Estimating the random effects first is accomplished by specifying 'ranefstart = TRUE'. Cluster-level parameter stability tests can be employed by specifying 'cluster = CHILDID', where CHILDID is het cluster indicator in the current study.

Table 2. *Two-factor design. The capitals indicate whether model estimation is initialized by estimating the tree structure (T) or random effects (R) and whether observation- (O) or cluster-level (C) parameter stability tests are used.*

	Model estimation initialization	
	Tree structure	Random effects
Parameter stability test		
Observation-level	OT	OR
Cluster-level	CT	CR

Performance

The performance of the different models were assessed by means of predictive accuracy, and interpretability of the fitted trees. The predictive accuracy of the models was measured by calculating the mean-squared error (*MSE*) between observed and predicted reading scores using 10-fold cross-validation. For cross-validation, we employed cluster-level (i.e. child-level) sampling instead of

observation-level sampling. That is, measurements of a single child were either in the training or test dataset. The *MSE* between the models could be compared, lower values indicating better performance. Standard errors (*SE*) were calculated to assess if differences are significant or not. Tree depth, length and random-effect variances for the trees of the 10-fold cross-validation were also calculated. Means and standard deviations of tree depth, length and random-effect variances were inspected to assess complexity and interpretability of the fitted models. These performance statistics were used to assess the hypotheses.

The models were also fitted on the complete datasets and plotted to assess interpretability of the trees between models. The maximum depth for these GLMM trees was constrained to four, to aid interpretability. Substantial interpretations of our model on the complete dataset were compared to the results in Stegmann et al. (2018).

Results

Model specification

Different aspects of the model needed to be specified in order to be able to fit a RP-GCM. Below is discussed how we took into account the longitudinal nature of the data and how random effects and partitioning variables were specified.

Node-specific model

To account for the longitudinal nature of the data, reading ability was regressed on time. We chose to use the number of months passed after the first measurement occasion as the timing variable.

Initially, reading trajectories were assumed to be linear over time, but after plotting the trajectories, the increasing curve of the relation between months passed and reading ability seemed to flatten over time (see Figure 1, panel A). The goal of this study is to be able to fit a RP-GCM with GLMM tree; that is, model-based recursive partitioning based on (generalized) linear mixed model. The terminal nodes of the trees therefore consist of a (generalized) linear model (Fokkema et al., in press). For this reason we needed to transform the timing metric for the association with the response to become approximately linear. When the increase flattens over time a log transformation is often used to account for the non-linearity (e.g., Long & Ryoo, 2010). We expected the data to show an approximately linear trend after this transformation, but as shown in Figure 1, panel B, the increase in reading ability now increased over time. We therefore applied a power function to produce an approximately linear trend in the trajectories. A square root was used to obtain an approximately linear trend, see Figure 1, panel C.

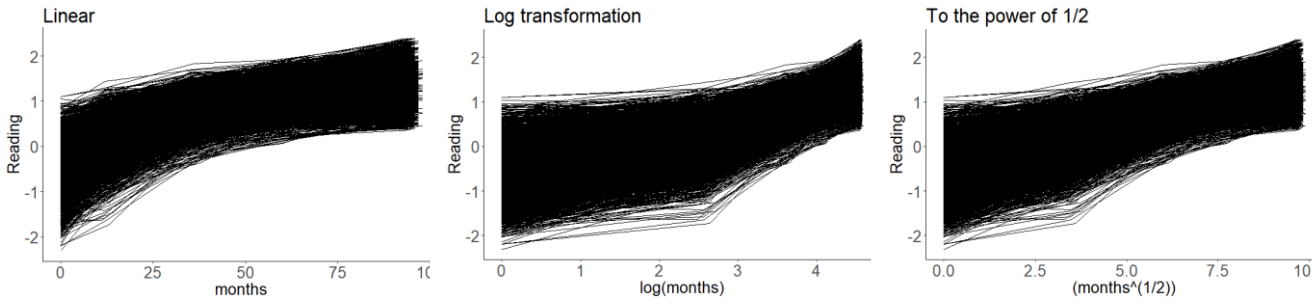
As with reading, linearity of the trajectories of math and science ability were checked. For math, the same transformation as with reading resulted in an approximately linear trend, and for science, months to the power of ⅔ was needed. In Appendix B plots for math (Figure B1) and science (Figure B2) are available.

Specifying random effects

We transposed the data from wide to long form so that each row would contain the variables for one measurement occasion for one child. Because there are several measurement occasions for one child, the rows are dependent. To account for this clustered nature, random effects were specified. A random intercept with respect to the indicator for child was specified, allowing for a different intercept value (i.e., baseline reading ability) for every child.

We also considered whether we needed to specify a child-specific random effect of time, so that a slope for reading ability is estimated per child. As we are interested in finding subgroups of children who differ in terms of their growth in reading ability over time, specifying such a random effect may on

one hand obfuscate the effects of interest, because the differential effects of time between groups of children will possibly be accounted for by the random effects and consequently not picked up by the tree, but on the other hand may provide more accurate results. We decided to both fit a child-specific random intercept (RI) model and a child-specific random intercept and slope (RIS) model to compare the results on all three datasets.



A. B. C.

Figure 1. *Different growth curves of time versus reading ability. Panel A shows the trajectories without transformations. Panel B shows a natural log transformation. Panel C shows the trajectories with a square root transformation.*

Partitioning variables

The partitioning variables gender, race, socioeconomic status, gross motor skills, fine motor skills, interpersonal skills, self-control, whether it was the first time attending kindergarten, internalizing and externalizing problem behaviour, and age at baseline were specified as potential partitioning variables. Below we will first discuss the results of the 10-fold cross-validation for RI and then for RIS models. Then we will fit the best fitting RI and RIS models on the complete datasets and interpret the resulting trees.

Random intercept models

Reading ability

In Table 3 model performance statistics of RI models are given in terms of predictive accuracy, tree size and variance of random intercepts. The largest differences in *MSE* are visible between using cluster- and observation-level parameter stability tests: Model CT and CR (for both *MSE* = .135) outperform model OT and OR (for both *MSE* = .162). The difference in *MSE* values is .027 and the standard errors (for all *SE* = .002) show that this difference is statistically significant. Note that we found no differences between initializing model estimation with the tree structure or random effects.

A similar pattern can be seen when the average number of nodes of the trees is assessed: models using observation-level parameter stability tests have substantially more nodes (*number of nodes* = 314.2) than the models in which cluster-level parameter stability tests were used (*number of nodes* = 104.8). Inspecting the trees fitted on the complete dataset showed that models OT and OR, and CT and CR are identical. The depth of CR is 9 and 11 for OR, resulting in a much larger tree with many more splits. Splits on the first four levels are similar in both models, indicating that both models identify similar partitioning variables to be of high importance for identifying subgroups within the data. Again, there are no differences between tree-structure and random-effects initialization. Thus, in this dataset, using cluster-level parameter stability tests result in a lower number of subgroups and higher predictive accuracy. The variance of the random intercept is slightly larger when cluster-level stability tests are used than when observation-level stability tests are used (s^2 = .048 vs. s^2 = .042, respectively).

Math ability

As with reading, in the math ability dataset differences in model performance are present between models fitted with cluster- and observation-level parameter stability tests (see Table 3 for performance statistics). Models CT and CR (both *MSE* = .193) outperform models OT and OR (both *MSE* = .221). No differences are visible between using tree-structure or random-effect initialization. The tree size is almost halved when cluster-level parameter stability tests (*number of nodes* = 307.0) compared to when observation-level parameter stability tests are used (*number of nodes* = 574.6), which indicates that substantially more spurious splits occur when observation-level tests are used. Again, the variance of the random effects are larger when cluster-level parameter stability tests are used (s^2 = .068 vs. s^2 = .057, respectively).

Science ability

As with reading and math, the performance statistics for models (in Table 3) of the science ability dataset show cluster-level parameter stability tests outperform models in which observation-level stability tests were used. In contrast to the findings of reading and math, the performance of the models do differ between the different initialization approaches. The best performing model is CR (*MSE* = .405), with CT performing only slightly worse (*MSE* = .406). The tree sizes differ slightly (*number of nodes*$_{CR}$ = 98.4 and *number of nodes*$_{CT}$ = 99.0), but the variances of random effects are similar (s^2 = .207). Both OT and OR perform statistically significantly worse (*MSE*$_{OT}$ = .505 and *MSE*$_{OR}$ = .459), yield substantially larger trees (*number of nodes*$_{\mathrm{OT}}$ = 236.4, *number of nodes*$_{\mathrm{OR}}$ = 148.4) and have smaller variances of random effects (s^2_{OT} = .187, s^2_{OR} = .197).

