PARCC Cognitive Complexity: Analysis 1, 2, and 3 Results

Steve Ferrara, Jeffrey Steedle, and Amy Kinsman

Pearson Research and Innovation Network

July 27, 2015

Abstract

We report results from the following three analyses of PARCC cognitive complexity measures, based on 2014 field test item and task development and field test data.

We conducted classification and regression tree analyses using 2014 PARCC field test data to do the following:

- Predict item difficulty and discrimination for samples of PARCC ELA and mathematics items and tasks and evaluate the interpretability and usability of the results for assembling operational test forms
- Evaluate the explanatory power of the four ELA and five mathematics cognitive complexity sources and propose final explanatory models
- Demonstrate how regression tree results can be used to assemble operational test forms

In addition, we surveyed cognitive complexity coders at Pearson to capture their insights and recommendations for PARCC on the cognitive complexity framework, training of complexity coders, the coding process and decision making, and future considerations for cognitive complexity. We also conducted a focus groups with ELA/literacy and mathematics coders to examine more closely responses to selected survey questions.

Presentation slides we used to conduct briefings on this report on July 23, 2015 for the ELA/literacy and mathematics Operational Working Groups appear in Appendix O.

Table of Contents

Additional Appendices Provided as Separate Documents:

- **Appendix J Survey Monkey Questions**
- **Appendix K Survey Summary**
- **Appendix L PARCC Survey Data Sheet**
- **Appendix M PARCC CC analysis I Results**
- **Appendix N PARCC CC for CA SIG Session**
- **Appendix O PARCC CC Slides July Briefings**

1. INTRODUCTION

PARCC has expressed interest in using its cognitive complexity measures to assemble field test forms for 2014 and operational test forms for 2015. The goal is to produce test forms with approximately equivalent measurement precision for low, moderate, and high proficiency students whose test performance is likely to place them in the low, middle, and high ranges of the test score scales, respectively. PARCC's strategy to achieve that goal is to create test forms in which items and tasks of low, medium, and high difficulty and cognitive complexity are approximately uniformly distributed across the ELA and mathematics test score scales.

Ordinarily, item difficulty statistics and parameters are used to pursue this goal, in a process called test information targeting (e.g., Yen & Fitzpatrick, 2006). An assumption underlies the use of cognitive complexity measures instead of or in addition to item difficulty indicators: There is a modest relationship between cognitive complexity and item difficulty, suggesting that cognitive complexity indicates something about item and task response demands (e.g., Ferrara, Svetina, Skucha, & Davidson, 2011) in addition to how difficult and discriminating they are. So, using item cognitive complexity to assemble PARCC test forms enables PARCC to pursue equivalent task complexity for all examinees. And using both item difficulty and cognitive complexity to assemble test forms enables PARCC to pursue both equivalent measurement precision and equivalent task complexity for all students.

In June 2012, PARCC and Pearson met to discuss development of item cognitive complexity frameworks for ELA/literacy and mathematics. In summer 2012, Pearson and ETS jointly created these frameworks and delivered final, updated reports in October. In summer 2013, Pearson and ETS worked separately to develop items and tasks and to code items and tasks for cognitive complexity. Both organizations also implemented a judgmental weighting scheme, determined by PARCC, for creating overall item and task complexity measures.[1]

While the judgmental weighting process is adequate for assembling field test forms, an empirical process for estimating item and task cognitive complexity is preferable for assembling operational test forms. The advantages of an empirical process include (a) Empirical weights reflect the relationship between cognitive complexity and item difficulty and thus, indicate the degree to which difficulty and complexity provide supplementary information to guide the test forms assembly process; and (b) Empirical modeling of cognitive complexity can account for interactions among the individual complexity sources in their relationships with item difficulty. In addition, the empirical modeling process provides information to evaluate the absolute and relative importance of each individual cognitive complexity source. Estimating and investigating empirical weights by predicting item difficulty statistics (and discrimination statistics) from item codes (e.g., content requirements, other response demands) is commonly referred to as item difficulty modeling (e.g., Gorin & Embretson, 2006). Analytic methods used in item difficulty modeling in published studies include ordinary least squares regression and latent class analysis. The PARCC assessment Technical Advisory Committee proposed using classification and

[1] ELA weights are TC=0.5, CTE=0.225, RM=0.225, PD=0.05; decision rules are Low {1.0,1.6}; Moderate {1.7,2.2}; High {2.3,3.0}; mathematics weights are MC=0.3, MP=0.4, SM=0.1, RM=0.1, PD=0.1; decision rules are Low {1.0,1.4}; Moderate {1.5,2.1}; High {2.2,3.0}.

regression tree analysis because of the advantages they saw in creating empirically weighted complexity measures and accounting for interactions among complexity sources.

In this study, we conducted classification and regression tree analyses using 2014 PARCC field test data to do the following:

- Predict item difficulty and discrimination for samples of PARCC ELA and mathematics items and tasks and evaluate the interpretability and usability of the results for assembling operational test forms
- Evaluate the explanatory power of the four ELA and five mathematics cognitive complexity sources and propose final explanatory models
- Demonstrate how regression tree results can be used to assemble operational test forms

In addition, we surveyed cognitive complexity coders at Pearson to capture their insights and recommendations for PARCC on the cognitive complexity framework, training of complexity coders,[2] the coding process and decision making, and future considerations for cognitive complexity. We also conducted a focus groups with ELA/literacy and mathematics coders to examine more closely responses to selected survey questions.

2. METHOD AND PROCEDURES

Cognitive Complexity Source Codes and Test Items and Tasks included in the Analyses

The tasks analyzed in this study included all phase 1 PARCC items and tasks that were coded for either the four ELA/literacy cognitive complexity sources or the five mathematics cognitive complexity sources.

ELA/literacy tasks were coded for the following source codes; all four codes were included in overall item and task cognitive complexity measures:

> **Text Complexity:** Defined in a separate process using a composite of Source Rater, Reading Maturity Metric, Lexile automated, quantitative text complexity measures plus qualitative estimates of text complexity by experts using the Literary and Informational text complexity rubrics developed specifically for PARCC assessments.[3]
>
> **Command of Textual Evidence**: Defined as the amount of text that an examinee must process (i.e., select and understand) in order to respond correctly to an assessment item. This category focuses on the numbers of details in one or more texts that must be processed in order to respond to the requirements of items. The amount of text processed is influenced by both the cognitive complexity of items and tasks and the complexity of the text or texts.[4]

[2] The limited information currently documented on coder training and validity check sets agreement rates appear on slides 10-12 in Appendix N.

[3] The text complexity code, based on a composite of the quantitative and qualitative measures, were determined in a separate process and included in the item cognitive complexity judgment. This Text Complexity measure is applied only to reading selections and graphical material, not to items.

[4] In contrast to Text Complexity, Command of Textual Evidence is a judgment about the cognitive complexity of items and tasks themselves, based on the complexity of processing reading selections and graphical material required to respond to the item or task.

Response Mode: The way in which examinees are required to complete assessment activities influences an item's cognitive complexity, including selecting a response from among given choices and generating an original response.

Processing Demands: Include linguistic demands and reading load in item stems, instructions for responding to an item, and response options. Linguistic demands include vocabulary choices, phrasing, and other grammatical structures. Length of item stems, instructions for responding to an item, and response choices define reading load.

Mathematics tasks were coded for the following cognitive complexity source codes; all four codes were included in overall task cognitive complexity measures:

Mathematical Content: Typical expectations for mathematical knowledge at the grade level, including new mathematical concepts and skills that require small or large shifts from previously learned concepts and skills.

Mathematical Practices: What students are asked to do with mathematical content, such as engage in application and analysis of the content, based on expectations of a typical student at a grade level and the content reflected in the Common Core State Standards.

Stimulus Material: The number of different pieces of stimulus material in a task and the role of technology tools in the task.

Response Mode: Defined as above: The way in which examinees are required to complete assessment activities influences an item's cognitive complexity, including selecting a response from among given choices and generating an original response.

Processing Demands: Defined as above: Include linguistic demands and reading load in item stems, instructions for responding to an item, and response options. Linguistic demands include vocabulary choices, phrasing, and other grammatical structures. Length of item stems, instructions for responding to an item, and response choices define reading load.

The source codes for most tasks were extracted from Item Tracker-Test Builder (ITTB). When codes were not available from ITTB, they were extracted from the raw coding data and averaged across multiple coders if applicable. For each task, an overall cognitive complexity measure was calculated by converting the source codes (low, moderate, high) to integers (e.g., 1, 2, 3) and applying judgmental weights.

ELA/literacy source codes were weighted 0.5 (Text Complexity), 0.2 (Command of Textual Evidence), 0.2 (Response Mode), and 0.1 (Processing Demands). The mathematics source codes were weighted 0.3 (Mathematical Content), 0.4 (Mathematical Practices), 0.1 (Stimulus Material), 0.1 (Response Mode) and 0.1 (Processing Demands). The resulting cognitive complexity measures ranging from 1.0 to 3.0 were converted to low, moderate, or high according to the following rules, which were based on the observed probabilities of each complexity score:

ELA/literacy – Low (1.0, 1.6); Moderate (1.7, 2.2); High (2.3, 3.0)

Mathematics: Low (1.0, 1.4); Moderate (1.5, 2.1); High (2.2, 3.0)

In analyses 1 and 2, items and tasks were omitted if they were missing the outcome variable (e.g., if they were not administered during 2014 field-testing), if they were marked as DNU (do not use), or if they were administered off-grade-level (e.g., a grade 6 task administered to grade 7 students for the purposes of vertical scaling). In order to avoid statistical dependencies in the data (and resulting over-emphasis of certain tasks in the results), duplicate tasks were removed. Specifically, if a task was administered in both computer based test (CBT) and paper based test (PBT) modes, the CBT task was

retained. (Generally, there were many more CBT than PBT tasks.) If a task was administered in both PBA and end of year (EOY) components, the EOY version was retained. The final task counts are provided in Table 2.1.

Table 2.1 *Task Counts*

ELA literacy	N	Mathematics	N
Grade 3	156	Grade 3	324
Grade 4	222	Grade 4	328
Grade 5	169	Grade 5	279
Grade 6	197	Grade 6	241
Grade 7	162	Grade 7	256
Grade 8	167	Grade 8	258
Grade 9	109	Algebra I	223
Grade 10	71	Algebra II	239
Grade 11	192	Geometry	226
		Integrated Math 1	80
		Integrated Math 2	68
		Integrated Math 3	67

Variables

Several statistics based on data from the 2014 PARCC field test were available from the item calibrations conducted by ETS to serve as measures of task difficulty and discrimination. As a measure of difficulty, the p-value (or "P+") indicates the average proportion of total points scored on a task. In addition, item response theory (IRT) estimates of difficulty were available for the one-, two-, and three-parameter logistic IRT models. However, only the one-parameter model difficulties were considered for this study because item difficulties from the two- and three-parameter models cannot be compared across tasks in a simple manner. The point-biserial (or item-total) correlation indicates the correlation between a task score and the overall test score. Biserial (for dichotomous items) and polyserial (for polytomous items) correlations indicate the correlation between a task score and the overall test score, except that they assume a continuous distribution underlying the dichotomous or polytomous scores. The biserial or polyserial correlations were calculated in two ways: using only the base-test operational tasks and using the base-test and field-test tasks.

Besides the cognitive complexity source codes, numerous variables describing tasks were available to serve as predictors of task difficulty, task discrimination, or overall cognitive complexity. For ELA/literacy, the potential predictors were test administration mode (CBT or PBT), number of score categories, test component (EOY or PBA), PARCC item type, response type, interaction type, technology-enhanced item type, task type, PARCC evidence statement, PARCC sub-claim, PARCC task model, passage identifier, media type, PARCC number of points, set identifier, passage word count, passage type, and PARCC stimulus identifier.

In mathematics, these variables included mode (CBT or PBT), number of score categories, component (EOY or PBA), PARCC item type, response type, interaction type, technology-enhanced item type, PARCC evidence statement, PARCC sub-claim, PARCC task model, companion materials, PARCC number of points, calculator code, PARCC stimulus identifier, and Common Core State Standards identifier.

Regression Tree Overview

Classification and regression tree (CART) analysis comprises a family of multivariate statistical techniques that are used to create binary decision trees (Breiman, Friedman, Olshen, & Stone, 1984). Decision trees are "grown" using a data set including a series of predictor variables X_1, X_2,…, X_p and an outcome variable Y. Once grown, the decision tree can be used to predict the value of Y for new observations based on their X_1, X_2,…, X_p values.

At each branching node within a decision tree, the user applies a binary test to one of the predictor variables (e.g., Does X_2 equal a certain categorical variable value? Is X_4 greater than a certain value?). The result of that test tells the decision tree user to move to the left or right in the tree. Eventually, the user reaches a terminal node (or a "leaf"), at which point the tree provides a prediction of Y. Classification trees are used when the outcome variable is categorical, and regression trees are applied to continuous outcome variables.

As an example, the Figure 2.1 shows a regression tree based on automobile data from the April 1990 issue of *Consumer Reports*. In this data set, the outcome variable is price, and the predictors include country of origin, reliability rating (1–5), fuel economy (miles per gallon), and car type (compact, large, medium, small, sporty, or van). The "root" node of the tree shows a split based on whether the vehicle is a small car. If it is a small car, the regression tree user would move to the left and reach a terminal node, where the predicted price is $7,682 (the mean of all the cars that belong in that node). If it is not a small car, the user would move right to a branching node that is split on country of origin. If the car is from France, Germany, Japan, and Sweden, the user would move to the right, and the predicted price would be $16,086. Otherwise, the user would move left to a branching node that is split on car type. Depending on car type, the predicted price would be either $11,056 or $14,183.

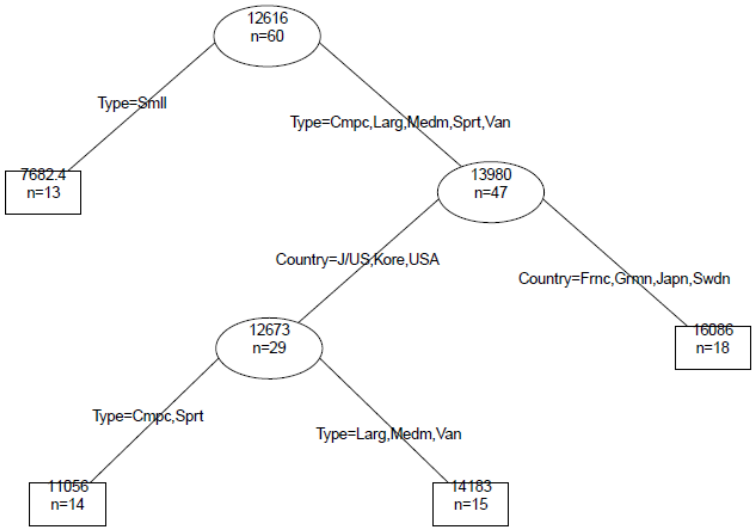


Figure 2.1. Regression tree for the April 1990 *Consumer Reports* data.