Table 3. *Performance statistics of RI models for reading, math and science ability with observation-/cluster-level parameter stability tests and initialization of the tree structure or random effects. Bold statistics indicate the best performing model. The sample variance (s^2) of reading ability is .66, s^2 = .73 for math and s^2 = .89 for science.*

Model	Dataset	*MSE* (*SE*)	Mean number of nodes[a] (*SD*)	Variance of random intercept[a] (*SE*)
Observation-level tests				
Tree initialization (OT)	Reading	.162 (.002)	314.2 (32.73)	.042 (< .001)
	Math	.221 (.002)	574.6 (39.64)	.057 (< .001)
	Science	.505 (.005)	236.4 (12.93)	.187 (< .001)
Random-effects initialization (OR)	Reading	.162 (.002)	314.2 (32.73)	.042 (< .001)
	Math	.221 (.002)	574.6 (39.64)	.057 (< .001)
	Science	.459 (.005)	148.4 (5.97)	.197 (< .001)
Cluster-level tests				
Tree initialization (CT)	Reading	**.135 (.002)**	**104.8 (16.26)**	**.048 (< .001)**
	Math	**.193 (.002)**	**307.0 (45.57)**	**.068 (< .001)**
	Science	.406 (.005)	99.0 (15.29)	.207 (< .001)
Random-effects initialization (CR)	Reading	**.135 (.002)**	**104.8 (16.26)**	**.048 (< .001)**
	Math	**.193 (.002)**	**307.0 (45.57)**	**.068 (< .001)**
	Science	**.405 (.005)**	**98.4 (16.63)**	**.207 (< .001)**

[a] Average over 10-fold cross-validation

Random intercept and slope models

Reading ability

In Table 4 model performance statistics of RIS models are given in terms of predictive accuracy, tree size, variance of random intercepts and variance of random slopes. The best performing model in terms of *MSE* is model OR (*MSE* = .117). Model CT (*MSE* = .137) performed slightly better than model CR (*MSE* = .148). Both CT and CR performed significantly worse than model OR. Model OT performs worst with an *MSE* of .162. A possible explanation for this is that differences between children in growth over time are already accounted for by estimating a slope for every child, which might explain why observation-level outperform cluster-level parameter stability tests. Initializing model estimation with random effects might ensure that the differences between children in the effect of time are accounted

for by the random slopes, yielding a lower number of subgroups to be found by the tree, resulting in more accurate results.

A similar pattern is present for the average number of nodes in the trees: better performing models have substantially less nodes. Model OR has the smallest tree with 21.2 nodes on average. The difference in number of nodes is substantial between OR, and CT (*number of nodes* = 105.2) and CR (*number of nodes* = 252.2). The tree of model OT is largest with 336.2 nodes on average. In this dataset, using observation-level parameter stability tests and initializing model estimation with the random effects results in less subgroups model and higher predictive accuracy.

Inspecting the variances of the random effects shows that a low *MSE* value corresponds with a larger variance both for the random intercept and slope, but the differences are small; e.g. s^2_{OR} = .124 vs. s^2_{OT} = .101 for the random intercept and s^2_{OR} = .001 vs. s^2_{OT} < .001 for the random slope.

Math ability

As with reading ability, model OR outperforms the other models based on both *MSE* (= .130) and tree size (*number of nodes* = 28.6) in the math ability dataset. Model OT performs worst with largest *MSE* values (= .227) and highest tree size (*number of nodes* = 617.6). Models CT and CR perform significantly worse than OR, but better than OT. There is a slight difference in *MSE* between model CT (*MSE* = .182) and CR (*MSE* = .178). Initializing model estimation with the random effects yields in this dataset a slightly better result when cluster-level parameter stability tests are used and a large difference when observation-level parameter stability tests are used. Again, as with reading, model CT (*number of nodes* = 237.8) is substantially smaller than model CR (*number of nodes* = 452.4) and if *MSE* is smaller the variance of the random effect is slightly larger than when *MSE* is larger; e.g. s^2_{OR} = .097 vs. s^2_{OT} = .069 for the random intercept and all random slope variances are < .001.

Science ability

Similar to the pattern of reading and math ability, model OR outperforms the other models (*MSE* = .382, *number of nodes* = 47.8) and model OT performs worst (*MSE* = .500, *number of nodes* = 236.2). Model CT (*MSE* = .414, *number of nodes* = 97.8) performs significantly better than model CR (*MSE* = .470, *number of nodes* = 259.8). Cluster-level stability tests yield better model performance than OT, but worse than OR. Variances of the random effects are smaller when *MSE* values are larger and vice versa; e.g. s^2_{OR} = .162 vs. s^2_{OT} = .133 for the random intercept and all random slope variances are < .001.

Table 4. *Performance statistics of RIS models for reading, math and science ability with observation-/cluster-level parameter stability tests and initialization of the tree structure or random effects. Bold statistics indicate the best performing model. The sample variance (s^2) of reading ability is .66, s^2 = .73 for math and s^2 = .89 for science.*

Model	Dataset	*MSE* (*SE*)	Mean number of nodes[a] (*SD*)	Variance of random intercept[a] (*SE*)	Variance of random slope[a] (*SE*)
Observation-level tests					
Tree initialization (OT)	Reading	.162 (.002)	336.2 (23.99)	.101 (< .001)	< .001 (< .001)
	Math	.227 (.003)	617.6 (51.22)	.069 (< .001)	< .001 (< .001)
	Science	.500 (.005)	236.2 (13.47)	.133 (< .001)	< .001 (< .001)
Random-effects initialization (OR)	Reading	**.117 (.002)**	**21.2 (2.74)**	**.124 (< .001)**	**.001 (< .001)**
	Math	**.130 (.002)**	**28.6 (4.20)**	**.097 (< .001)**	**< .001 (< .001)**
	Science	**.382 (.005)**	**47.8 (7.07)**	**.162 (< .001)**	**< .001 (< .001)**
Cluster-level tests					
Tree initialization (CT)	Reading	.137 (.002)	105.2 (10.00)	.112 (< .001)	< .001 (< .001)
	Math	.182 (.002)	237.8 (30.51)	.078 (< .001)	< .001 (< .001)
	Science	.414 (.005)	97.8 (17.31)	.149 (< .001)	< .001 (< .001)
Random-effects initialization (CR)	Reading	.148 (.002)	252.2 (31.09)	.108 (< .001)	< .001 (< .001)
	Math	.178 (.002)	452.4 (36.06)	.079 (< .001)	< .001 (< .001)
	Science	.470 (.005)	259.8 (14.30)	.139 (< .001)	< .001 (< .001)

[a] Average over 10-fold cross-validation

Interpretation of the best fitting model

The best fitting RI and RIS models were specified as mentioned before, but now on the complete dataset instead of using 10-fold cross-validation. The only difference in the RI and RIS models is that in the RIS model a child-specific random slope is added. In Table 5 statistics are given for the full trees.

In the RI models fitted on the complete datasets of reading, math and science ability, the trees of CT and CR are identical. For the science dataset this is in contrast to what is expected, because models CT and CR had slightly different performance levels. A possible explanation is that RP-GCM yields slightly different models in the 10-fold cross-validation, because only 90% of the data is used in each fold. For the interpretation of the best fitting model we used 100% of the data, which means that there

is more information available for the analysis to find the 'correct' model. Because models CT and CR are identical we will only show results of model CR, but these results also apply to model CT.

The number of nodes in the science ability tree is slightly larger than the average tree of the 10-fold cross-validation. This is because we now fit the model on 100% of the data instead of 90% of the data in the cross-validation. With a larger sample, significant splits are more easily accomplished, resulting in more splits. Remarkably, the tree for math ability is much larger when fitted on the full dataset than when 10-fold cross-validation was used. This can be explained by the larger *SD* of the number of nodes in the math ability dataset. The number of nodes in the reading ability tree is slightly smaller than when 10-fold cross-validation was used.

The RIS model tree sizes of reading and math ability also slightly increased as we fit the models on 100% of the data. Science ability, on the contrary, has less nodes in the tree than the average model of the 10-fold cross-validation.

Table 5. *Tree characteristics for the RP-GCM fitted on the full datasets of reading, math and science ability for best fitting RI (CR) and RIS (OR) models.*

	RI model			**RIS model**		
Tree characteristics	Reading	Math	Science	Reading	Math	Science
Number of nodes	101	435	101	25	29	47
Average group size	123.08	29.87	129.90	482.85	434.13	276.04
Range group sizes	5 – 428	4 – 282	9 – 430	45 – 1247	20 - 1145	13 – 866
First partitioning variable	WKSESL	WKSESL	RACE	WKSESL	WKSESL	RACE
Times used as partitioning variable:						
GENDER	4	14	7	--	1	3
RACE	6	16	4	4	2	4
WKSESL	12	32	13	4	1	4
C1GMOTOR	1	23	1	--	--	1
C1FMOTOR	12	22	10	2	3	4
T1INTERN	2	17	2	--	--	--
T1EXTERN	2	13	--	--	--	--
T1INTERP	7	16	2	--	--	1
T1CONTRO	--	18	2	--	--	--
P1FIRKDG	1	--	2	1	--	1
AGEBASELINE	3	46	7	1	7	5

As expected, average group size in the terminal nodes is larger when the number of nodes is smaller in both RI and RIS models. All reading and math ability models have socioeconomic status as the first partitioning variable. For science race is the first partitioning variable. This indicates that socioeconomic status is an important variable in identifying subgroups for reading and math trajectories and race is important for science ability trajectories. Race, fine motor skills and age at baseline also seem to be important partitioning variables, although the latter not as much as for reading ability. Whether it was the first time attending kindergarten does not seem to be an important variable, because it is not often picked up by the RP-GCMs. Below we will interpret the best fitting RI and RIS model on the reading ability dataset. Best fitting trees for math and science ability can be found in Appendices C and D (C contains figures for the restricted models and D for full models).

RI model

In Figure 2 the best performing tree on the reading ability data is plotted (model CR) with maximum tree depth of four. Adding this restriction adjusts the estimation process, which results in different splits than when we do not add this restriction. Inspecting the restricted (see Appendix C) and full (see Appendix D) trees shows the splits are similar in the first four levels of most of the trees. Only the restricted math tree in Figure C1 to the full tree in Figure D2 shows that in node 13 of the restricted tree age at baseline is picked as a partitioning variable instead of socioeconomic status in the full tree. Because the trees are similar, we decided to restrict tree size to aid in interpretation of the fitted models. We will focus on the trees from the reading dataset, but the trees of math and science can be interpreted in the same way.

Again, only the CR model is plotted, because the CT model is identical and using cluster-level parameter stability tests outperform models in which observation-level parameter stability tests were used. Socioeconomic status is the first splitting variable in the tree. In the tree, the children are divided on a socioeconomic status of around .23. Further splits on the second level are made on fine motor skills. On the third level, splits are made on race and socioeconomic status.

The main differences between the subgroups obtained from the analyses are differences in intercept values. The slope values are very similar in all subgroups/terminal nodes. Reading ability increases at the same rate in all subgroups within the tree. This means that splits are mostly driven by differences in intercept values. Children with lower socioeconomic status and worse fine motor skills seem to have lower intercept values, thus start off with lower reading ability levels, than those with a higher socioeconomic status and better fine motor skills. Children who are white (non-Hispanic), Hispanic (race specified), Asian or who have more than one race (non-Hispanic) have higher intercept values.

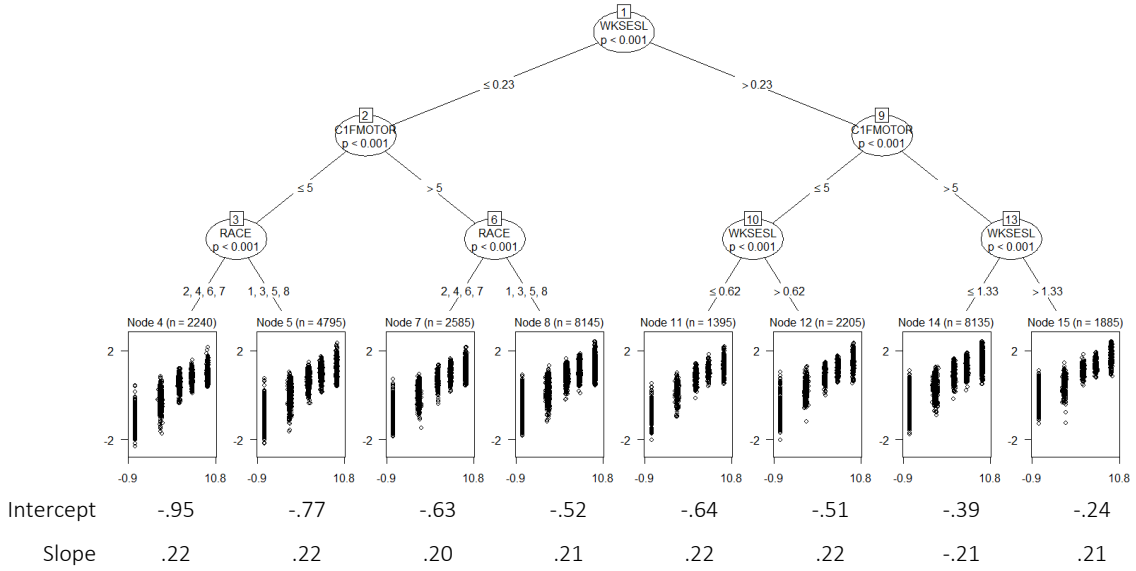


Figure 2. *Plot of restricted reading ability tree from the RI model CR on full dataset with maximum tree depth of four. Node-specific intercept and slope estimates are given below the tree.*

RIS model

Again, we restricted the maximum depth of the tree to four, to aid interpretation. In Figure 3 the tree of RIS model OR is depicted of the reading dataset. The top of the tree is on many aspects quite similar to the CT RI model tree, as both trees contain socioeconomic status as the first splitting variable, fine motor skills as second-level splitting variables and socioeconomic status on the third level. Although different settings were used in fitting the models, this part of the tree yields similar results. Differences are visible on the third level of the tree. The RIS model has one less split, the value on the splitting variable socioeconomic status in node 11 is lower than the value in node 13 of the RI model and race is not present as a partitioning variable (in the restricted tree).

As in the RI model, splits are mostly driven by differences in intercept values. The slope values are similar between the subgroups. Higher socioeconomic status, better fine motor skills and being white (non-Hispanic), Hispanic (race specified), Asian or having more than one race (non-Hispanic) are important predictors for starting off with higher reading ability levels.

Restricted and full trees for the RIS models of the math and science datasets can be found in Appendix C and D, respectively.

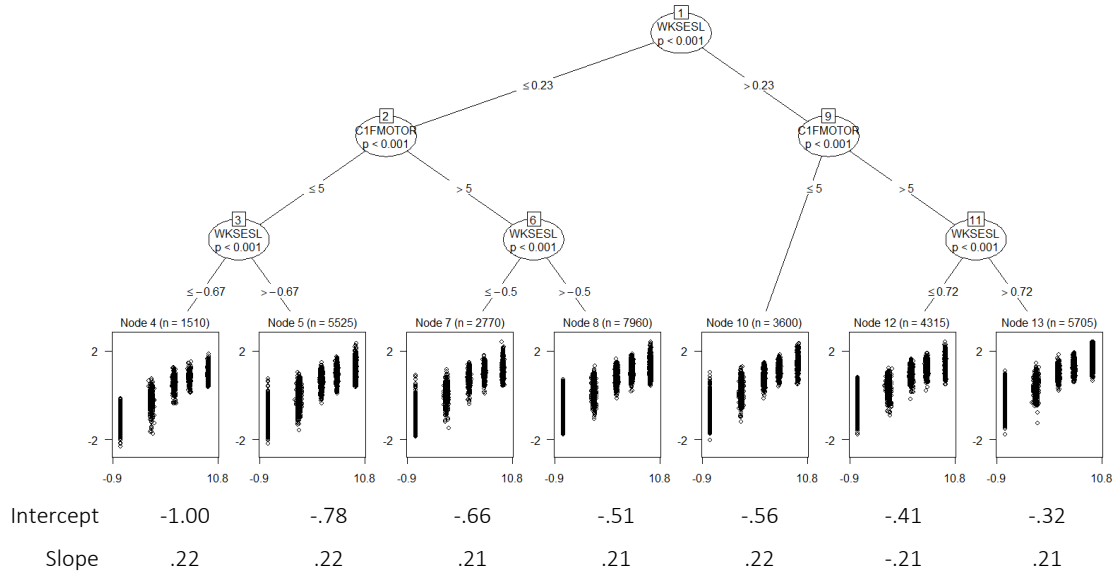


Figure 3. *Plot of restricted reading ability tree from the RIS model OR on full dataset with maximum tree depth of four. Node-specific intercept and slope estimates are given below the tree.*

Hypotheses

In the case of the RI model our results do not support the hypothesis that initializing model estimation with the random effects yields more accurate results, because no (or neglectable) differences between the initializing approaches were found. Our results confirm the hypothesis that cluster-level parameter stability tests yield more accurate results, as the model statistics showed more accurate results than the models with observation-level parameter stability tests.

In contrast to the RI model, in the RIS model the results confirm the hypothesis that initializing model estimation with the random effects yields more accurate results when observation-level parameter stability results are used. Largest significant differences are present between model OT and OR. We can reject the hypothesis that cluster-level parameter stability tests yield more accurate results, because the best performing RIS model was OR, which performed significantly better than the other models, thus in the RIS model observation-level parameter stability tests yield more accurate results.