Unlike other statistical modeling approaches such as OLS regression, it is not necessary in CART to specify the interactions between predictor variables. Indeed, CART automatically seeks out and detects

complex interactions. In the decision tree, the sequence of branching nodes reflects the interactions between the variables. For example, after separating the small cars from other car types in the root node, country of origin becomes the most important variable for predicting car prices. Then, after controlling for country of origin, car type again becomes the most important predictor of car prices.

To grow the decision tree, recursive partitioning is employed to split the outcome space into progressively smaller regions. The algorithm determines which split to make by searching for the value of one of the X variables which, if used as the splitting criterion, would minimize the variance in *Y* in the resulting nodes (i.e., so that the observations in the resulting nodes are similar on the outcome variable *Y*). The splitting process could proceed until each observation is in its own leaf. However, this maximum tree would not likely generalize well to new data sets. In this study, a stopping criterion was employed such that splits were not added to trees unless they reflected statistically significant improvements in the predictive accuracy of the tree (at the α = .05 level).

CART analyses provide indicators of the relative importance of variables as predictors. In the automobile example, the importance statistics are 100 for car type, 56 for fuel economy, 44 for country of origin, and 4 for reliability. These statistics are scaled to have a maximum of 100, but other transformations of this scale are common (e.g., adding up to 100 or 1.0). Regardless, the importance statistics can be interpreted as ratio scaled (e.g., car type is approximately twice as important as fuel economy). CART can identify variables as important even if they do not appear in decision nodes (e.g., fuel economy). This is achieved by examining the quality if the tree had other ("surrogate") variables been used to split the data. In this study, these values were used as empirical weights for the cognitive complexity sources and also to indicate which variables were the most important predictors of item statistics and overall cognitive complexity.

CART is viewed as advantageous because it (a) produces readily interpretable decision trees, (b) is nonparametric, meaning that it does not require making assumptions about the distributions of variables or specifying a statistical model, (c) performs variable selection automatically, and (d) deals easily with noisy data and outliers. One notable disadvantage of decision trees is that each node can involve only one variable, so a large and complex tree is sometimes required to model relatively simple data structures. Moreover, decision trees may be sensitive to the training data and, thus, limited in their generalization to other data sets.

In this study, an advanced version of regression tree analysis was used: conditional random forests. A conditional tree (Hothorn, Hornik, & Zeileis, 2012) improves upon a basic regression tree because it corrects for possible bias in variable selection (i.e., categorical variables with many possible values are more likely to be selected by chance). A "random forest" of conditional trees is created by growing many conditional trees based on random samples of predictor variables and random samples of data (Breiman, 2001). A prediction for an individual observation reflects the aggregation of predictions from all trees in a forest. Unbiased (or "out-of-bag") estimates of predictive accuracy (R^2) are obtained by considering only predictions based on trees in the forest that were not grown using a given observation.

Analysis 1 Procedures

The major goal of Analysis 1 was to examine the relationship between the individual cognitive complexity source codes and task difficulty. These procedures were also applied in an exploratory fashion to examine the prediction of task discrimination measures. As a first step, descriptive statistics (mean, standard deviation, and frequency) were calculated for each cognitive complexity variable. In addition, each cognitive complexity variable was correlated with the outcome variable. To help illustrate

results and appropriate interpretations, example results from the analysis of grade 3 mathematics tasks are shown throughout this section. Example descriptive statistics and correlations for grade 3 mathematics are shown in Table 2.2. Note that the cognitive complexity source codes correlated negatively with p-values because lower p-values indicate greater task difficulty. This table also shows descriptive statistics and correlations for the overall cognitive complexity measures based on the judgmental weights. The numeric judgmental cognitive complexity (CC) measure in Table 2.2 equaled the judgmental weights multiplied by the individual source code values (1 for low, 2 for moderate, and 3 for high). The ordinal values are rounded versions of the numeric measures (using the rounding rules described above). The empirical cognitive complexity measures (numeric and ordinal) were based on the empirical weights derived from the conditional random forest analyses described below.

The initial descriptive analysis illustrated by Table 2.2 was repeated for p-values and for one-parameter IRT model difficulty estimates. Conditional random forest R^2 values were examined to determine which of those values was better predicted by cognitive complexity source codes. Only results for that variable (p-values) are reported here and in the results section.

Table 2.2
Descriptive Statistics for Predicting Grade 3 Mathematics P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Mathematical Content (MC)	1.78	0.62	107	182	35	-.359	***	.129
Mathematical Practices (MP)	1.28	0.52	242	72	10	-.356	***	.127
Stimulus Material (SM)	1.56	0.82	212	44	68	.010		.000
Response Mode (RM)	1.15	0.52	296	6	22	-.301	***	.090
Processing Demands (PD)	1.52	0.56	164	150	10	-.158	**	.025
Judgmental CC Measure (ordinal)	1.51	0.62	181	122	21	-.373	***	.139
Judgmental CC Measure (numeric)	1.47	0.38				-.429	***	.184
Empirical CC Measure (ordinal)	1.53	0.67	185	107	32	-.384	***	.148
Empirical CC Measure (numeric)	1.47	0.41				-.443	***	.196

* $p < .05$, ** $p < .01$, *** $p < .001$

A conditional random forest with 1,000 trees was constructed using the data for each grade level or course within a subject area. Each conditional tree in the random forest was grown from three randomly selected cognitive complexity source codes for a random sample of available tasks. Importance statistics from the random forest were scaled to sum to 1.0 to serve as empirical weights for the cognitive complexity source codes. To obtain the empirical cognitive complexity measure (see Table 2.3 for an example), the numeric weighted composites were rounded using the same rounding rules as the judgmental cognitive complexity measures. As an example, weights from the analysis of grade 3 mathematics p-values are shown in Table 2.3. For example, the highest weight was .39 for mathematical content (MC), which indicates that mathematical content was the most important predictor of task difficulty for grade 3. With a weight of .20, response mode (RM) was about half as important as mathematical content. Note that the empirical weights are not regression coefficients, so they cannot be used to calculate an expected cognitive complexity measure. They reflect only the

relative importance of the predictors across the conditional trees in a random forest. Combining them in the manner described above provides an overall measure that reflects task difficulty.

Note also that the interpretation of these weights is conditional on the successful prediction of p-values from the cognitive complexity source codes (see conditional random forest R^2). That is, if R^2 is very low, the weights cannot be meaningfully interpreted. In this study, R^2 values below .10 were considered too low to interpret.

Table 2.3. *Empirical Weights for Mathematics Based on Analysis of P-Values*

							R^2	
	N	MC	MP	SM	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	324	.39	.32	.04	.20	.05	.204	.131	.161

Table 2.3 shows two other R^2 values: conditional tree and conditional tree cross-validation. Before fitting the conditional random forest, a single conditional tree was fit to the data because it may suggest recommendations about how to manipulate cognitive complexity source codes to impact task difficulty. Note that the conditional random forest offers optimal prediction, but it provides no single tree for interpretation like that in Figure 2.2 (i.e., there are 1,000 trees based on a smaller number of predictors). Note that the variables in Figure 2.2 are mathematical content (MC), mathematical practices (MP), and response mode (RM). Ten-fold cross-validation was used to obtain an unbiased estimate of a single tree's predictive accuracy. In that process, 10 different conditional trees were be grown, where, for a given tree, 9/10 of the data were used to grow the tree, and the other 1/10 was withheld for validation. In Table 2.3 (and throughout the results section below), the cross-validation R^2 values are notably lower than the conditional tree R^2 values, which suggests that the conditional tree likely reflects overfitting of the data and would only generalize to new items to the extent indicated by the cross-validation R^2.

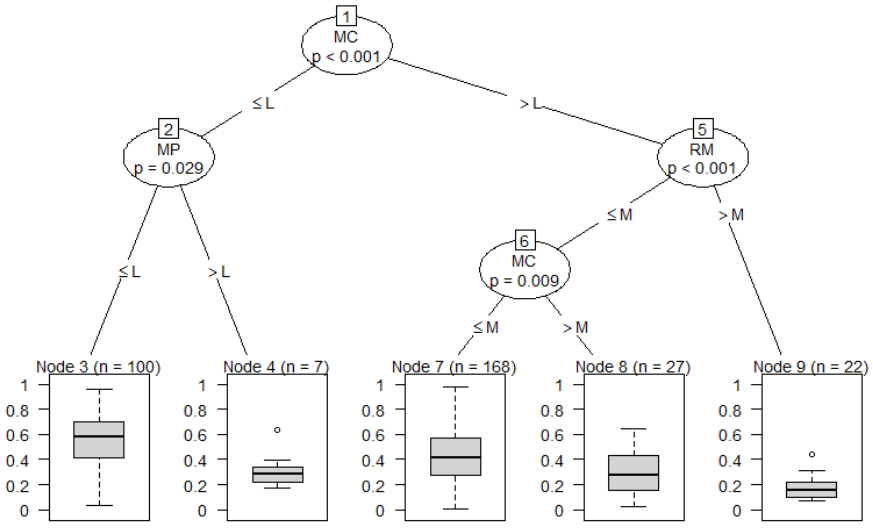


Figure 2.2. Conditional tree for predicting p-values from cognitive complexity source codes.

R^2 values like those in Table 2.3 indicate the degree to which cognitive complexity source codes explain variance in task difficulty. However, without a point a reference, it is challenging to decide whether the cognitive complexity source codes are doing a "good" job of explaining variance in task difficulty. A reasonable point of reference would be the proportion of variance explained by a much larger set of variables that are possibly associated with task difficulty. So, in the final step of Analysis I, a conditional random forest was fit to predict task difficulty from the cognitive complexity source codes, overall cognitive complexity, and all available metadata variables. (See Appendix M for a list of the item metadata variables.) In addition to providing an important point of reference, results also indicate what variables (e.g., metadata or cognitive complexity) are the most important predictors of task difficulty. The values shown in Table 2.4 are importance statistics. Following common practice, the values in 2.4 were scaled to have a maximum of 100. This differs from Table 2.3 where the importance statistics were scaled to sum to 1.0 to make them comparable to the judgmental weights. As before, a single conditional tree (with 10-fold cross-validation) was fit to suggest recommendations about how to manipulate cognitive complexity and metadata to impact task difficulty. The example shown in Figure 2.3 includes the variables Common Core State Standards Identifier, mathematical content (MC), overall cognitive complexity measure (CCM), and number of score categories (ScoreCat).

Table 2.4. *Importance Statistics for Predictors of Mathematics Task P-Values for Grade 3 Mathematics*

	Grade 3 Importance
Math Content	28
Math Practices	16
Stimulus Material	3
Response Mode	5
Processing Demands	4
Overall Cognitive Complexity	35
Mode (CBT or PBT)	0
Number of Score Categories	19
Component (EOY or PBA)	0
PARCC Item Type	26
Response Type	7
Interaction Type	12
TEI Type	46
PARCC Evidence Statement 1	100
PARCC Sub-claim	6
PARCC Task Model 1	84
Companion Materials	0
PARCC Number of Points	18
Calculator Code	0
PARCC Stimulus Identifier	1
CCSS Identifier 1	54
CCSS Identifier 2	1

Table 2.4. *Importance Statistics for Predictors of Mathematics Task P-Values for Grade 3 Mathematics*

	Grade 3 Importance
R^2 Cond. Tree	.372
R^2 Cond. Tree Cross-Val.	.199
R^2 Cond. Random Forest	.402

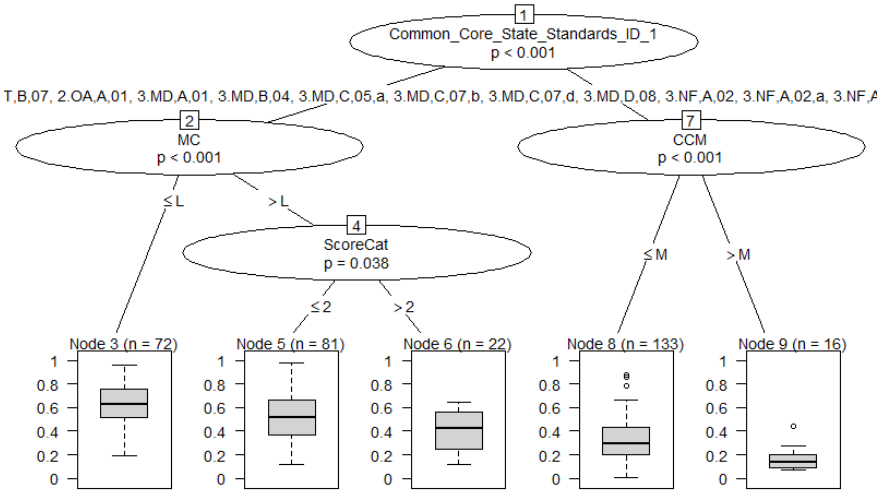


Figure 2.3. Conditional tree for predicting p-values from cognitive complexity source codes and metadata variables.

Analysis 2 Procedures

Conditional random forests also were used in Analysis II, which had the goal of determining the most important predictors of overall cognitive complexity. The methodology of Analysis 2 was identical to Analysis I, except that the outcome variable was overall cognitive complexity based on the judgmental weights and only metadata were used as predictors. Importance statistics were derived from the conditional random forest (see Table 2.5 for an example), and a single conditional tree (with 10-fold cross-validation) was fit to suggest recommendations about how to manipulate metadata to impact overall cognitive complexity (see Figure 2.4 for an example).

Table 2.5. *Importance Statistics for Predictors of Mathematics Task Overall Cognitive Complexity for Grade 3 Mathematics*

	Grade 3
Mode (CBT or PBT)	0
Number of Score Categories	9
Component (EOY or PBA)	0

Table 2.5. *Importance Statistics for Predictors of Mathematics Task Overall Cognitive Complexity for Grade 3 Mathematics*

	Grade 3
PARCC Item Type	34
Response Type	82
Interaction Type	100
TEI Type	56
PARCC Evidence Statement 1	55
PARCC Sub-claim	14
PARCC Task Model 1	36
Companion Materials	0
PARCC Number of Points	24
Calculator Code	0
PARCC Stimulus Identifier	1
CCSS Identifier 1	43
CCSS Identifier 2	3
R^2 Cond. Tree	.534
R^2 Cond. Tree Cross-Val.	.336
R^2 Cond. Random Forest	.449

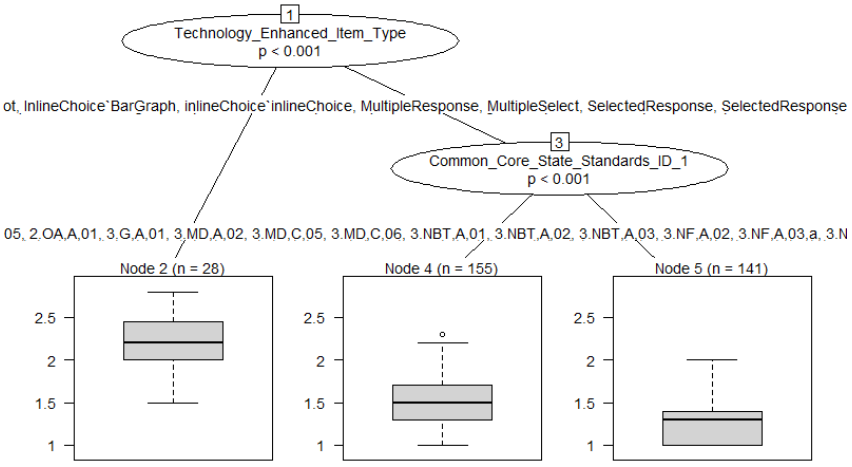


Figure 2.4. Conditional tree for predicting overall cognitive complexity from metadata variables.

Analysis 3 Procedures

To supplement the results and conclusions from analyses 1 and 2, PARCC requested that we gather insights from content developers who coded items for cognitive complexity about the complexity codes,

training, coding process, areas for improvement, and ongoing use of the cognitive complexity framework. We surveyed 56 content developers at Pearson who coded PARCC ELA and mathematics items and tasks during the periods of development leading up to the spring 2014 field test. We also conducted focus group interviews with selected subsets of the ELA and mathematic coders.

Goals of the Analysis

We surveyed and interviewed Pearson cognitive complexity coders to address the following questions.

- What difficulties did item complexity coders encounter in coding items for cognitive complexity? Are the definitions of the cognitive complexity sources clear and relevant to the PARCC items? Are distinctions between high, medium, and low levels of complexity clear? Was making judgments about individual sources of complexity versus overall judgments of complexity more or less challenging?
- What should PARCC learn from implementing the cognitive complexity framework? Would item complexity coders recommend any changes to the framework and how it is used? How can item complexity measures be used in future item and task development?

Survey Topics and Questions

Pearson's study team brainstormed candidate survey questions that would address the research questions (above). We evaluated candidate questions and selected the most important questions that could be included in a survey with an expected average response time of 20 minutes. The team then decided which questions should be posed as selected response questions, which as open ended responses questions. Finally, we sorted the selected questions into topical areas: respondent background, other involvement in PARCC, cognitive complexity training, coding and decision making processes, and recommendations for the future.

Data Sources

We surveyed 56 cognitive complexity coders in ELA and mathematics in the elementary, middle, and high school grade bands using the online survey service, Survey Monkey. A PDF of the survey as it appeared to respondents is viewable in Appendix J.

We reviewed response frequencies and open ended responses to select survey questions to probe more extensively in focus group interviews. We conducted one focus group interview in ELA and one in mathematics via conference call and WebEx.

3. ANALYSIS 1 RESULTS

As noted in the previous section, several measures of task difficulty and discrimination were available for analysis. Initially, separate analyses using the five available outcome measures were conducted. Each

analysis generated R^2 statistics, which indicated how well the cognitive complexity source codes predicted task difficulty or discrimination. With perfect consistency across grade levels, courses, and content areas, R^2 was highest when predicting p-values as measures of task difficulty and point-biserial correlations as measures of task discrimination. This pattern of results is apparent, for example, by comparing the R^2 values in Table 3.1 (p-values) to those in Table A.1 (IRT B values) in Appendix A. Complete results for p-values and point-biserial correlations are reported in the following sections. Initial results for all other outcome variables are included in Appendix A.

Predicting Mathematics Task Difficulty

Appendix B provides the complete set of tables showing the descriptive statistics for the cognitive complexity source codes used to predict mathematics task p-values (Tables B.1–B.12). As indicated by the means and frequencies, most items were coded as having low to moderate cognitive complexity. As shown throughout Tables B.1–B.12, many of the cognitive complexity source codes were significantly correlated with p-values, but the magnitudes of those correlations were generally low. Relatively high correlations were observed at grade 4, where Mathematical Content, Mathematical Practices, and Response Mode each explained approximately 20% of the variance in p-values (Table B.2). At grade 3, those source codes explained between 9% and 13% of the variance in p-values (Table B.1). Mathematical Content explained 10% of the variance at grade 7 (Table B.5). In all other cases, the variance explained was less than 10%. In all, these results suggested that the cognitive complexity source codes, especially Mathematical Content, Mathematical Practices, and Response Mode, were likely to be useful predictors of p-values in the regression tree analyses for certain grades and subjects.

Table 3.1 shows the empirical weights derived from the conditional random forests used to predict p-values from the cognitive complexity source codes. These weights were calculated by scaling the importance statistics to sum to 1.0. To interpret these values, one must first consider the conditional random forest R^2 reported in the rightmost column of Table 3.1. In grades or subjects where the conditional random forest R^2 was low (e.g., below .10), the weights should be interpreted with caution, because the cognitive complexity source codes were not effective predictors of p-values. Low R^2 could reflect several factors, including low variance in the predictors (e.g., many items with low cognitive complexity), low variance in the outcome (e.g., many items that are very difficult), or lack of association between cognitive complexity and p-values. As shown in Table 3.2, the standard deviation of the p-values was greatest for tasks at the lower grade levels. With more variance to explain, there was greater potential to predict p-values at the lower grades.

Table 3.1. *Empirical Weights for Mathematics Based on Analysis of P-Values*

							R^2	
	N	MC	MP	SM	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	324	.39	.32	.04	.20	.05	.204	.131	.161
Grade 4	328	.31	.26	.15	.28	.01	.350	.326	.322
Grade 5	279	.00	.48	.02	.46	.04	.117	.081	.101
Grade 6	241	.28	.32	.07	.21	.12	.066	.037	.062
Grade 7	256	.65	.21	.04	.04	.06	.124	.066	.099
Grade 8	258	.26	.19	.17	.16	.22	.035	.000	.075
Algebra I	223	.19	.17	.64	.00	.00	.086	.000	.033
Algebra II	239	.50	.15	.20	.14	.00	.056	.000	.045

Table 3.1. *Empirical Weights for Mathematics Based on Analysis of P-Values*

							R^2	
Geometry	226	.18	.46	.00	.32	.04	.087	.064	.090
Integrated Math 1	80	.00	.21	.08	.00	.71	.000	.000	.000
Integrated Math 2	68	.00	.05	.06	.00	.89	.000	.000	.017
Integrated Math 3	67	.00	1.00	.00	.00	.00	.000	.000	.000

Table 3.2. *Summary Statistics for Mathematics P-Values*

	N	Mean	SD	Min.	25th %ile	Median	75th %ile	Max
Grade 3	324	.44	.23	.01	.26	.43	.60	.98
Grade 4	328	.42	.20	.00	.27	.41	.57	.89
Grade 5	279	.37	.20	.03	.21	.34	.51	.87
Grade 6	241	.32	.20	.00	.17	.28	.47	.91
Grade 7	256	.25	.18	.01	.10	.21	.36	.82
Grade 8	258	.24	.19	.02	.10	.19	.35	.88
Algebra I	223	.19	.17	.00	.04	.16	.30	.71
Algebra II	239	.21	.16	.00	.08	.19	.31	.77
Geometry	226	.20	.18	.00	.06	.15	.30	.88
Integrated Math 1	80	.21	.15	.00	.08	.19	.32	.65
Integrated Math 2	68	.21	.17	.00	.05	.17	.31	.70
Integrated Math 3	67	.20	.14	.00	.07	.20	.30	.69

Using a criterion of $R^2 \geq .10$, the empirical weights for grade 3, grade 4, grade 5, and grade 7 ($R^2 = .099$) may be interpreted as indictors of the relative importance of the cognitive complexity source codes as predictors of task difficulty. The weights varied across grade levels, but Mathematical Content, Mathematical Practices, and Response Mode tended to be the most important. Their average weights across the four grades were .34, .32, and .25, respectively. Weights for stimulus material and processing demands were generally below .10. Recall that the judgmental weights for the mathematics source codes were 0.3 (Mathematical Content), 0.4 (Mathematical Practices), 0.1 (Stimulus Material), 0.1 (Response Mode) and 0.1 (Processing Demands). The following statements summarize the comparison of judgmental and empirical weights.

- The judgmental and empirical weights for **Mathematical Content** were similar.
- The empirical weights for **Mathematical Practices** were lower than the corresponding judgmental weights.
- The judgmental and empirical weights for **Stimulus Material** were similar.
- The empirical weights for **Response Mode** were higher than the corresponding judgmental weights.
- The judgmental and empirical weights for **Processing Demands** were similar.

Compared to the judgmental weights, the empirical weights suggested higher importance for Response Mode and lower importance for Mathematical Practices. Otherwise, the judgmental weights were similar to the empirical weights.

We used the empirical weights to calculate a new measure of overall cognitive complexity for each task. The descriptive statistics tables in Appendix B provide distributional information about these measures, including their correlations with p-values. Because these measures were derived from regression trees predicting task difficulty, they were expected to correlate more strongly with task difficulty than the judgmental cognitive complexity measures. This pattern of results is apparent in Tables B.1, B.2, B.3, and B.5, which correspond to grades 3, 4, 5, and 7, but the differences tended to be small (.02–.04). As indicated by the frequencies in these tables, the judgmental and empirical weights resulted in similar distributions of overall cognitive complexity (i.e., a similar number of tasks classified as low, moderate, and high). Grade 7 results were somewhat different, with more tasks being classified as moderately complex by the empirical measures. At grade 7, the empirical weight for Mathematical Content was quite high (.65), so the distribution of overall cognitive complexity measures matched perfectly with the Mathematical Content distribution.

As noted in the method section, a single conditional tree was also fit to the data. The conditional tree R^2 in Table 3.1 reflects the predictive accuracy of that single conditional tree, and the conditional tree cross-validation R^2 is the corresponding 10-fold cross-validation measure of predictive accuracy. These conditional trees may support recommendations about how to manipulate cognitive complexity source codes to impact task difficulty (see Figures D.1 through D.12 in Appendix D). Note, however, that many of the cross-validation R^2 values were nearly zero, which indicates that those trees would not be useful for predicting p-values for new tasks.

To provide a point of reference for interpreting the magnitude of the R^2 values in Table 3.1, additional conditional trees and random forests were fit to the data. In these analyses, cognitive complexity and available metadata variables were used to predict task difficulty. Table 3.3 shows the importance statistics and R^2 values for individual conditional trees and conditional random forests fit to the data. The corresponding conditional trees are provided for reference in Appendix F (Figures F.1 through F.12). The R^2 values listed in Table 3.3 are substantially higher than those listed in Table 3.1, which reveals that cognitive complexity accounted for a fraction of the variance in p-values that could be explained by all available variables. The difference was smallest at grade 4, where the cognitive complexity source codes alone accounted to 32.2% of the variance. Including the overall cognitive complexity and metadata increased the percentage of variance explained to 49.8%. These R^2 values are consistent with findings in other item difficulty modeling studies, where R^2 values can range from 5% to 60% of variance explained (Ferrara, Svetina, Skucha, & Murphy, 2011). As shown in Table 3.3, several variables were important predictors of p-values across grades or courses, especially variables related to item type and content alignment. Note that the importance of mode (computer based testing versus paper based testing) was nearly zero across all mathematics assessments, which suggests that, controlling for other task variables, tasks administered in different modes were similarly difficult. In this study, however, only one version of each task was analyzed (usually the computer based testing version). A more rigorous comparability study would examine performance on the same tasks administered in different modes.

Table 3.3. *Importance Statistics for Predictors of Mathematics Task P-Values*

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Algebra II	Geometry	Int. 1	Int. 2	Int. 3	Mean
Math Content	28	40	3	11	8	2	1	3	2	0	0	0	8
Math Practices	16	36	13	3	2	2	1	1	6	0	1	0	7
Stimulus Material	3	5	1	0	0	2	2	1	1	0	0	0	1
Response Mode	5	19	6	1	0	1	0	1	4	0	0	0	3
Processing Demands	4	2	6	6	0	2	0	0	1	0	5	0	2
Overall Cognitive Complexity	35	100	10	22	3	7	1	1	7	0	0	0	15
Mode (CBT or PBT)	0	0	0	1	0	0	0	0	0	2	10	14	2
Number of Score Categories	19	51	6	2	6	1	0	1	5	0	0	1	8
Component (EOY or PBA)	0	5	2	9	1	2	0	0	1	0	0	1	2
PARCC Item Type	26	72	19	12	2	6	0	1	4	1	0	0	12
Response Type	7	20	36	40	15	16	22	23	38	19	12	23	23
Interaction Type	12	13	38	32	9	13	16	21	31	21	10	21	20
TEI Type	46	71	100	100	100	100	100	100	100	100	100	100	93
PARCC Evidence Statement 1	100	61	40	69	30	58	18	26	15	18	7	2	37
PARCC Sub-claim	6	48	7	10	1	3	0	2	2	0	1	1	7
PARCC Task Model 1	84	45	36	58	23	53	14	21	12	14	7	9	31
Companion Materials	0	0	0	5	1	2	1	1	4	3	1	0	2
PARCC Number of Points	18	16	3	2	4	2	1	1	12	4	0	1	5
Calculator Code	0	0	0	4	1	2	0	1	4	1	0	0	1
PARCC Stimulus Identifier	1	1	0	1	0	2	0	0	0	0	0	0	0
CCSS Identifier 1	54	64	24	31	17	33	5	7	12	7	16	0	22
CCSS Identifier 2	1	0	2	0	0	0	1	2	0	1	0	1	1
R^2 Cond. Tree	.372	.483	.484	.430	.494	.465	.555	.400	.413	.627	.471	.141	.445
R^2 Cond. Tree Cross-Val.	.199	.354	.168	.044	.287	.252	.406	.206	.263	.000	.265	.000	.204
R^2 Cond. Random Forest	.402	.498	.385	.328	.436	.466	.472	.328	.361	.332	.259	.132	.367

Predicting Mathematics Task Discrimination

The analyses to predict task discrimination should be considered exploratory because there was no *a priori* expectation that the cognitive complexity source codes should correlate with task discrimination. Appendix B provides the complete set of tables showing the descriptive statistics for the cognitive complexity source codes used to predict mathematics task point-biserial correlations (Tables B.13–B.24). Many of the cognitive complexity source codes were significantly correlated with point-biserial correlations, but the magnitudes of those correlations were generally low. Across the grades and subjects, Response Mode tended to correlate most highly with task discrimination, with a maximum of .505 for Integrated Math 1. After Response Mode, Processing Demands correlated significantly most often with task discrimination. These results suggested that the cognitive complexity source codes, especially Response Mode and Processing Demands, were likely to be useful predictors of task discrimination in the regression tree analyses for certain grades and subjects. Generally, Response Mode should be expected to correlate with task discrimination because tasks with moderate or high Response Mode complexity are likely to be constructed-response items with polytomous scoring models. With a greater number of possible scores, polytomous tasks tend have greater variance and greater potential to discriminate between examinees of low and high ability.