Discussion

The goal of our study was to determine how to best fit a RP-GCM with glmertree on longitudinal data, in order to detect subgroups with different growth trajectories. To answer this question a large dataset of the ECLS-K study was used to model reading, math and science ability trajectories for children in kindergarten through eighth grade.

GLMM trees require several different aspects to be specified in the model formula: the subgroup-specific GLM (i.e., the response variable and regressors), the random effects and partitioning variables (Fokkema et al, in press). In the subgroup-specific GLM, the response variables were regressed on the number of months passed after the first measurement occasion. Inspecting the association between months passed and reading ability appeared to be non-linear. The timing metric needed to be transformed with a power function in order to obtain an approximate linear association between months passed and each of the responses (reading, math and science ability).

Next, we investigated how the random effects needed to be specified. To account for the dependency of observations within the same child, both a model with child-specific random intercept (RI) and a model with a child-specific random intercept and slope (RIS) were specified. Although fitting a child-specific random slope can possibly obfuscate the effects of interest, it may also improve model accuracy.

Lastly, only time-invariant partitioning variables can be used when partitioning growth curve models. In our study, the partitioning variables comprised gender, race, socioeconomic status, gross motor skills, fine motor skills, interpersonal skills, self-control, whether it was the first time attending kindergarten, internalizing and externalizing problem behaviour, and age at baseline.

We investigated which setting of two aspects of the GLM algorithm yielded more accurate results in both the RI and RIS model: initialization of model estimation (random effects vs. tree structure) and parameter stability tests (cluster- vs. observation-level).

RI models using cluster-level parameter stability tests (C) performed better than models in which observation-level parameter stability tests (O) were used. Their predictive accuracy is higher, tree size is lower and the variance of the random intercepts is slightly larger. These variances are possibly larger when the cluster-level parameter stability tests are used because clustering is taken into account during recursive partitioning, reducing the power to detect splits, so effects that can be accounted for by the tree or the random effects will be accounted for by the random effects.

No differences were found between initializing model estimation with the tree structure (T) or random effects (R) in the reading and math dataset. Only small differences were found in the science dataset. These differences might be due to the lower number of measurement occasions for the science abilities. In the CT and CR models the tree structures are identical for all three datasets. Thus, using a

different initialization process for estimating the model seems to have little to no effect on the results for RI models.

In contrast to the RI model, we found a model with observation-level stability tests and model initialization with the random effects (OR model) to perform best in all three datasets in the RIS model. The accuracy of the OR model was higher, tree size lower and variances of the random intercepts slightly larger. A possible explanation for the different findings between the RI and RIS models is that the random effects in the RIS model already account for a large part of variation between clusters (children). Thus, even when the power for detecting splits is (too) high, as is likely the case with observation-level parameter stability tests in clustered data, the splits will not be detected when the differences between clusters are accounted for by the random effects. This also explains why in the RIS models, the initialization approach did affect performance.

Although different settings were used for the best fitting RI and RIS models, the substantial conclusions are very similar. Both models found socioeconomic status, fine motor skills and race to be important in predicting subgroup-specific trajectories in all three dataset. Remarkably, age at baseline was highly important only in the math dataset. Inspecting the subgroup-specific parameters in the terminal nodes of the best fitting models, showed that the partitions are mostly driven by differences in intercept values between the subgroups. Only small (neglectable) differences in slopes were observed.

In our study, the RIS model seems preferable over the RI model, because predictive accuracy is higher and tree size is substantially lower, making the tree easier to interpret and apply in practice.

Substantial interpretation

The results indicate that the children develop reading, math and science abilities at a similar rate across subgroups, but they differ in the ability level at which they start in kindergarten. Thus, if a child is a poorer reader in kindergarten, he/she will be a poorer reading in eighth grade, vice versa. The pre-existing differences between children may persist over time. The most important predictors for initial ability level are socioeconomic status, fine motor skills, race in the three datasets and for math also age at baseline.

A possible explanation for why socioeconomic status is an important predictor, is that children from families with a lower socioeconomic status may not have been read to as much as children from high socioeconomic backgrounds. As a result they may have less knowledge of stories, making it harder to learn to read (Juel, 1988).

Fine motor skills are important for, for instance, controlling eye movement. Having worse fine motor skills, might make it harder for the children to read and make cognitive learning more difficult (Grissmer, Grimm, Aiyer, Murrah, & Steele, 2010).

We found children of approximately 6 years and older at baseline to have a higher math ability score at baseline than younger children. Math ability is related to executive functions of the brain such

as inhibition, shifting and working memory (Cragg & Gilmore, 2014). Executive functioning increases throughout childhood and adolescence (Blakemore & Choudhury, 2006), but significant improvements are found to occur between ages 5 and 14. More specifically, the important executive functioning for math show significant improvements between ages 5 and 8 for inhibition, 5 and 6 for shifting and 4 and 15 for working memory (Best, Miller, & Jones, 2009). This may explain why older children have higher math ability scores.

With our model we know that when a child comes from a low socioeconomic background and has less developed fine motor skills compared to the other children they will likely perform worse on reading, math and science. Specific interventions could be implemented to minimize the effects of these variables. A child could for instance receive additional training in developing fine motor skills, in order to catch up with its peers.

Comparison with non-linear longitudinal recursive partitioning of Stegmann et al. (2018)

To build our model, we used the model from Stegmann et al. (2018) as a starting point. We fitted a similar model, but additionally included age at baseline as a partitioning variable and used months passed since baseline as a timing metric instead of the age of the child. Stegmann et al. (2018) used their non-linear longitudinal recursive partitioning (nLRP) method to analyse the data. The difference with glmertree is that this method estimates a non-linear mixed models in each terminal node, while GLMM tree fits generalized linear mixed models, with fixed effects estimated in each terminal node and random effects estimated globally. With nLRP, Stegmann et al. (2018) were able to model an exponential component in the growth curve model to account for the non-linear reading trajectories. In contrast, we transformed the data with a power function before the analysis.

The goal of Stegmann et al. (2018) was similar to ours. They investigated whether they could identify groups of individuals with similar growth trajectories and find predictors of these growth trajectories. Their final model consisted of three splits on fine motor skills at 5.5. Children with a higher or equal score of 5.5 were further subdivided on race. White or non-Hispanic children were then split on gender. Similar to these results, we found fine motor skills and race of high importance in predicting growth trajectories of reading ability. On the contrary, in our RI and RIS models for reading scores we found gender not to be a partitioning variable of high importance, but instead socioeconomic status was an important predictor for growth trajectories.

Stegmann et al. (2018) found their high fine motor skill, white or non-Hispanic and female group to have a higher reading scores at baseline than the other groups. No clear differences were found in the rate of increase in all terminal nodes. This is in agreement with our results. We found the splits to be mainly driven by differences in the random intercept and found no or negligible differences in the

random slope. The initial reading scores in the other terminal nodes in the Stegmann et al. (2018) did not differ between terminal nodes.

The current and Stegmann's et al. (2018) studies differ in the tree size obtained from the analyses and it is not clear which model has better accuracy as the nLRP algorithm is implemented in an R package that does not allow for generating predictions from the fitted model. Thus, predictive accuracy of nLRP models cannot be evaluated, nor compared with that of glmertree. The differences possibly arise from using different samples from the entire dataset. Stegmann et al. (2018) used a subsample of 591 children. Our sample was over ten times larger and consisted of 6,277 children. More data yields higher power to detect splits in recursive partitioning, thus likely resulting in a larger tree. Comparing the results of Stegmann et al. (2018) corroborated most of our results, which supports that we succeeded in fitting an accurate RP-GCM with glmertree on longitudinal data.

Limitations and future research

The substantial conclusions of the current study may not be generalizable to the entire population, as we analysed only the children who had data on all measurement occasions. However, the central aim of our study was to evaluate the performance of glmertree, not to obtain results that can be used in practice, for instance to improve reading trajectories. Future research is necessary on how to fit a RP-GCM with glmertree, for example on how to deal with missing data.

In our study, we found a RIS model to provide the most accurate results. At the start of the study, we hypothesised that fitting a child-specific random slope might obfuscate effects of interest. Substantial conclusions were similar for the RI and RIS models, which would suggest that the effects are not obfuscated by adding a child-specific random slope to the model. An interesting topic for further (simulation) research would be to use datasets in which strong subject-specific slope effects are present to investigate whether fitting a RIS model with glmertree is still able to pick up these effects correctly.

Another topic that we did not investigate in this study is including time-variant partitioning variables in the model. These need to be incorporated in the model in a different way. A new study could investigate how these variables can be included in the model.