Table 3.4 shows the "empirical weights" derived from the conditional random forests used to predict point-biserial correlations from the cognitive complexity source codes. Looking first at the conditional random forest R^2 values, it is apparent that the cognitive complexity source codes were not useful predictors of task discrimination for all grades and subjects. R^2 was approximately .10 or higher for grade 3, grade 5, grade 6, grade 7, grade 8, and Integrated Math 1. There was somewhat more variance in the outcome variable at certain grades and subjects (Table 3.5), but it did not seem to be associated with higher R^2 values.

Table 3.4. *Empirical Weights for Mathematics Based on Analysis of Point-Biserial Correlations*

							R^2	
	N	MC	MP	SM	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	324	.00	.01	.57	.09	.33	.189	.174	.163
Grade 4	328	.00	.00	.13	.69	.17	.078	.070	.058
Grade 5	279	.18	.04	.04	.66	.08	.121	.074	.096
Grade 6	241	.09	.02	.00	.70	.20	.148	.138	.133
Grade 7	256	.04	.00	.00	.95	.01	.173	.163	.140
Grade 8	258	.00	.01	.09	.84	.05	.132	.082	.099
Algebra I	223	.03	.13	.20	.64	.01	.131	.044	.051
Algebra II	239	.02	.03	.02	.86	.06	.088	.061	.043
Geometry	226	.08	.05	.07	.80	.00	.113	.065	.067
Integrated Math 1	80	.08	.00	.00	.92	.00	.239	.143	.110
Integrated Math 2	68	.00	.00	.52	.48	.00	.000	.000	.000
Integrated Math 3	67	.00	.00	.00	.00	.00	.000	.000	.000

Table 3.5. *Summary Statistics for Mathematics Point-Biserial Correlations*

	N	Mean	SD	Min.	25th %ile	Median	75th %ile	Max
Grade 3	324	.47	.14	.03	.40	.49	.56	.82
Grade 4	328	.49	.12	.06	.42	.49	.57	.82
Grade 5	279	.45	.13	-.04	.38	.46	.54	.85
Grade 6	241	.45	.14	.01	.37	.46	.56	.80
Grade 7	256	.45	.16	.01	.33	.45	.57	.80
Grade 8	258	.43	.14	.08	.33	.42	.53	.79
Algebra I	223	.31	.16	-.02	.19	.29	.40	.85
Algebra II	239	.37	.16	-.03	.27	.38	.48	.90
Geometry	226	.41	.15	.10	.30	.40	.51	.87
Integrated Math 1	80	.39	.18	.01	.26	.39	.52	.82
Integrated Math 2	68	.30	.15	-.03	.20	.28	.38	.69
Integrated Math 3	67	.34	.17	-.04	.19	.35	.45	.66

Using a criterion of $R^2 \geq .10$, the empirical weights for grade 3, grade 5 ($R^2 = .096$), grade 6, grade 7, grade 8 ($R^2 = .099$), and Integrated Math 1 may be interpreted as indictors of the relative importance of the cognitive complexity source codes as predictors of task discrimination. As expected based on the correlations in Appendix B, the weight for Response Mode was consistently the highest. Considering only the six aforementioned grades and subjects, the average weight for Response Mode was .69. The average weights for Stimulus Material and Processing Demands were approximately .10. When the empirical weights were used to calculate overall cognitive complexity, the resulting cognitive complexity measures were nearly identical to the Response Mode codes, as is apparent in the frequencies shown in Tables B.13 through B.24 in Appendix B. Because the cognitive complexity source codes were not necessarily expected to correlate with task discrimination, the empirical weights should not be compared to the judgmental weights.

Individual conditional trees were also fit to the data (Figures D.13 through D.24 in Appendix D). The root node (i.e., the first split) in nearly all of those trees involves splitting on Response Mode. The conditional tree and cross-validation R^2 in Table 3.4 reflect the predictive accuracy of the trees shown in Appendix D. Recall that such trees should not be used to predict the discrimination of new items if the cross-validation R^2 is low.

Additional regression trees were fit to the data to examine how much better the prediction of task discrimination could be when including metadata as predictor variables. Table 3.6 shows the importance statistics and R^2 values for individual conditional trees and conditional random forests fit to the data. The corresponding conditional trees are provided for reference in Appendix F (Figures F.13 through F.24). The R^2 values listed in Table 3.6 are substantially higher than those listed in Table 3.4, with an average difference of .25. As expected, given that polytomous items tend to be more discriminating, the number of score categories was the most important predictor of discrimination (Table 3.6). Some other variables associated with item type (e.g., TEI Type, Response Type, Interaction Type) and content alignment (e.g., PARCC Evidence Statement, CCSS Identifier) were also important.

Table 3.6. *Importance Statistics for Predictors of Mathematics Task Point-Biserial Correlations*

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Algebra II	Geometry	Int. 1	Int. 2	Int. 3	Mean
Math Content	0	1	1	3	1	0	1	0	4	3	2	0	1
Math Practices	0	4	0	0	2	0	5	5	4	0	9	0	2
Stimulus Material	16	1	1	0	1	3	1	0	0	0	13	0	3
Response Mode	0	3	6	9	8	5	1	3	4	28	13	0	7
Processing Demands	9	1	1	3	0	0	0	2	0	0	0	0	1
Overall Cognitive Complexity	0	1	1	2	1	0	0	0	6	0	4	0	1
Mode (CBT or PBT)	0	0	0	0	0	0	0	0	0	0	6	0	1
Number of Score Categories	46	84	63	100	78	100	100	100	100	84	20	52	77
Component (EOY or PBA)	1	1	4	3	0	2	22	1	0	4	7	0	4
PARCC Item Type	10	34	24	11	20	29	54	18	55	81	6	0	29
Response Type	12	29	49	39	100	42	27	84	72	100	58	8	52
Interaction Type	10	20	34	29	76	34	23	64	65	67	98	8	44
TEI Type	53	100	76	23	39	61	25	56	76	61	89	75	61
PARCC Evidence Statement 1	100	58	100	42	18	44	59	99	22	6	97	100	62
PARCC Sub-claim	18	22	11	8	7	20	44	12	19	49	21	0	19
PARCC Task Model 1	79	47	71	32	13	34	55	93	18	6	100	88	53
Companion Materials	0	0	0	17	2	21	0	1	3	0	6	1	4
PARCC Number of Points	55	66	60	100	66	89	62	44	73	25	9	66	60
Calculator Code	0	0	0	15	2	20	0	1	0	3	0	0	3
PARCC Stimulus Identifier	6	3	3	2	11	6	0	0	2	0	0	1	3
CCSS Identifier 1	52	26	65	27	10	24	20	68	4	7	0	77	32
CCSS Identifier 2	4	4	2	3	1	5	4	26	2	4	0	20	6
R^2 Cond. Tree	.525	.357	.497	.407	.485	.518	.342	.458	.442	.328	.000	.224	.382
R^2 Cond. Tree Cross-Val.	.380	.311	.321	.334	.399	.315	.290	.180	.275	.300	.000	.000	.259
R^2 Cond. Random Forest	.444	.397	.418	.383	.445	.431	.338	.330	.354	.314	.000	.052	.326

Predicting ELA/Literacy Task Difficulty

Appendix C provides the complete set of tables showing the descriptive statistics for the cognitive complexity source codes used to predict ELA literacy task p-values (Tables C.1–C.9). As would be expected, average Text Complexity tended to increase with grade level. In terms of Command of Textual Evidence, tasks were most often coded as moderately complex. Response Mode and Processing Demands were mostly split between low and moderate complexity. In many cases, the cognitive complexity source codes were significantly correlated with p-values. The only notable trend in the correlations was that correlations between Text Complexity and p-values tended to be the smallest in magnitude (average of -.11). The other correlations were similar on average across the grades (-.25 for Command of Textual Evidence, -.24 for Response Mode, and -.21 for Processing Demands). In some cases, a single source code explained more than 10% of the variance in p-values (e.g., Response Mode at grades 5, 6, 7, and 9, Command of Textual Evidence at grades 6 and 10, and Processing Demands at grades 7 and 9). In all, descriptive statistics suggested that the cognitive complexity source codes were likely to be useful predictors of p-values in the regression tree analyses for certain grades and subjects.

Table 3.7 shows the "empirical weights" derived from the conditional random forests used to predict p-values from the cognitive complexity source codes. The conditional tree R^2 values were quite low at several grades (e.g., 3, 4, 10, and 11), so the associated weights should be interpreted with caution. Table 3.8 provides descriptive statistics for the outcome variable. Note that the standard deviation of the p-values tended to increase as grade increased, but larger standard deviations were not necessarily associated with higher R^2.

Table 3.7. *Empirical Weights for ELA/Literacy Based on Analysis of P-Values Values*

						R^2	
	N	TC	CTE	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	156	.00	.00	.00	1.00	.049	.007	.000
Grade 4	222	.00	.37	.00	.63	.071	.000	.025
Grade 5	169	.05	.18	.64	.13	.157	.080	.124
Grade 6	197	.02	.20	.56	.22	.215	.141	.159
Grade 7	162	.03	.06	.11	.80	.266	.244	.232
Grade 8	167	.55	.14	.23	.09	.117	.000	.092
Grade 9	109	.04	.06	.29	.62	.280	.188	.230
Grade 10	71	.62	.04	.34	.00	.000	.000	.000
Grade 11	192	.16	.55	.19	.09	.096	.023	.038

Table 3.8. *Summary Statistics for ELA/Literacy P-Values*

	N	Mean	SD	Min.	25th %ile	Median	75th %ile	Max
Grade 3	156	.42	.13	.04	.33	.43	.52	.66
Grade 4	222	.46	.14	.16	.37	.46	.53	.83
Grade 5	169	.43	.15	.13	.35	.44	.51	.82
Grade 6	197	.47	.14	.18	.38	.46	.54	.85
Grade 7	162	.47	.14	.18	.38	.47	.55	.83
Grade 8	167	.45	.17	.09	.35	.42	.53	.86
Grade 9	109	.46	.17	.07	.35	.45	.53	.87
Grade 10	71	.46	.17	.01	.35	.46	.56	.89
Grade 11	192	.44	.18	-.17	.35	.44	.52	.86

Using a criterion of $R^2 \geq .10$, the empirical weights for grade 5, grade 6, grade 7, and grade 9 may be interpreted as indictors of the relative importance of the cognitive complexity source codes as predictors of task difficulty. The weights varied across grade levels, but Response Mode and Processing Demands tended to be the most important. Their average weights across the four grades were .40 and .44, respectively. Weights for Text Complexity and Command of Textual Evidence were generally below .10. Recall the judgmental weights for the ELA literacy source codes were 0.5 (text complexity), 0.2 (command of textual evidence), 0.2 (response mode), and 0.1 (processing demands). The following statements summarize the comparison of judgmental and empirical weights.

- The empirical weights for **Text Complexity** were much lower than the corresponding judgmental weights.
- The judgmental and empirical weights for **Command of Textual Evidence** were similar.
- The empirical weights for **Response Mode** were higher than the corresponding judgmental weights.
- The empirical weights for **Processing Demands** were higher than the corresponding judgmental weights.

Compared to the judgmental weights, the empirical weights suggested higher importance for Response Mode and Processing Demands and much lower importance for Text Complexity.

We used the empirical weights to calculate a new measures of overall cognitive complexity. The descriptive statistics tables in Appendix C provide distributional information about these measures, including their correlations with p-values. As expected, the empirical measures were more highly correlated with task difficulty than the judgmental measures. This pattern of results is apparent in Tables C.3, C.4, C.5, and C.7, which correspond to grades 5, 6, 7, and 9, and some of the differences were large (as high as .26 at grade 7). Such results suggest that the ELA literacy judgmental weights were not very useful for generating an overall cognitive complexity measure reflective of task difficulty. The empirical weights, having been derived from a statistical model for predicting task difficulty, were better in this regard. As shown in Tables C.3, C.4, C.5, and C.7, the frequency distributions of the empirical cognitive complexity measures were similar to the Response Mode and Processing Demands distributions, which would be expected given their high empirical weights.

Each conditional tree and cross-validation R^2 in Table 3.7 reflects the predictive accuracy of a single conditional tree fit to the data. These conditional trees may support recommendations about how to manipulate cognitive complexity source codes to impact task difficulty (see Figures E.1 through E.9 in Appendix E). Note, however, that many of the cross-validation R^2 values were nearly zero, which indicates that these trees such trees were not useful predictors of p-values.

To provide a point of reference for interpreting the magnitude of the R^2 values in Table 3.7, overall cognitive complexity and available metadata variables were added to the conditional trees and random forest analyses. The importance statistics and R^2 values for these trees are shown in Table 3.9. The corresponding conditional trees are provided for reference in Appendix G (Figures G.1 through G.9). With the exception of grade 9, the R^2 values listed in Table 3.9 are higher than those listed in Table 3.7, but only by an average of .06. At several grade levels, Processing Demands were relatively important predictors of p-values compared to the metadata variables. At grades 5, 6, and 7, mode (i.e., administration on paper or online) and component (performance-based assessment or end-of-year) were important predictors of p-values. Other important predictors were related to item type, content alignment, and item set or passage. Note that component is very likely correlated with item type.

Table 3.9. *Importance Statistics for Predictors of ELA/Literacy Task P-Values*

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Mean
Text Complexity	6	0	3	2	8	51	3	7	1	9
Command of Textual Evidence	5	12	10	40	8	7	4	0	28	13
Response Mode	0	0	27	83	11	12	15	13	16	20
Processing Demands	73	29	14	26	100	5	100	0	0	39
Overall Cognitive Complexity	25	1	28	6	9	100	13	9	19	23
Mode (CBT or PBT)	0	1	0	3	0	0	0	0	0	1
Number of Score Categories	0	0	71	79	88	3	3	0	15	29
Component (EOY or PBA)	1	0	2	21	3	5	0	0	12	5
PARCC Item Type	4	0	58	80	61	7	0	0	21	26
Response Type	5	0	45	79	59	37	5	12	35	31
Interaction Type	7	1	48	55	47	32	11	7	26	26
TEI Type	0	15	100	94	49	38	26	8	82	46
Task Type	0	6	1	19	1	5	11	1	11	6
PARCC Evidence Statement 1	69	21	2	6	1	36	0	36	0	19
PARCC Evidence Statement 2	100	16	28	43	39	44	6	61	61	44
PARCC Evidence Statement 3	86	33	16	62	22	7	12	3	55	33
PARCC Sub-claim	88	100	65	32	47	30	15	6	88	52
PARCC Task Model 1	11	1	1	21	3	10	6	2	6	7
1st Passage Identifier	81	41	0	100	64	82	0	82	98	61
Media Type	1	3	0	11	1	0	0	0	0	2
PARCC Number of Points	2	0	16	16	29	2	1	0	17	9
Set Identifier	70	35	0	75	53	62	0	100	100	55
Passage Word Count	71	31	0	68	46	56	0	58	98	48
Passage Type	5	53	0	0	4	23	0	17	0	11

Table 3.9. *Importance Statistics for Predictors of ELA/Literacy Task P-Values*

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Mean
PARCC Stimulus Identifier	54	21	2	38	33	48	0	45	58	33
R^2 Cond. Tree	.364	.282	.397	.508	.500	.468	.352	.505	.351	.414
R^2 Cond. Tree Cross-Val.	.000	.051	.104	.098	.204	.183	.029	.000	.000	.074
R^2 Cond. Random Forest	.053	.122	.172	.315	.369	.196	.144	.000	.104	.164

Predicting ELA/Literacy Task Discrimination

As a reminder, the analyses to predict task discrimination should be considered exploratory. Appendix C provides the complete set of tables showing the descriptive statistics for the cognitive complexity source codes used to predict ELA literacy task point-biserial correlations (Tables C.10–C.18). Response Mode was most often strongly positively correlated with task discrimination (at every grade except grade 3), but Response Mode should be expected to correlate with task discrimination because tasks with moderate or high Response Mode complexity are likely to have polytomous scoring models. Command of Textual Evidence was also positively correlated with task discrimination at several grade levels (4, 5, 6, 8, 9, and 11). More often than not, Text Complexity and Processing Demands were negatively correlated with task discrimination. That is, tasks with higher Text Complexity or Processing Demands tended to be less discriminating. It may be that tasks with high Processing Demands tended to be very difficult, and therefore have little score variance and little potential to discriminate between students of low and high ability. This explanation could not apply to Text Complexity because it correlated so weakly with task difficulty.

Table 3.10 shows the empirical weights derived from the conditional random forests used to predict point-biserial correlations from the cognitive complexity source codes. Looking first at the R^2 values, the conditional random forests were successful in predicting point-biserial correlations at all grades except 3 and 10. There was somewhat more variance in the outcome variable at higher grade levels (Table 3.11), but it was not associated with higher R^2 values.

Table 3.10. *Empirical Weights for ELA/Literacy Based on Analysis of Point-Biserial Correlations*

						R^2		
	N	TC	CTE	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	156	.48	.00	.00	.52	.000	.000	.000
Grade 4	222	.00	.06	.94	.00	.409	.397	.381
Grade 5	169	.00	.00	.88	.12	.408	.393	.377
Grade 6	197	.04	.03	.91	.02	.458	.455	.411
Grade 7	162	.00	.01	.09	.90	.422	.407	.263
Grade 8	167	.01	.03	.88	.07	.469	.465	.412
Grade 9	109	.02	.07	.87	.04	.557	.549	.462
Grade 10	71	.23	.09	.51	.18	.018	.000	.000

Table 3.10. *Empirical Weights for ELA/Literacy Based on Analysis of Point-Biserial Correlations*

						R^2		
	N	TC	CTE	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 11	192	.01	.20	.79	.00	.354	.347	.246

Table 3.11. *Summary Statistics for ELA/Literacy Point-Biserial Correlations*

	N	Mean	SD	Min.	25th %ile	Median	75th %ile	Max
Grade 3	156	.42	.13	.04	.33	.43	.52	.66
Grade 4	222	.46	.14	.16	.37	.46	.53	.83
Grade 5	169	.43	.15	.13	.35	.44	.51	.82
Grade 6	197	.47	.14	.18	.38	.46	.54	.85
Grade 7	162	.47	.14	.18	.38	.47	.55	.83
Grade 8	167	.45	.17	.09	.35	.42	.53	.86
Grade 9	109	.46	.17	.07	.35	.45	.53	.87
Grade 10	71	.46	.17	.01	.35	.46	.56	.89
Grade 11	192	.44	.18	-.17	.35	.44	.52	.86

Using a criterion of $R^2 \geq .10$, the empirical weights for grades 4, 5, 6, 7, 8, 9, and 11 may be interpreted as indictors of the relative importance of the cognitive complexity source codes as predictors of task discrimination. With an average weight of .77, Response Mode was the most important predictor of task discrimination. With such high weights applied to Response Mode, the distributions of empirical measures of overall cognitive complexity were generally identical to Response Mode, as is apparent in Tables C.10 through C.18. The only notable exception to this trend in results was grade 7, where Processing Demands was weighted .90. This may have occurred because of low variance in Response Mode at grade 7, where 80% of the items were coded as having moderate Response Mode complexity. Recall that the empirical weights should not be compared to the judgmental weights.

Individual conditional trees were also fit to the data (Figures E.10 through E.18 in Appendix E). The root node (i.e., the first split) in nearly all of those trees involved splitting on Response Mode. The conditional tree and cross-validation R^2 in Table 3.10 reflect the predictive accuracy of the trees shown in Appendix E. Recall that such trees should only be used to predict the discrimination of new items if the cross-validation R^2 is non-negligible.

Additional regression trees were fit to the data to examine how much better the prediction of task discrimination could be when including metadata as predictor variables. Table 3.12 shows the importance statistics and R^2 values for individual conditional trees and conditional random forests fit to the data. The corresponding conditional trees are provided for reference in Appendix G (Figures G.10 through G.18). The R^2 values listed in Table 3.12 are slightly higher than those listed in Table 3.10, with an average difference of .09. This suggests that the cognitive complexity source codes (Response Mode in particular) accounted for much of the variance in task discrimination that could be explained by all

available variables. Component was consistently an important predictor of task discrimination, but this is likely a reflection of PBA tasks having more possible score points than EOY tasks. Several other variables related to item type were also very important (e.g., PARCC Item Type, Response Type, and Interaction Type). Other important predictors were related to content alignment and item set or passage.

Table 3.12. *Importance Statistics for Predictors of ELA/Literacy Task Point-Biserial Correlations*

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Mean
Text Complexity	5	0	0	1	1	1	1	0	1	1
Command of Textual Evidence	0	2	0	1	0	1	0	1	4	1
Response Mode	0	46	23	33	1	30	28	0	14	19
Processing Demands	37	0	6	1	21	4	1	0	0	8
Overall Cognitive Complexity	0	5	0	0	1	1	0	12	1	2
Mode (CBT or PBT)	3	0	0	2	1	0	0	0	0	1
Number of Score Categories	0	45	54	50	31	46	45	16	61	39
Component (EOY or PBA)	0	4	0	2	5	1	3	17	1	4
PARCC Item Type	0	100	100	100	100	100	100	32	100	81
Response Type	0	90	80	92	88	76	74	52	95	72
Interaction Type	8	60	66	64	63	62	69	30	88	57
TEI Type	0	75	55	56	51	59	50	89	76	57
Task Type	38	6	0	2	4	0	5	28	3	9
PARCC Evidence Statement 1	0	3	0	10	0	0	1	0	3	2
PARCC Evidence Statement 2	100	0	0	3	7	6	8	4	27	17
PARCC Evidence Statement 3	13	8	7	13	2	4	10	7	8	8
PARCC Sub-claim	0	24	26	28	24	25	21	12	31	21
PARCC Task Model 1	41	3	0	1	10	1	3	15	3	9
1st Passage Identifier	0	27	0	26	32	11	6	100	38	27
Media Type	0	0	0	1	0	0	0	0	0	0
PARCC Number of Points	0	35	35	31	32	30	30	26	41	29
Set Identifier	0	23	0	22	24	11	3	80	39	22
Passage Word Count	0	18	0	15	18	8	2	55	21	15
Passage Type	7	1	0	1	1	0	0	2	1	1
PARCC Stimulus Identifier	0	16	0	16	14	5	2	46	23	14
R^2 Cond. Tree	.211	.488	.426	.462	.566	.657	.466	.330	.525	.459
R^2 Cond. Tree Cross-Val.	.000	.370	.409	.454	.413	.463	.537	.000	.411	.340
R^2 Cond. Random Forest	.000	.456	.372	.534	.494	.468	.535	.053	.460	.375

4. ANALYSIS 2 RESULTS

In terms of methodology, Analysis 2 was much like the second portion of Analysis 1 (i.e., predicting an outcome from metadata variables). In Analysis 2, the outcome variable was overall cognitive complexity based on the judgmental weights. The numeric weighted composite was analyzed rather than the rounded ordinal measures (low, moderate, high) because the numeric value had greater variance. Moreover, the numeric measure was presumably more precise. That is, the numeric measure could indicate a difference between a low-moderate complexity task and a moderate-high complexity task, whereas both tasks might have been classified as moderately complex. Results of this analysis could reveal associations between task characteristics and cognitive complexity that could guide task authors to create tasks that accurately target a desired level of cognitive complexity.

Predicting Mathematics Task Cognitive Complexity

Table 3.13 provides descriptive statistics for mathematics task overall cognitive complexity. The average task was in the low to moderate complexity range. Indeed, for all grades and subjects, more than 75% of tasks has cognitive complexity measures below 1.7.

Table 3.13. *Summary Statistics for Mathematics Cognitive Complexity Measures*

	N	Mean	SD	Min.	25th %ile	Median	75th %ile	Max
Grade 3	324	1.47	.38	1.00	1.20	1.40	1.70	2.80
Grade 4	328	1.42	.40	1.00	1.10	1.30	1.70	2.90
Grade 5	279	1.38	.39	1.00	1.10	1.30	1.60	2.90
Grade 6	241	1.33	.38	1.00	1.10	1.20	1.40	2.90
Grade 7	256	1.31	.30	1.00	1.10	1.20	1.40	2.30
Grade 8	258	1.38	.38	1.00	1.10	1.30	1.60	3.00
Algebra I	223	1.32	.33	1.00	1.10	1.20	1.50	2.70
Algebra II	239	1.29	.36	1.00	1.00	1.10	1.40	2.80
Geometry	226	1.40	.41	1.00	1.10	1.30	1.60	2.90
Integrated Math 1	80	1.35	.35	1.00	1.10	1.25	1.50	2.40
Integrated Math 2	68	1.28	.31	1.00	1.00	1.20	1.40	2.40
Integrated Math 3	67	1.28	.28	1.00	1.00	1.30	1.40	2.40

Table 3.14 shows the importance statistics and R^2 values for individual conditional trees and conditional random forests fit to the data. The corresponding conditional trees are provided for reference in Appendix H. Except for the Integrated Mathematics courses, for which there were many fewer tasks to analyze, the conditional random forests accounted for at least 32% of the variance in mathematics task cognitive complexity. The most important predictors of cognitive complexity were variables associated with item type (e.g., PARCC Item Type, Response Type, Interaction Type, TEI type) and content alignment (e.g., PARCC Evidence Statement, PARCC Sub-claim, CCSS Identifier).

Table 3.14. *Importance Statistics for Predictors of Mathematics Task Cognitive Complexity Measures*

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Algebra I	Algebra II	Geometry	Int. 1	Int. 2	Int. 3	Mean
Mode (CBT or PBT)	0	0	0	0	1	0	0	0	0	0	0	0	0
Number of Score Categories	9	40	11	33	38	12	6	32	26	13	11	61	25
Component (EOY or PBA)	0	1	0	1	3	1	0	5	2	0	1	0	1
PARCC Item Type	34	68	40	92	100	75	42	100	87	70	4	0	59
Response Type	82	58	100	53	55	100	100	53	64	78	100	0	70
Interaction Type	100	67	82	41	72	100	49	38	100	100	52	0	67
TEI Type	56	100	55	100	44	57	68	58	55	74	68	0	61
PARCC Evidence Statement 1	55	31	36	34	25	34	33	31	20	16	7	0	27
PARCC Sub-claim	14	22	10	29	65	14	25	37	16	27	7	0	22
PARCC Task Model 1	36	22	26	20	18	16	20	21	13	8	0	0	17
Companion Materials	0	0	0	3	6	3	7	1	5	0	0	30	5
PARCC Number of Points	24	35	9	22	18	16	9	20	18	7	6	0	16
Calculator Code	0	0	0	2	6	2	5	0	0	0	0	0	1
PARCC Stimulus Identifier	1	2	0	3	2	5	0	0	1	0	0	0	1
CCSS Identifier 1	43	36	16	16	8	10	20	22	7	5	0	29	18
CCSS Identifier 2	3	2	6	0	3	0	1	5	2	2	4	100	11
R^2 Cond. Tree	.534	.526	.199	.531	.349	.367	.356	.507	.452	.370	.165	.000	.363
R^2 Cond. Tree Cross-Val.	.336	.234	.131	.389	.290	.195	.121	.295	.220	.333	.000	.000	.212
R^2 Cond. Random Forest	.449	.440	.327	.476	.332	.358	.322	.418	.405	.235	.056	.000	.318

Predicting ELA/Literacy Task Cognitive Complexity

Table 3.15 provides descriptive statistics for ELA literacy task overall cognitive complexity. Across grades, the average task was moderately complex, and average cognitive complexity tended to increase with grade level. The standard deviation of the ELA literacy cognitive complexity measures ranged from .33 to .52.