Conclusion

Several aspects are important when fitting a RP-GCM on clustered data. The researcher needs to decide on how to correctly specify the effect of time, random effects and partitioning variables. In case of only fitting a child-specific random intercept (RI), cluster-level parameter stability tests outperform observation-level parameter stability tests. Initializing model estimation with the tree or random effects yielded similar results. When both a child-specific random intercept and slope (RIS) are modelled, a combination of observation-level parameter stability tests and model initialization of the random effects

yielded the most accurate results. The difference in tree-structure or random-effects initialization approaches is smaller in the RI models, but in the RIS models the choice of initialization approach is important. In all instances, it may therefore be better to initialize model estimation with the random effects, because it either yields similar or better results. In our study, the RIS model was preferred over the RI model, because of its performance, but further research is needed to investigate whether the RIS model would still allow for detecting subgroups which show differences in growth over time.

Bibliography

Beauchaine, T. P., Webster-Stratton, C., & Reid, M. J. (2005). Mediators, moderators, and predictors of 1-year outcomes among children treated for early-onset conduct problems: A latent growth curve analysis. *Journal of Consulting and Clinical Psychology*, *73*(3), 371-388.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. New York: Routledge.

Best, J. R., Miller, P. H., & Jones, L. L. (2009). Executive functions after age 5: Changes and correlates. *Developmental Review*, *29*(3), 180-200.

Blakemore, S.-J., & Choudhury, S. (2006). Development of the adolescent brain: Implications for executive function and social cognition. *Journal of Child Psychology and Psychiatry*, *47*(3-4), 296-312.

Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, *3*(2), 63-68.

Duncan, T. E., & Duncan, S. C. (2009). The ABC's of LGM: An introductory guide to latent variable growth curve modeling. *Social and Personality Psychology Compass*, *3*(6), 979-991.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (in press). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*.

Fokkema, M., & Zeileis, A. (2016). Glmertree: Generalized linear mixed model trees. Retrieved from http://R-Forge.R-project.org/R/?group_id=261 (R package version 0.1-2).

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development*, *82*(5), 1357-1371.

Grissmer, D., Grimm, K. J., Aiyer, S. M., Murrah, W. M., & Steele, J. S. (2010). Fine motor skills and early comprehension of the world: Two new school readiness indicators. *Developmental Psychology*, *46*(5), 1008-1017.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313-1328.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R*. New York: Springer.

Jones, J. D., Marsiske, M., Okun, M. S., & Bowers, D. (2015). Latent growth-curve analysis reveals that worsening Parkinson's disease quality of life is driven by depression. *Neuropsychology*, *29*(4), 603-609.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, *80*(4), 437-447.

King, M. W., & Resick, P. A. (2014). Data mining in psychological treatment research: A primer on classification and regression trees. *Journal of Consulting and Clinical Psychology*, *82*(5), 895-905.

Long, J., & Ryoo, J. (2010). Using fractional polynomials to model non-linear trends in longitudinal data. *British Journal of Mathematical and Statistical Psychology*, *63*(1), 177-203.

National Center for Education Statistics. (2016). Early Childhood Longitudinal Program: Kindergarten class of 1998-1999 (ECLS-K). Available from National Center of Education Statistics: https://nces.ed.gov/ecls/kindergarten.asp.

Reddy, R., Rhodes, J. E., & Mulhall, P. (2003). The influence of teacher support on student adjustment in the middle school years: A latent growth curve study. *Development and Psychopathology*, *15*(1), 119-138.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016. https://www.R-project.org/.

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.

Stegmann, G., Jacobucci, R., Serang, S., & Grimm, K. J. (2018). Recursive Partitioning with Nonlinear Models of Change. *Multivariate Behavioral Research*, 1-12.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492-514.

Appendix A: Coding of covariates

Gender (GENDER)

- 1 = Male
- 2 = Female

Race (RACE)

- 1 = White, non-Hispanic
- 2 = Black or African American, non-Hispanic
- 3 = Hispanic, race specified
- 4 = Hispanic, race not specified
- 5 = Asian
- 6 = Native Hawaiian, other pacific islander
- 7 = American Indian or Alaska native
- 8 = More than one race, non-Hispanic

Socioeconomic status (WKSESL)

- -5 – 3, higher values indicating better socioeconomic status

Gross motor skills (C1GMOTOR)

- 0 – 8, higher values indicating better gross motor skills

Fine motor skills (C1FMOTOR)

- 0 – 9, higher values indicating better fine motor skills

Interpersonal skills (T1INTERP)

- 1 – 4, higher values indicating better interpersonal skills

Self-control (T1CONTRO)

- 1 – 4, higher values indicating better self-control

First time in kindergarten (P1FIRKDG)

- 1 = yes
- 2 = no

Internalizing problem behaviour (T1INTERN)

- 1 – 4, higher values indicating less internalising problem behaviour

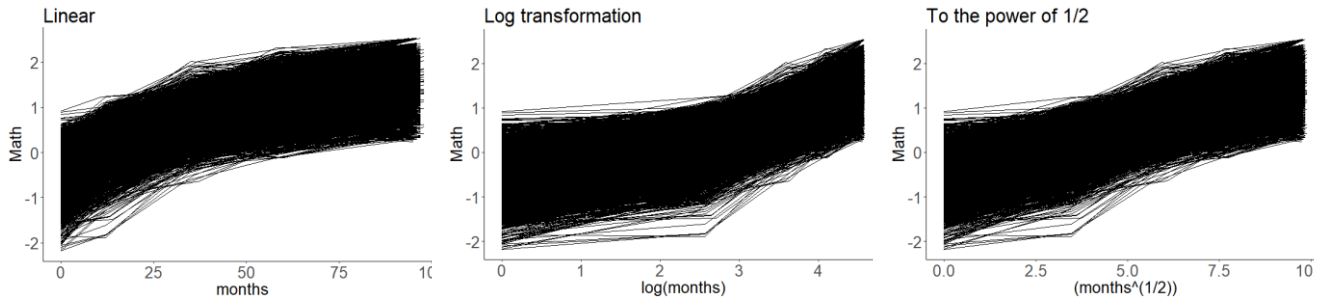
Externalizing problem behaviour (T1EXTERN)

- 1 – 4, higher values indicating less externalising problem behaviour

Age at baseline (AGEBASELINE)

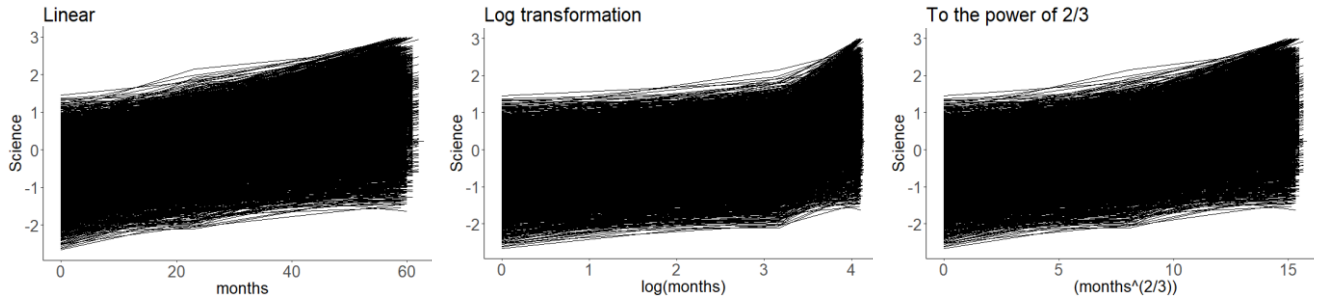
- 1 unit stands for one month

Appendix B: Growth curves of math and science ability



A. B. C.

Figure B1. *Different growth curves of time vs. math ability. Panel A shows the trajectories without transformations. Panel B shows a natural log transformation. Panel C shows the trajectories with a square root function.*

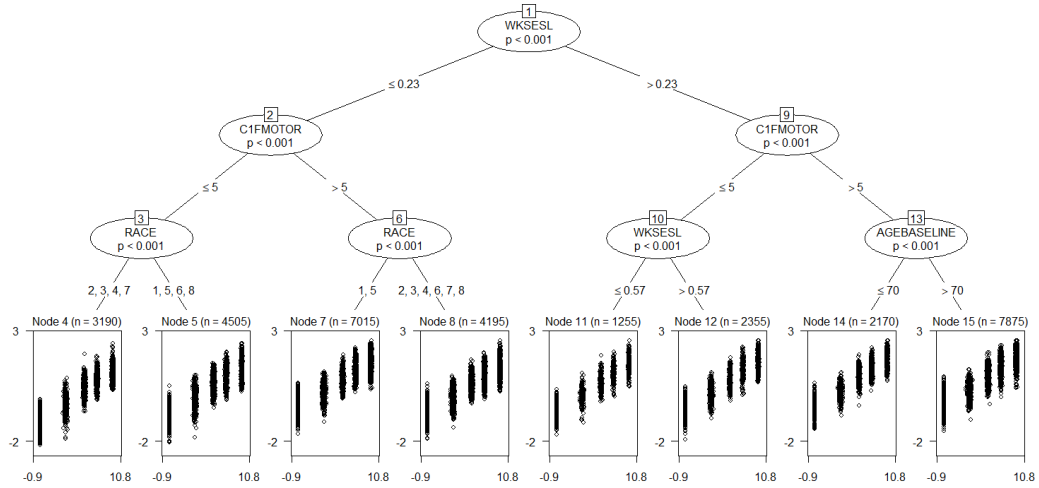


A. B. C.

Figure B2. *Different growth curves of time vs. science ability. Panel A shows the trajectories without transformations. Panel B shows a natural log transformation. Panel C shows a power function of ⅔.*

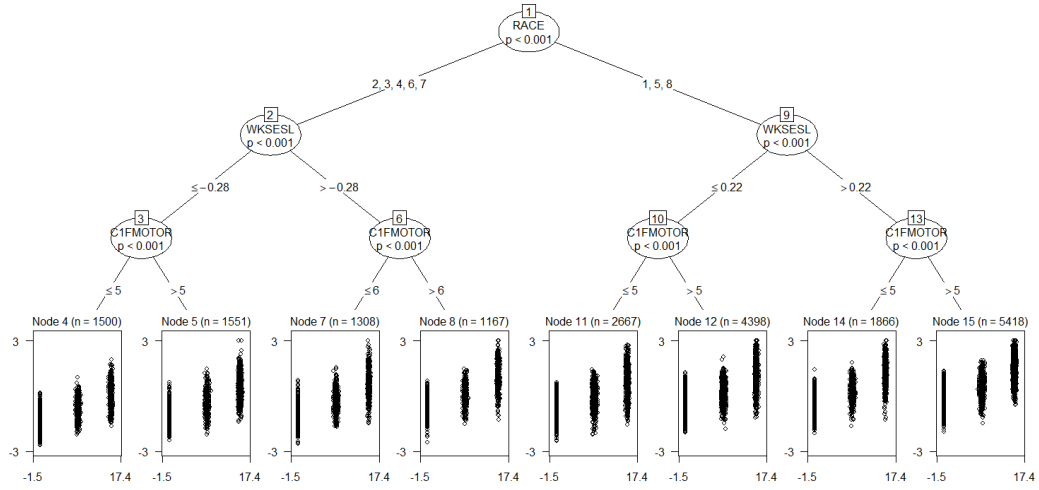
Appendix C: Trees with maximum depth is four

RI models



Intercept	-1.01	-.78	-.49	-.69	-.71	-.56	-.53	-.32
Slope	.23	.22	.22	.22	.23	.23	.23	.22

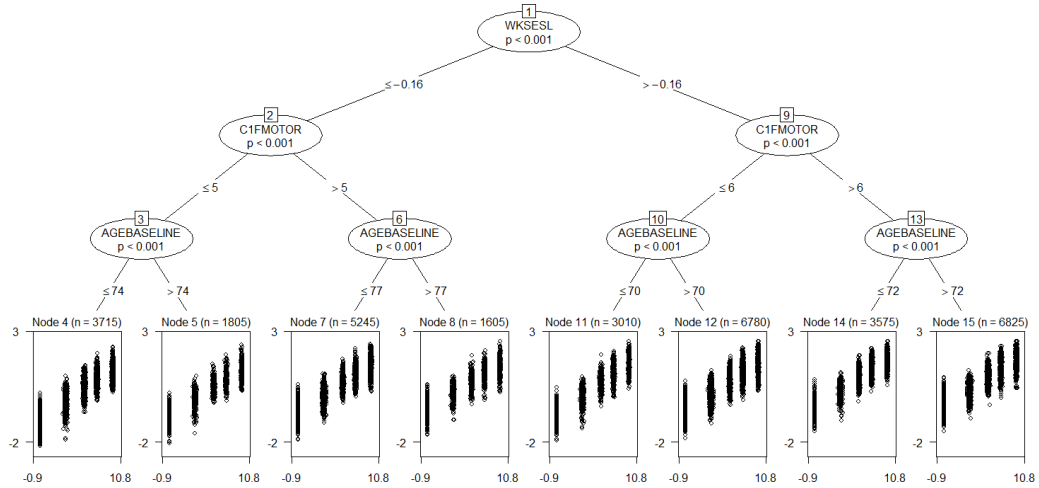
Figure C1. *Plot of restricted math ability tree from the RI model CR on full dataset with maximum tree depth of four. Node-specific intercept and slope estimates are given below the tree.*



Intercept	-1.34	-.99	-.88	-.59	-.70	-.44	-.42	-.21
Slope	.09	.10	.10	.10	.09	.10	-.10	.11

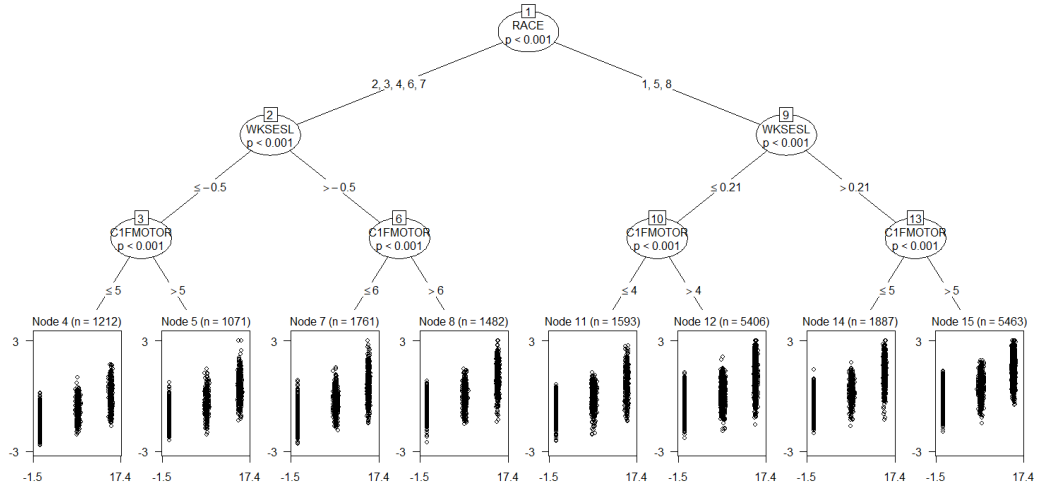
Figure C2. *Plot of restricted science ability tree from the RI model CR on full dataset with maximum tree depth of four. Node-specific intercept and slope estimates are given below the tree.*

RIS models



Intercept	-.98	-.82	-.65	-.48	-.72	-.54	-.47	-.32
Slope	.23	.21	.22	.20	.24	.22	.23	.21

Figure C3. *Plot of restricted math ability tree from the RIS model OR on full dataset with maximum tree depth of four. Node-specific intercept and slope estimates are given below the tree.*



Intercept	-1.38	-1.06	-.94	-.63	-.77	-.47	-.42	-.21
Slope	.09	.10	.09	.10	.09	.10	-.10	.11

Figure C4. *Plot of restricted science ability tree from the RIS model OR on full dataset with maximum tree depth of four. Node-specific intercept and slope estimates are given below the tree.*