Table 3.15. *Summary Statistics for ELA/Literacy Cognitive Complexity Measures*

	N	Mean	SD	Min.	25th %ile	Median	75th %ile	Max
Grade 3	156	1.59	.37	1.00	1.30	1.50	1.80	2.90
Grade 4	222	1.65	.33	1.00	1.50	1.70	1.90	2.40
Grade 5	169	1.91	.38	1.00	1.70	1.90	2.10	2.90
Grade 6	197	1.93	.42	1.00	1.70	2.00	2.30	2.90
Grade 7	162	1.98	.52	1.00	1.70	2.00	2.40	3.00
Grade 8	167	1.82	.45	1.00	1.50	1.80	2.20	2.80
Grade 9	109	1.95	.42	1.00	1.70	2.00	2.20	2.90
Grade 10	71	1.95	.44	1.20	1.60	2.00	2.30	2.70
Grade 11	192	2.08	.37	1.20	1.80	2.00	2.40	2.90

Table 3.16 shows the importance statistics and R^2 values for individual conditional trees and conditional random forests fit to the data. The corresponding conditional trees are provided for reference in Appendix I. The conditional random forests accounted for at least 51% of the variance in ELA literacy task cognitive complexity. At every grade level, "1st Passage Identifier" was the most important predictor of cognitive complexity. The next three most important predictors were also associated with the passage (Set Identifier, Passage Word Count, and PARCC Stimulus Identifier). This finding could indicate that passage is a major determiner of cognitive complexity, or it could indicate that the cognitive complexity coders were greatly influenced by the passages. Other, less important predictors were related to item type (e.g., TEI Type, PARCC Item Type).

Table 3.16. *Importance Statistics for Predictors of ELA/Literacy Task Cognitive Complexity Measures*

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Mean
Mode (CBT or PBT)	0	0	0	0	0	0	0	0	0	0
Number of Score Categories	0	32	0	6	2	1	33	0	11	9
Component (EOY or PBA)	1	1	3	1	6	1	1	0	1	2
PARCC Item Type	0	29	11	14	2	5	30	2	11	12
Response Type	1	29	8	9	4	6	18	0	11	10
Interaction Type	0	22	4	7	3	3	18	1	7	7
TEI Type	34	84	84	28	4	17	46	7	55	40
Task Type	3	4	9	12	11	4	30	0	4	8
PARCC Evidence Statement 1	4	21	0	1	1	3	5	11	5	6

Table 3.16. *Importance Statistics for Predictors of ELA/Literacy Task Cognitive Complexity Measures*

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Mean
PARCC Evidence Statement 2	12	18	17	3	1	4	16	3	5	9
PARCC Evidence Statement 3	8	15	6	2	1	2	4	1	9	5
PARCC Sub-claim	14	43	14	10	4	3	12	3	8	13
PARCC Task Model 1	9	4	7	7	11	5	67	0	3	13
1st Passage Identifier	100	100	100	100	100	100	100	100	100	100
Media Type	0	1	0	1	0	0	0	0	0	0
PARCC Number of Points	1	28	7	7	4	4	36	2	22	12
Set Identifier	76	78	79	70	84	85	83	74	72	78
Passage Word Count	54	61	48	54	59	67	57	55	46	56
Passage Type	1	9	0	2	5	1	1	7	1	3
PARCC Stimulus Identifier	46	41	45	40	46	51	33	41	43	43
R^2 Cond. Tree	.687	.767	.665	.767	.695	.814	.473	.499	.685	.672
R^2 Cond. Tree Cross-Val.	.478	.606	.316	.710	.533	.700	.161	.133	.415	.450
R^2 Cond. Random Forest	.529	.694	.582	.713	.642	.821	.510	.482	.666	.626

5. ANALYSIS 3 RESULTS

At the request of PARCC, we surveyed cognitive complexity coders at Pearson to capture their insights and recommendations for PARCC on the cognitive complexity framework, training of complexity coders, the coding process and decision making, and future considerations for cognitive complexity. We also conducted a focus groups with ELA/literacy and mathematics coders to examine more closely responses to selected survey questions.

Survey and Focus Group Interview Results

A total of 48 content developers in ELA and mathematics responded to the survey: approximately 28 in ELA and 17 in mathematics.[5] This represents a response rate of 86% of the total 56 coders who received survey invitations. Pearson content developers who coded PARCC items and tasks for cognitive complexity were asked to participate in the survey. The frequencies of responses to all selected response survey items are summarized in Survey Monkey output, which appears in Appendix K. Raw responses to open ended response survey items for all respondents appear in Appendix L. We summarize responses to selected key survey questions in the following sections and incorporate focus group interview results where appropriate.

[5] Responses to the surveys were anonymous. We used responses to survey question 17, *How clear were the code definitions you were assigned (e.g., for Command of Textual Evidence or for Mathematical Content)?* to identify survey respondents who coded ELA and mathematics items.

Prior Familiarity with the Common Core State Standards and Item Cognitive Complexity

As the responses to survey question 6 indicate, most respondents reported that they were at least slightly or moderately familiar with the Common Core State Standards prior to conducting complexity coding. In fact, only six of 46 respondents (13%) reported no familiarity with the standards in their content area. Familiarity with the Common Core standards was not an eligibility requirement for complexity coding and the standards are not part of the cognitive complexity measures, except for Mathematical Content and Practices. Coders were trained on the Mathematical Content and Practices standards as part of training for coding these two sources of item cognitive complexity.

Survey Question 6

How familiar were you with the Common Core State Standards in your coding area (i.e., ELA or math) before your involvement in PARCC cognitive coding?						
Answer Options	Very Familiar	Moderately Familiar	Slightly Familiar	Not at All Familiar	Rating Average	Response Count
	14	16	10	6	2.17	46
					answered question	46
					skipped question	2

Responses to survey questions 7 and 8 indicate that a considerable number of complexity coders were not at all familiar with Bloom's taxonomy and the Depth of Knowledge frameworks that are used often as indicators of item cognitive complexity. Fourteen of 46 respondents (30%) reported no familiarity with Bloom's taxonomy; 16 of 47 respondents (34%) reported no familiarity with the Depth of Knowledge framework. We posed these questions out of concern that familiarity with these frameworks would interfere with making judgments using PARCC's cognitive complexity framework. In response to survey question 26, six of 23 respondents (26%) reported moderate difficulty with disregarding the Bloom and Depth of Knowledge frameworks while coding PARCC items for cognitive complexity.

Survey Question 7

How familiar were you with Bloom's taxonomy before your involvement in PARCC cognitive complexity coding?						
Answer Options	Very Familiar	Moderately Familiar	Slightly Familiar	Not at all Familiar	Rating Average	Response Count
	22	5	5	14	2.24	46
					answered question	46
					skipped question	2

Survey Question 8

How familiar were you with the Depth of Knowledge framework before your involvement in PARCC cognitive complexity coding?						
Answer Options	Very Familiar	Moderately Familiar	Slightly Familiar	Not at all Familiar	Rating Average	Response Count
	14	12	5	16	2.49	47
					answered question	47
					skipped question	1

Responses to survey question 26 suggest that complexity coders found it slightly or moderately difficult (16 of 23 responses, 70%) to disregard Bloom's taxonomy or the Depth of Knowledge framework and focus on the PARCC cognitive complexity codes.

Survey Question 26

If you answered previously that you have worked with Bloom's taxonomy or the Depth of Knowledge framework before your involvement in PARCC cognitive coding, how difficult was it to disregard them and focus on the PARCC cognitive complexity code definitions?						
Answer Options	Very Difficult	Moderately Difficult	Slightly Difficult	Not at All Difficult	Rating Average	Response Count
	0	6	10	7	3.04	23
					answered question	23
					skipped question	25

In the ELA focus group discussion, coders speculated that having experience in working with Bloom's taxonomy and the Depth of Knowledge framework may lead coders to draw parallels to the PARCC cognitive complexity framework as a way of learning and understanding it. The math coders suggested that thinking about the other frameworks could be helpful to coders who are able to "compartmentalize" their thinking (i.e., think about the other frameworks as a help and focus only on the PARCC framework when making coding decisions); otherwise, experience in working with the other frameworks could interfere with applying the PARCC coding framework. One math coder suggested that it can be helpful to think about the verbs in the Depth of Knowledge framework because the verbs highlight application of knowledge and skills, and that the PARCC framework does not emphasize the use of verbs to represent application of knowledge and skills. Another math coder suggested that it might be helpful to compare and contrast Bloom, Depth of Knowledge, and PARCC cognitive complexity during coder and reviewer training.

A response to survey question 29, *What changes would you recommend to (a) the framework_ (b) the coding process (including training), or (c) how PARCC uses the framework?* is relevant here.

> I think for content specialists to understand the value of cognitive complexity, a foundational knowledge of Bloom's and Webb's DOK is essential. This [PARCC cognitive complexity] is not something that is really an easy thing to learn and understand. And the

two [i.e., Bloom and DOK], while used interchangeably, differ in their applicability to instruction and assessment.

Training

A primary focus in training was to ensure that complexity coders understood clearly the definitions of each source of complexity and the distinctions among high, medium, and low levels of complexity for each source. The clarity of their understandings should have been influenced by the definitions and distinctions as they are written in PARCC documents (e.g., the original cognitive complexity documents from October 2012, complexity training slides) and the execution of the training on the definitions and distinctions. Responses to survey questions 17 and 18 on this topic appear in the tables below. The responses to survey question 17 indicate that most of the coders found that the definitions were moderately or very clear for the three ELA sources (e.g., 19 or 20 of the respondents in ELA) and of the five mathematics sources (e.g., 30 of 31 respondents regarding Mathematical Content, 25 of 29 respondents regarding Stimulus Material).

Survey Question 17

How clear were the code definitions you were assigned (e.g., for Command of Textual Evidence or for Mathematical Content)?					
Answer Options	**Very Clear**	**Moderately Clear**	**Slightly Clear**	**Not at All Clear**	**Response Count**
ELA: Command of Textual Evidence	11	8	1	0	20
ELA: Response Mode	11	9	0	0	20
ELA: Processing Demands	5	14	1	0	20
Math: Mathematical Content	13	17	1	0	31
Math: Mathematical Practices	12	14	3	0	29
Math: Stimulus Material	11	13	3	1	28
Math: Response Mode	12	13	4	0	29
Math: Processing Demands	9	15	4	0	28

	Question Totals
If you responded Slightly or Not at All, please explain in detail.	7
answered question	**42**
skipped question	**6**

Responses to the survey question 18 regarding the clarity of distinctions among high, medium, and low complexity were similar. Most respondents in ELA and mathematics reported that the distinctions were moderately to very clear.

Survey Question 18

How clear were the distinctions between high, medium, and low levels of complexity for the code definitions you were assigned (e.g., for Command of Textual Evidence or for Mathematical Content)?

Answer Options	Very Clear	Moderately Clear	Slightly Clear	Not at All Clear	Response Count
ELA: Command of Textual Evidence	8	10	2	0	20
ELA: Response Mode	10	9	1	0	20
ELA: Processing Demands	4	14	2	0	20
Math: Mathematical Content	8	20	1	0	29
Math: Mathematical Practices	8	19	2	0	29
Math: Stimulus Material	7	18	3	0	28
Math: Response Mode	9	17	3	0	29
Math: Processing Demands	5	19	3	0	27

	Question Totals
If you responded Slightly Or Not at All, please explain in detail.	6
answered question	41
skipped question	7

Responses to survey questions 21 and 22 provide guidance on clarifying the definitions of Mathematical Content and Practices, which coders reported to be particularly challenging during coder training session. Some salient recommendations include:

- Provide more examples of items to illustrate the definitions and distinctions among high, medium, and low complexity and give trainees more time to collaborate during training sessions and more time to practice making coding decisions.
- Help coders understand differences between item difficulty and item cognitive complexity.
- Clarify the definitions.

In the focus group discussions, mathematics coders also recommended distinguishing what may seem cognitively complex to adult content experts from what is cognitively complex to students.

Coding Process and Decisions

We posed questions to the complexity coders about how difficult they perceived the process of reviewing items and tasks and coding them for cognitive complexity. The perceived difficulty they reported provides additional information about the clarity of definitions, the effectiveness of training, the cognitive load[6] of the judgmental task, and other concerns.

In phase 1 of PARCC item and task development, Pearson complexity coders were trained to make judgments about only one source of complexity (e.g., Command of Textual Evidence in ELA, Mathematical Practices). Psychometricians then applied judgmental weights (determined by PARCC) to each individual complexity source judgment (including Text Complexity in ELA, which was evaluated in a separate process) to create an overall measure of cognitive complexity for each item. (We used these individual complexity source judgments in Analysis 1.) In phase 2 of item and task development, Pearson complexity coders were trained to make judgments on the three ELA sources or five mathematics sources and then combine those judgments holistically into an overall judgmental measure of item cognitive complexity. In order to evaluate the cognitive load of this more involved complexity judgment, we asked complexity coders directly about the difficulty they perceived in making this judgment. As the response frequencies for survey question 23 indicate, fewer than half of the ELA and mathematics coders responded. Of those who did, most found the task slightly or moderately difficult. Few reported that the task was easy; one math coder reported that the task was very difficult. Based on these limited responses, it appears that the holistic approach to judging overall cognitive complexity of PARCC items may be a reasonable task for content experts to undertake. Rater accuracy rates (i.e., agreement with expert judgments of cognitive complexity) from coder training in summer 2013 bear this out; most agreement rates in ELA/literacy and mathematics were in the range 70-100% (with some exceptions; see Ferrara et al., 2014).

Survey Question 23

FOR PHASE 2 CODERS: How difficult was it to combine your judgments about the:

Answer Options	Very Difficult	Moderately Difficult	Slightly Difficult	Not at All Difficult	Response Count
three ELA complexity codes into a single complexity judgment?	0	1	7	2	10
five mathematical complexity codes into a single complexity judgment?	1	5	10	0	16

[6] Defined as the amount of effort required in working memory.