Appendix D: Full trees

Reading ability RI model

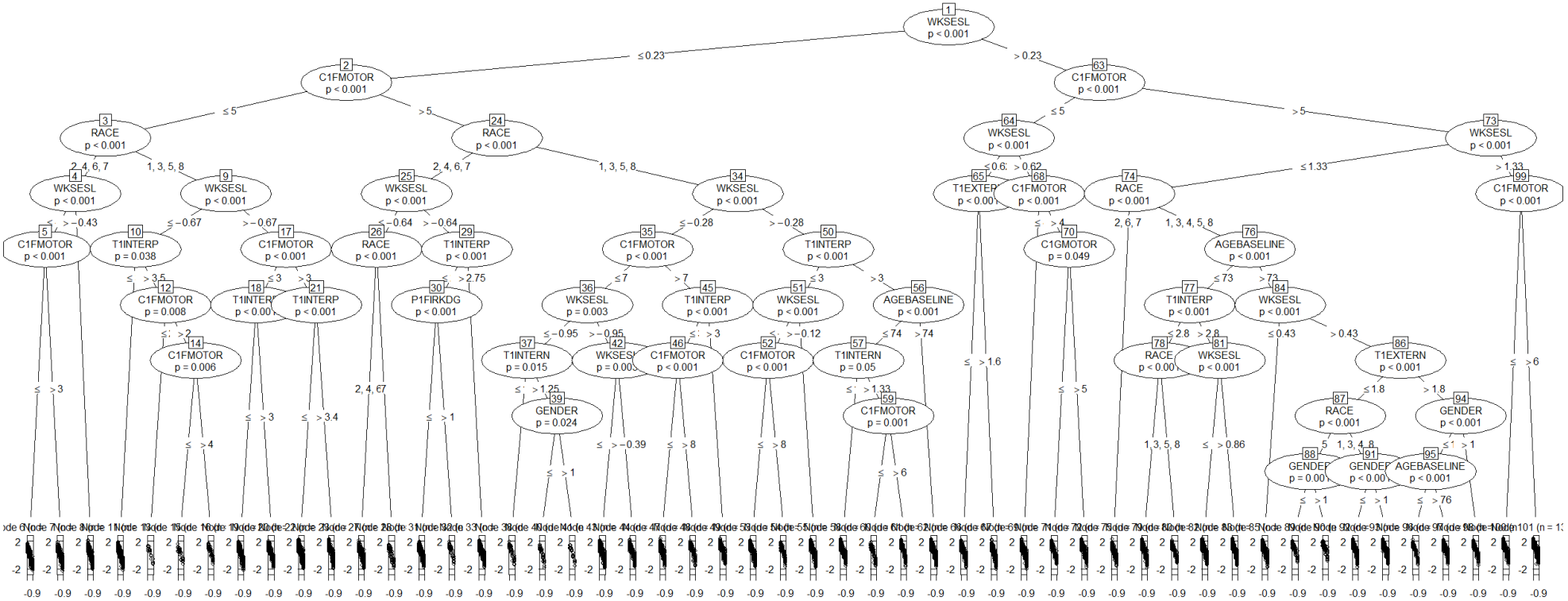


Figure D1. *Full CR tree of reading ability RI model.*

Math ability RI model

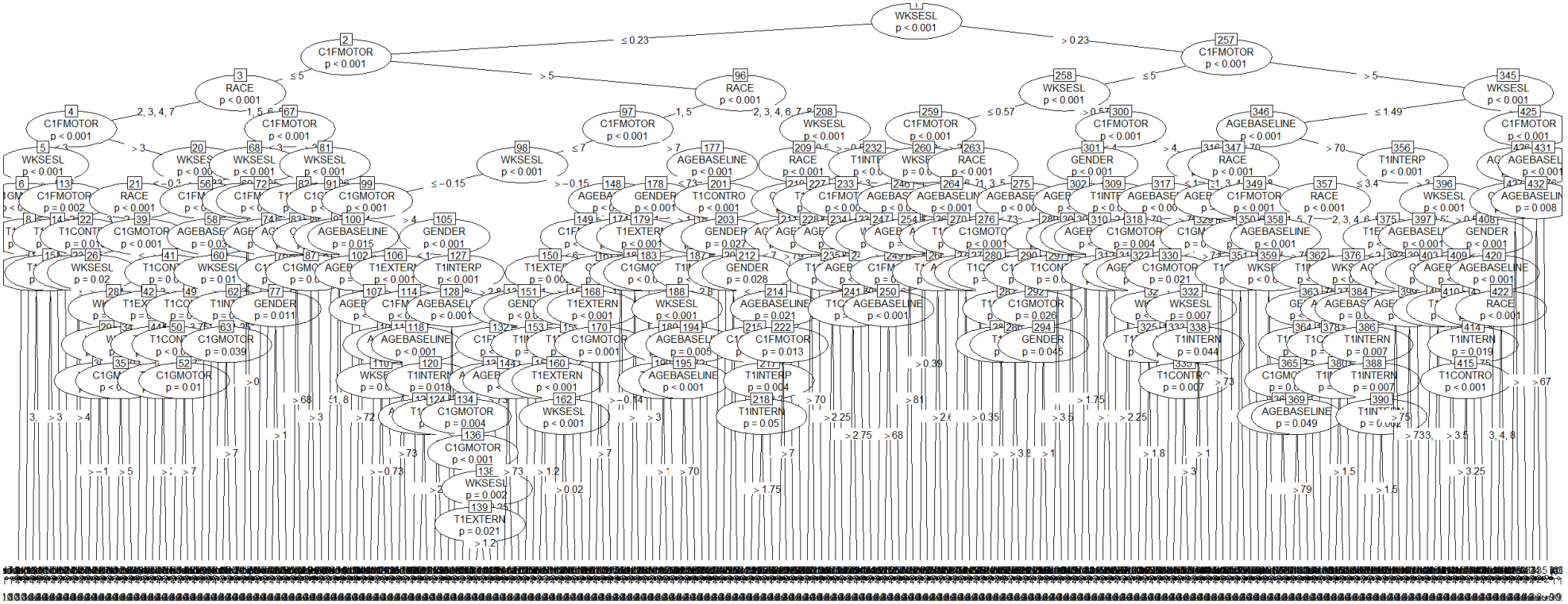


Figure D2. *Full CR tree of math ability RI model.*

Science ability RI model

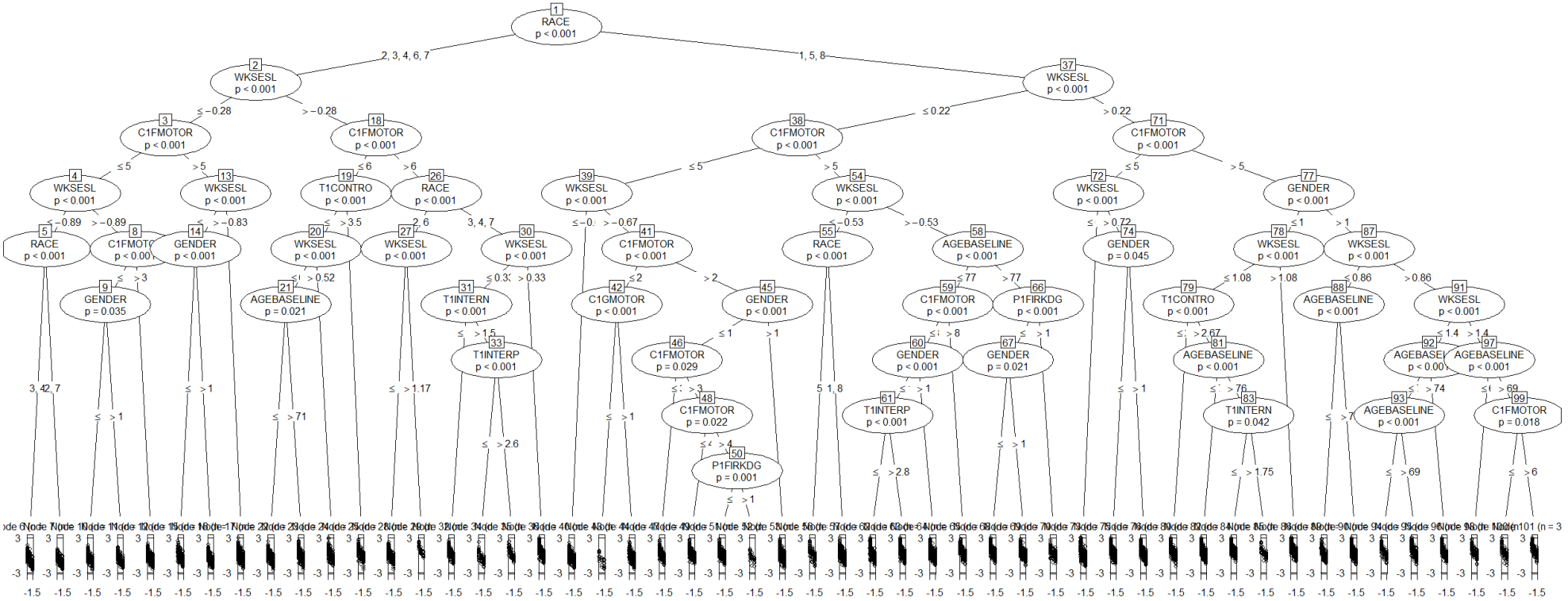


Figure D3. *Full CR tree of science ability RI model.*