FOR PHASE 2 CODERS: How difficult was it to combine your judgments about the:					
Answer Options	Very Difficult	Moderately Difficult	Slightly Difficult	Not at All Difficult	Response Count / Question Totals
If you responded Very or Moderately Difficult, please explain in detail.					6
answered question					22
skipped question					26

Fifty percent of the complexity coders responded to survey question 27, *What difficulties did you encounter in coding items for cognitive complexity?* The most salient reported difficulties include:

- The PARCC cognitive complexity framework is new and requires learning, processing, and applying lots of complex information.
- Coders struggled with understanding distinctions among the complexity codes, deciding how judgments about codes may differ across grade levels, understanding Processing Demands, and making decisions about "composite items."

The mathematics focus group offered no insights beyond their open ended responses to this question. The ELA focus group engaged in an extensive discussion and suggested the following:

- Working on short timelines and making judgments: Analysis of item response demands and making holistic complexity judgments takes time; determining the complexity level for each complexity source and integrating the weights adds to the required time. Some coders may have chosen to speed through making judgments by finding shortcuts that could lead to errors in judgment.
- Number of text structures and Processing Demands: Differences in the number of text structures, number of prepositional phrases, and other processing demands made it difficult to make Processing Demands complexity judgments.
- Different complexity considerations for different examinee subgroups: One coder expressed concerns about bias and English learners. This coder observed that you might judge an item as low or medium in cognitive complexity but other, similar items might have been judged as higher complexity because of challenges to English learners. This coder questioned how to address considerations for English learners into cognitive complexity judgments (and, by logical extension, considerations for students with disabilities and other struggle learners).

Recommendations on the PARCC Cognitive Complexity Framework and Its Use in the Future

In the final section of the survey, we asked complexity coders to recommend for PARCC consideration how the cognitive complexity frameworks might be improved and ways they can and should be used.

Sixteen of 48 survey respondents (33%) responded to survey question 28, *What was the most important thing you learned from implementing the cognitive complexity framework that would be important for PARCC to know about and consider in future item development?* Some of the most salient responses include:

- The PARCC cognitive complexity framework is "more comprehensive" (or perhaps more detailed or nuanced) than Bloom's Taxonomy and the Depth of Knowledge framework.
- It helps "set the tone for rigor" of items.
- Some items may be complex because of the standards they align to, even when text is Readily Accessible.
- It may difficult to produce medium and high complexity items associated with low complexity passages, primarily due to the weight of Text Complexity in the overall cognitive complexity measure.

Twenty of 48 survey respondents (42%) provided recommendations in response to survey question 29, *What changes would you recommend to (a) the framework, (b) the coding process (including training), or (c) how PARCC uses the framework?* Some of the most salient responses include:

- Perhaps consider differentiating complexity codes and definitions for the elementary, middle, and high school levels.
- Training should include more example items to illustrate high, medium, and low complexity for each complexity source; training should include examples for each schooling level and perhaps provide should be separate for coders of items from the grades 3-5, 6-8, and high school tests.
- PARCC should reevaluate the value of the Processing Demands complexity source. In contrast, in the regression tree analyses above, the Processing Demands empirical weights were very low for mathematics but high in ELA/literacy (0.44 on average).
- Review committees, including the OWG, should be trained in cognitive complexity.

The mathematics focus group offered no insights beyond their open ended responses to this question. The ELA focus group engaged in an extensive discussion and suggested the following:

- Familiarity with Bloom's taxonomy and the Depth of Knowledge framework may help in understanding and applying PARCC cognitive complexity even though PARCC items are quite different from items in other assessment programs.
- Differentiating codes across grade levels may not be necessary. However if there are plans to revise the cognitive complexity framework, PARCC could place more emphasis on vocabulary and language (i.e., linguistic demands) for items in lower grade levels, separate from middle and high school grade level items.
- There may not be time during committee reviews to consider cognitive complexity judgments across items (e.g., whether the judgments are consistent across similar items). So emphasize in training for coders and review committees the role that cognitive complexity plays in test forms construction.

Eighteen of 48 survey respondents (38%) provided recommendations to survey question 30, *For what purposes can item cognitive complexity measures be used in future item and task development?* Some of the most salient responses include:

- "To standardize test administration," which may be a reference to multi-stage, adaptive testing or to the PARCC test forms construction specifications to distribute item difficulty and cognitive complexity uniformly across the test reporting and proficiency scale.
- Require item writers to produce percentages of items at each complexity level.
- Avoid "artificially adjusting" complexity, a reference to avoiding making items more cognitively complex in construct irrelevant ways (e.g., making Processing Demands needlessly complex).
- Reading passages should be selected to support a range of cognitive complexity within item sets and across item sets in test forms.

Nineteen of 48 survey respondents (40%) provided recommendations in response to survey question 31, *Do you think that the cognitive complexity codes add value to the item metadata? In what ways?* Some of the most salient responses include:

- The framework may be more helpful at the lower grades, not at the higher grades. The ELA/literacy and mathematics focus groups offered no comments on this survey response.
- (Related to the "artificially adjusting" complexity response to survey question 30) The framework may help item developers to edit items so that they do not make items (needlessly) more difficult without item quality and adjusting complexity.
- The framework may be most valuable for selecting items for inclusion on "accommodated [test] forms."
- Only the overall complexity measure is needed.
- The Common Core standard to which items are aligned may be a better indicator of cognitive complexity than the complexity measure. Some evidence supports this view. For example, regression tree Importance statistics for CCSS identifiers and evidence statements in Tables 3.3 and 3.9 are relatively high but do not dominate other predictors of item and task difficulty.

Twenty of 48 survey respondents (42%) provided recommendations to survey question 32, *Would you recommend that PARCC items be coded at the overall cognitive complexity level only or at the individual complexity sources level as well? Please explain your answer.* Recommendations were divided:

- Use only the overall measure because it is simple.
- Use the individual sources because they "offer more insight" into item response demands.
- Use both because there is value in both the individual source codes and the overall measures.

One coder in the ELA focus group commented that cognitive complexity measures may have value as item metadata only for lower grade tests.

Sixteen survey respondents (33%) provided recommendations to survey question 33, *Would you recommend that PARCC use a different cognitive complexity coding framework for PARCC items and tasks? Which one and why?* Responses to this question were split:

- Three respondents indicated they prefer the PARCC cognitive complexity framework. Several others responded that they would not recommend another framework.
- One respondent recommended either simplifying the PARCC cognitive complexity framework or using the Depth of Knowledge framework. Another recommended using the Depth of Knowledge framework because it is understood by more the people who code and review items.

One coder in the mathematics focus group commented that it is easy to reduce Bloom's Taxonomy and Depth of Knowledge descriptions to verbs that indicate skills, and that PARCC cognitive complexity provides a more comprehensive way to evaluate items.

In the final two survey questions, we asked complexity coders to rate the validity of the PARCC cognitive complexity framework in comparison to Bloom's Taxonomy and the Depth of Knowledge framework. We expected that individual responses to these overall evaluations would be framed by responses to the previous open ended question (survey question 33). The majority of responses suggest the view that the PARCC complexity framework is similar in validity or more valid than Bloom's taxonomy (i.e., 17 of 19 countable responses to survey question 34) and the Depth of Knowledge framework (i.e., 19 of 20 countable responses to survey question 35). We note that nineteen respondents skipped these questions and that 9-10 indicated that they did not know enough about the other frameworks to make a fair judgment.

Survey Question 34

How would you rate the validity of the PARCC cognitive complexity framework in capturing the complexity of PARCC items compared to Bloom's taxonomy?						
Answer Options	Less Valid than Bloom's Taxonomy	Similar in Validity to Bloom's Taxonomy	More Valid than Bloom's Taxonomy	Don't Know Enough about Bloom's Taxonomy to Make a Fair Judgment	Rating Average	Response Count
	2	10	7	10	2.86	29
				answered question		29
				skipped question		19

Survey Question 35

How would you rate the validity of the PARCC cognitive complexity framework in capturing the complexity of PARCC items compared to the Depth of Knowledge framework?						
Answer Options	Less Valid than the Depth of Knowledge Framework	Similar in Validity to the Depth of Knowledge Framework	More Valid than the Depth of Knowledge Framework	Don't Know Enough about the Depth of Knowledge Framework to Make a Fair Judgment	Rating Average	Response Count
	1	11	8	9	2.86	29
				answered question		29
				skipped question		19

6. RECOMMENDATIONS AND DISCUSSION

We offer the following recommendations for PARCC consideration regarding the cognitive complexity framework. Our recommendations follow from the results of analyses 1, 2, and 3 and are based on experience in conducting item difficulty modeling research on other assessment programs. Additional recommendations for consideration appear through this report (e.g., in discussions in section 5, on survey and focus group results on coder training). In this section we highlight broader, and the most important recommendations.

Cognitive Complexity Sources, Weights, and Measures

PARCC should review and refine all complexity source definitions, refine distinctions among high, medium, and low complexity, and determine whether holistic complexity measures adequately support all intended uses of the cognitive complexity framework.

PARCC may want to review cognitive complexity sources that play limited roles in predicting item difficulty and discrimination in the regression tree results. Options could include (a) removing complexity sources from the framework because their regression tree Importance statistics are low; (b) retaining the codes as currently defined because they may play an important role in item and task specification, development, or review; or (c) revising the definitions and devising training enhancements for those codes.

> PARCC may want to delay any significant changes to the codes until a replication study is completed using operational data, after assessment design changes have been implemented and the PARCC program has matured.

PARCC may want to revise the judgmental weights assigned to each complexity source to align them with the empirical weights from these analyses.

> PARCC may want to delay any changes to the weights until a replication study is completed using operational data, after assessment design changes have been implemented and the PARCC program has matured.

Coding Items and Tasks for Coding Complexity

PARCC may want to consider using the holistic judgmental approach for operational coding of items, because it may be most efficient for coders, and coding individual cognitive complexity sources for monitoring and research purposes,[7] in follow-up studies. In order to make this decision, PARCC should (a) consider whether overall complexity measures provide adequate information to support all intended uses of the cognitive complexity framework, and (b) conduct studies to examine hypotheses about time requirements, cost, and information usefulness.

> Holistic coding appears to be more efficient than individual source coding, an important consideration for meeting challenging project time constraints, but that has not been confirmed.
>
> If holistic coding is more time efficient it may be the most cost effective approach, but that has not been confirmed.
>
> Holistic coding may provide information adequate to support all intended uses of the cognitive complexity framework, but that has not been confirmed.

PARCC may want to pursue procedures for automatic coding of some cognitive complexity sources. For example, components of the Text Complexity measure—that is, the TextEvaluator, Reading Maturity Metric, and Lexile measures—already are automated. And linguistic response demands coding—an element in the Processing Demands cognitive complexity source—can be coded automatically (paper forthcoming).

Other Recommendations

PARCC may want to offer training to other groups involved in the test development, review, and approval process. At least one complexity coder recommended in a response to an open ended survey question that the Operational Working Groups (OWG) should receive training in cognitive complexity. Members of the mathematics OWG proposed that during a July 23, 2015 briefing on this report, with the rationale that they should consider cognitive complexity measures as part of the item and task review process.

A complexity coder recommended in the focus group discussion that item and task cognitive complexity may differ for some items for English language learners. In fact, a similar concern would apply for students with disabilities and other struggling learners. PARCC may want to conduct a special study or working group to address this concern.

PARCC may want to consider reviewing and then explicating and publicizing intended interpretations and uses of the cognitive complexity frameworks.

> This recommendation also facilitates the earlier recommendation review and refine complexity sources and distinctions among high, medium, and low complexity and guide consideration of holistic versus individual complexity source coding.

[7] In phase 1 of item and task development, separate groups of coders coded for one cognitive complexity source. Subsequently, psychometricians combined those separate codes into a weighted composite, overall complexity measure. In phase 2 item and task development, coders made a holistic judgment of overall complexity by considering their judgments of the complexity of each individual source and the judgmental weighting scheme.

Additional Research

PARCC should replicate this study after assessment design changes have been implemented operationally, and when the PARCC assessment program matures, item data are stable, and student performance is growing incrementally.

> We acknowledge that item difficulties from 2014 and 2015 are highly correlated and that differences between mean difficulties are quite small. We make this caveat to the recommendation in anticipation that item difficulty and student performance could undergo shifts in the coming years, as teachers and students become even more familiar with the demands of the Common Core State Standards and PARCC items and tasks.
>
> If PARCC would prefer to conduct additional cognitive complexity studies using 2015 operational data, perhaps special focus studies could be conducted. Special focuses could include analyzing new items and comparing to the 2014 results in this project, focusing one type of ELA or mathematics task or TEI item functionality, focusing on selected claims and standards, and so forth.

PARCC may want to consider additional use of the results from analyses 1 and 2 that were not part of the scope of this project. In other studies (e.g., Ferrara et al., 2011; Ferrara & Steedle, 2015), we have discussed how regression tree results can be used to manipulate items to achieve item difficulty, discrimination, and complexity targets and construct relevant ways and to train item writers to do so. We can discuss how PARCC proceed on this recommendation upon request.

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

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

Ferrara, S., Dogan, E., Glazer, N., Gorin, J., Haberstroh, J., Hain, B., Huff, K., Larkin, J., Nichols, P., Piper, C., & Sheehan, P. (2014). Development of cognitive complexity measures for PARCC. In A Rupp (Chair), *Cognition and Assessment Special Interest Group Business Meeting* in the annual meeting of the American Educational Research Association, Philadelphia, PA.

Ferrara, S., Svetina, D., Skucha, S., & Murphy, A. (2011). Test design with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice, 30*(4), 3-15.

Ferrara, S., & Steedle, J. (2015 April). *Predicting item parameters using regression trees: Analyzing existing data to understand and improve item writing*. Presentation at the annual meeting of the National Council on Measurement in Education, Chicago.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394–411.

Hothorn, T., Hornik, K., & Zeileis, A. (2012). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651-674.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: American Council on Education/Praeger.