Reading ability RIS model

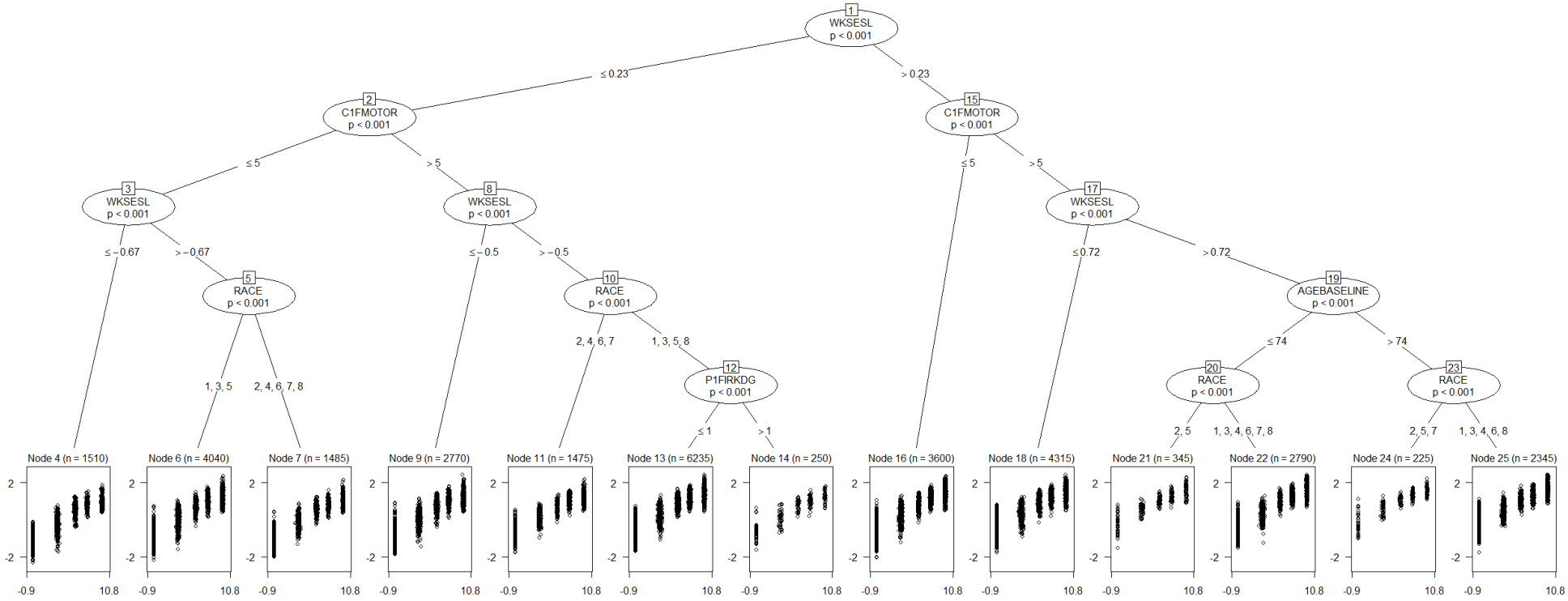


Figure D4. *Full OR tree of reading ability RIS model.*

Math ability RIS model

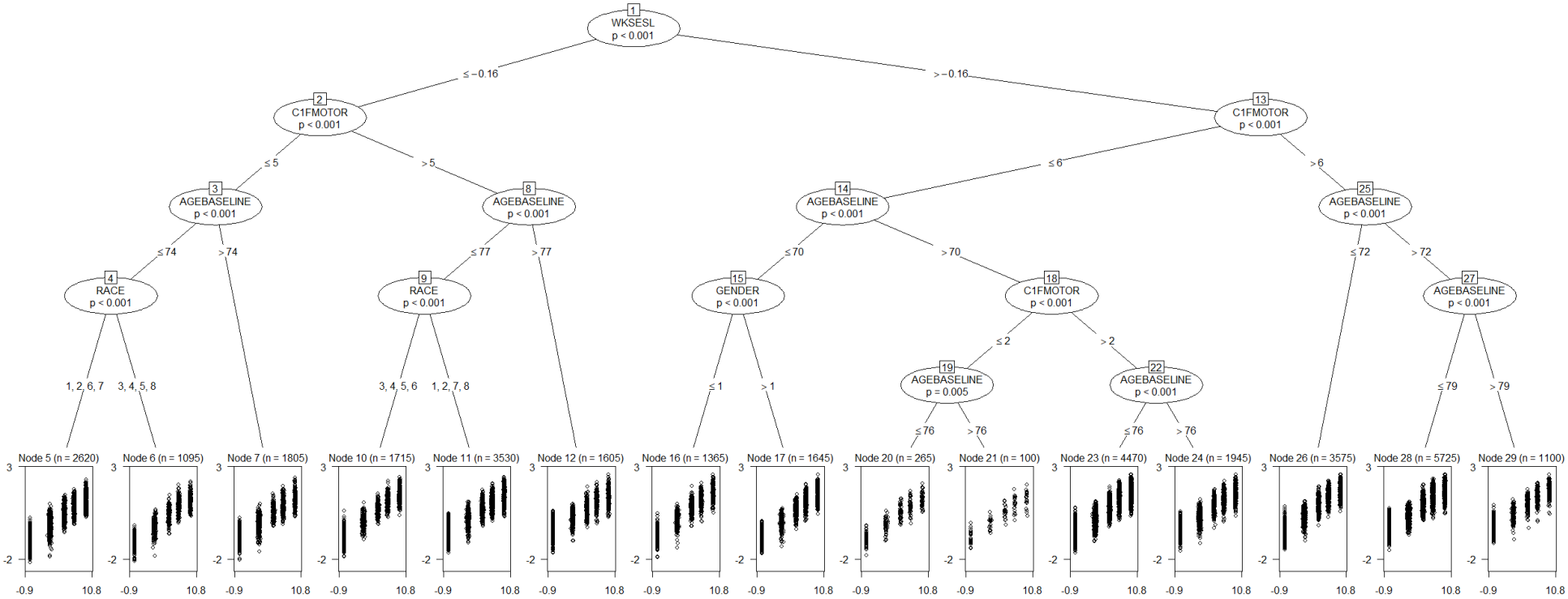


Figure D5. *Full OR tree of math ability RIS model.*

Science ability RIS model

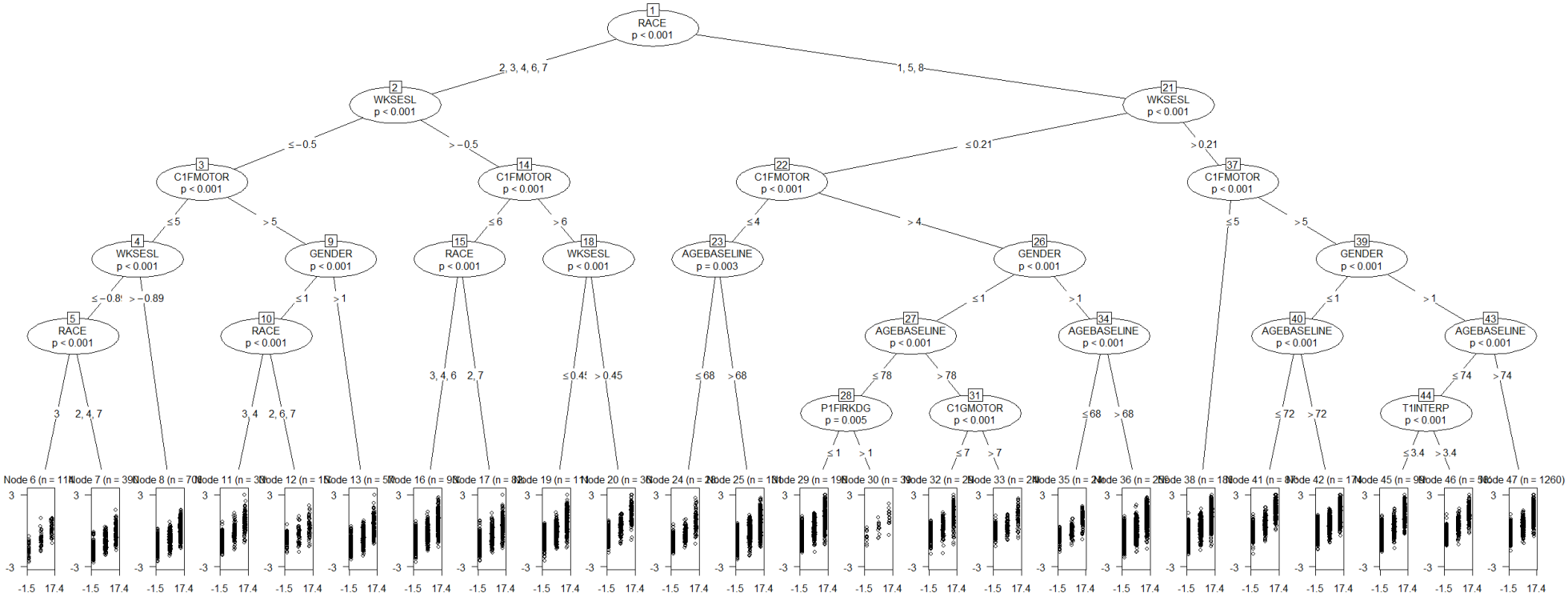


Figure D6. *Full OR tree of science ability RIS model.*