Appendix A: Additional Empirical Weights

Table A.1
Empirical Weights for Mathematics Based on Analysis of IRT B Values

							R^2	
	N	MC	MP	SM	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	323	.42	.39	.04	.09	.07	.133	.115	.138
Grade 4	327	.33	.27	.16	.23	.01	.326	.307	.296
Grade 5	277	.02	.65	.07	.25	.01	.081	.036	.069
Grade 6	239	.34	.47	.04	.01	.14	.040	.002	.032
Grade 7	251	.63	.13	.08	.00	.16	.063	.047	.059
Grade 8	253	.28	.23	.16	.00	.33	.000	.000	.058
Algebra I	209	.24	.00	.70	.06	.00	.000	.000	.006
Algebra II	229	.81	.00	.19	.00	.00	.043	.000	.012
Geometry	221	.32	.65	.00	.00	.03	.071	.020	.053
Integrated Math 1	73	.00	.00	.00	.00	.00	.000	.000	.000
Integrated Math 2	52	.00	.02	.00	.00	.98	.000	.000	.000
Integrated Math 3	62	.00	.00	.00	.01	.99	.000	.000	.000

Table A.2
Empirical Weights for Mathematics Based on Analysis of Biserial or Polyserial Correlations (Using Core Operational Tasks Only to Compute Total Scores)

							R^2	
	N	MC	MP	SM	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	323	.00	.00	.86	.02	.12	.099	.047	.058
Grade 4	327	.00	.00	.05	.85	.09	.041	.001	.006
Grade 5	277	.13	.11	.03	.69	.04	.103	.060	.074
Grade 6	239	.16	.02	.32	.32	.18	.036	.000	.027
Grade 7	251	.00	.00	.16	.84	.00	.000	.000	.000
Grade 8	253	.03	.00	.19	.62	.15	.064	.044	.042
Algebra I	209	.27	.26	.00	.38	.08	.023	.000	.011
Algebra II	229	.00	.00	.62	.38	.00	.000	.000	.000
Geometry	221	.19	.00	.55	.26	.00	.038	.023	.047
Integrated Math 1	73	.07	.00	.00	.75	.18	.207	.099	.111
Integrated Math 2	52	.64	.08	.02	.26	.00	.156	.000	.035
Integrated Math 3	62	.00	.00	.00	1.00	.00	.000	.000	.000

Table A.3
Empirical Weights for Mathematics Based on Analysis of Biserial or Polyserial Correlations (Using Core Operational and Non-Core Tasks to Compute Total Scores)

							R^2	
	N	MC	MP	SM	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	324	.00	.00	.86	.04	.09	.133	.109	.122
Grade 4	328	.00	.19	.00	.65	.16	.025	-.014	.005
Grade 5	279	.19	.00	.10	.69	.02	.053	.035	.032
Grade 6	241	.27	.03	.05	.41	.24	.074	.001	.054
Grade 7	256	.06	.05	.00	.89	.00	.061	.042	.027
Grade 8	258	.00	.01	.18	.81	.01	.074	.060	.039
Algebra I	223	.03	.21	.16	.60	.00	.104	.013	.064
Algebra II	239	.02	.07	.06	.70	.14	.048	.032	.029
Geometry	226	.13	.05	.15	.68	.00	.086	.072	.059
Integrated Math 1	80	.10	.02	.00	.77	.11	.174	.106	.069
Integrated Math 2	68	.00	.00	.36	.49	.14	.000	-.008	-.059
Integrated Math 3	67	.00	.00	.02	.98	.00	.000	-.033	-.061

Table A.4
Empirical Weights for ELA Based on Analysis of IRT B Values

						R^2		
	N	TC	CTE	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	155	.00	.00	.22	.78	.053	.007	.005
Grade 4	222	.00	.09	.00	.91	.095	.028	.039
Grade 5	169	.13	.20	.49	.17	.083	.000	.068
Grade 6	197	.01	.12	.38	.49	.135	.096	.108
Grade 7	162	.04	.08	.07	.80	.157	.127	.146
Grade 8	167	.42	.03	.11	.43	.126	.000	.055
Grade 9	109	.01	.02	.19	.78	.170	.143	.152
Grade 10	70	.25	.01	.01	.74	.000	.000	.000
Grade 11	192	.21	.43	.36	.00	.000	.000	.015

Table A.5

Empirical Weights for ELA Based on Analysis of Biserial or Polyserial Correlations (Using Core Operational Tasks Only to Compute Total Scores)

						R^2	
	N	TC	CTE	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	155	.36	.00	.49	.15	.000	.000	.000
Grade 4	222	.00	.09	.91	.00	.188	.178	.137
Grade 5	169	.00	.00	1.00	.00	.181	.125	.093
Grade 6	197	.02	.04	.93	.01	.236	.226	.187
Grade 7	162	.03	.05	.07	.85	.250	.242	.148
Grade 8	167	.00	.04	.96	.00	.278	.250	.175
Grade 9	109	.00	.01	.96	.03	.280	.250	.176
Grade 10	70	.12	.11	.49	.28	.018	.000	.000
Grade 11	192	.00	.15	.85	.00	.185	.148	.130

Table A.6

Empirical Weights for ELA Based on Analysis of Biserial or Polyserial Correlations (Using Core Operational and Non-Core Tasks to Compute Total Scores)

						R^2	
	N	TC	CTE	RM	PD	Cond. Tree	Cond. Tree Cross-Val.	Cond. Random Forest
Grade 3	156	.43	.00	.17	.40	.000	-.019	-.005
Grade 4	222	.00	.03	.97	.00	.336	.290	.301
Grade 5	169	.00	.00	.92	.08	.329	.313	.297
Grade 6	197	.05	.01	.87	.07	.360	.313	.328
Grade 7	162	.00	.01	.08	.91	.325	.312	.189
Grade 8	167	.02	.04	.87	.07	.403	.391	.343
Grade 9	109	.02	.04	.83	.11	.428	.409	.368
Grade 10	71	.24	.01	.52	.23	.027	-.097	-.006
Grade 11	192	.00	.18	.82	.00	.335	.326	.250

Appendix B: Mathematics Cognitive Complexity Descriptive Statistics

Table B.1
Descriptive Statistics for Predicting Grade 3 Mathematics P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.78	0.62	107	182	35	-.359	***	.129
Math Practices (MP)	1.28	0.52	242	72	10	-.356	***	.127
Stimulus Materials (SM)	1.56	0.82	212	44	68	.010		.000
Response Mode (RM)	1.15	0.52	296	6	22	-.301	***	.090
Processing Demands (PD)	1.52	0.56	164	150	10	-.158	**	.025
Judgmental CC Measure (ordinal)	1.51	0.62	181	122	21	-.373	***	.139
Judgmental CC Measure (numeric)	1.47	0.38				-.429	***	.184
Empirical CC Measure (ordinal)	1.53	0.67	185	107	32	-.384	***	.148
Empirical CC Measure (numeric)	1.47	0.41				-.443	***	.196

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.2
Descriptive Statistics for Predicting Grade 4 Mathematics P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.60	0.60	149	160	19	-.422	***	.178
Math Practices (MP)	1.35	0.55	227	88	13	-.442	***	.195
Stimulus Materials (SM)	1.29	0.62	262	37	29	-.226	***	.051
Response Mode (RM)	1.19	0.56	292	10	26	-.424	***	.180
Processing Demands (PD)	1.53	0.54	161	161	6	-.262	***	.069
Judgmental CC Measure (ordinal)	1.41	0.58	210	102	16	-.514	***	.264
Judgmental CC Measure (numeric)	1.42	0.40				-.569	***	.323
Empirical CC Measure (ordinal)	1.39	0.57	215	98	15	-.533	***	.284
Empirical CC Measure (numeric)	1.38	0.39				-.587	***	.344

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.3
Descriptive Statistics for Predicting Grade 5 Mathematics P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.52	0.57	146	122	11	-.207	***	.043
Math Practices (MP)	1.32	0.53	199	71	9	-.298	***	.089
Stimulus Materials (SM)	1.27	0.58	224	35	20	.039		.002
Response Mode (RM)	1.19	0.57	249	7	23	-.291	***	.085
Processing Demands (PD)	1.57	0.55	129	142	8	-.214	***	.046
Judgmental CC Measure (ordinal)	1.35	0.56	191	77	11	-.293	***	.086
Judgmental CC Measure (numeric)	1.38	0.39				-.324	***	.105
Empirical CC Measure (ordinal)	1.38	0.60	189	73	17	-.351	***	.123
Empirical CC Measure (numeric)	1.27	0.44				-.356	***	.127

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.4
Descriptive Statistics for Predicting Grade 6 Mathematics P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.47	0.59	140	89	12	-.244	***	.060
Math Practices (MP)	1.18	0.45	203	32	6	-.246	***	.061
Stimulus Materials (SM)	1.24	0.55	198	29	14	-.077		.006
Response Mode (RM)	1.24	0.60	206	13	22	-.222	***	.049
Processing Demands (PD)	1.68	0.59	92	134	15	-.219	***	.048
Judgmental CC Measure (ordinal)	1.29	0.54	183	47	11	-.327	***	.107
Judgmental CC Measure (numeric)	1.33	0.38				-.310	***	.096
Empirical CC Measure (ordinal)	1.29	0.57	187	39	15	-.316	***	.100
Empirical CC Measure (numeric)	1.34	0.39				-.314	***	.099

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.5
Descriptive Statistics for Predicting Grade 7 Mathematics P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.37	0.49	162	93	1	-.313	***	.098
Math Practices (MP)	1.19	0.39	208	48	0	-.207	***	.043
Stimulus Materials (SM)	1.26	0.57	205	35	16	.061		.004
Response Mode (RM)	1.28	0.69	218	4	34	-.181	**	.033
Processing Demands (PD)	1.66	0.54	95	153	8	.081		.007
Judgmental CC Measure (ordinal)	1.24	0.44	197	57	2	-.241	***	.058
Judgmental CC Measure (numeric)	1.31	0.30				-.273	***	.075
Empirical CC Measure (ordinal)	1.37	0.49	162	93	1	-.313	***	.098
Empirical CC Measure (numeric)	1.34	0.37				-.318	***	.101

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.6
Descriptive Statistics for Predicting Grade 8 Mathematics P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.50	0.59	141	104	13	-.154	*	.024
Math Practices (MP)	1.21	0.45	210	43	5	-.191	**	.036
Stimulus Materials (SM)	1.58	0.74	147	72	39	.071		.005
Response Mode (RM)	1.28	0.65	215	14	29	-.181	**	.033
Processing Demands (PD)	1.61	0.60	117	125	16	.090		.008
Judgmental CC Measure (ordinal)	1.38	0.56	170	78	10	-.210	***	.044
Judgmental CC Measure (numeric)	1.38	0.38				-.166	**	.027
Empirical CC Measure (ordinal)	1.49	0.61	146	97	15	-.089		.008
Empirical CC Measure (numeric)	1.45	0.39				-.098		.010

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.7
Descriptive Statistics for Predicting Algebra I P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.35	0.51	149	70	4	-.133	*	.018
Math Practices (MP)	1.28	0.51	167	50	6	-.187	**	.035
Stimulus Materials (SM)	1.31	0.62	171	34	18	.181	**	.033
Response Mode (RM)	1.22	0.59	195	8	20	-.090		.008
Processing Demands (PD)	1.50	0.54	117	101	5	-.077		.006
Judgmental CC Measure (ordinal)	1.33	0.53	156	60	7	-.127		.016
Judgmental CC Measure (numeric)	1.32	0.33				-.170	*	.029
Empirical CC Measure (ordinal)	1.34	0.64	168	34	21	.149	*	.022
Empirical CC Measure (numeric)	1.31	0.45				.095		.009

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.8
Descriptive Statistics for Predicting Algebra II P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.31	0.49	168	68	3	-.236	***	.056
Math Practices (MP)	1.28	0.52	179	52	8	-.223	***	.050
Stimulus Materials (SM)	1.22	0.51	196	33	10	.132	*	.017
Response Mode (RM)	1.20	0.60	214	2	23	-.203	**	.041
Processing Demands (PD)	1.43	0.55	143	89	7	-.081		.007
Judgmental CC Measure (ordinal)	1.28	0.53	181	49	9	-.203	**	.041
Judgmental CC Measure (numeric)	1.29	0.36				-.255	***	.065
Empirical CC Measure (ordinal)	1.33	0.51	166	68	5	-.235	***	.055
Empirical CC Measure (numeric)	1.27	0.36				-.219	***	.048

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.9
Descriptive Statistics for Predicting Geometry P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.37	0.54	149	70	7	-.224	***	.050
Math Practices (MP)	1.38	0.58	152	63	11	-.291	***	.085
Stimulus Materials (SM)	1.56	0.73	132	62	32	-.113		.013
Response Mode (RM)	1.26	0.67	195	3	28	-.276	***	.076
Processing Demands (PD)	1.60	0.57	100	117	9	.031		.001
Judgmental CC Measure (ordinal)	1.41	0.60	147	66	13	-.270	***	.073
Judgmental CC Measure (numeric)	1.40	0.41				-.311	***	.096
Empirical CC Measure (ordinal)	1.45	0.67	146	58	22	-.346	***	.120
Empirical CC Measure (numeric)	1.35	0.47				-.331	***	.110

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.10
Descriptive Statistics for Predicting Integrated Mathematics 1 P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.40	0.54	50	28	2	.051		.003
Math Practices (MP)	1.26	0.47	60	19	1	-.088		.008
Stimulus Materials (SM)	1.34	0.64	60	13	7	-.015		.000
Response Mode (RM)	1.28	0.66	67	4	9	-.089		.008
Processing Demands (PD)	1.59	0.52	34	45	1	.108		.012
Judgmental CC Measure (ordinal)	1.34	0.57	57	19	4	-.001		.000
Judgmental CC Measure (numeric)	1.35	0.35				-.027		.001
Empirical CC Measure (ordinal)	1.59	0.52	34	45	1	.108		.012
Empirical CC Measure (numeric)	1.50	0.41				.074		.006

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.11

Descriptive Statistics for Predicting Integrated Mathematics 2 P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.25	0.44	51	17	0	.124		.015
Math Practices (MP)	1.28	0.48	50	17	1	-.174		.030
Stimulus Materials (SM)	1.35	0.75	55	2	11	.157		.025
Response Mode (RM)	1.13	0.49	63	1	4	-.061		.004
Processing Demands (PD)	1.46	0.53	38	29	1	.274	*	.075
Judgmental CC Measure (ordinal)	1.24	0.49	54	12	2	-.031		.001
Judgmental CC Measure (numeric)	1.28	0.31				.019		.000
Empirical CC Measure (ordinal)	1.46	0.53	38	29	1	.274	*	.075
Empirical CC Measure (numeric)	1.44	0.48				.277	*	.077

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.12

Descriptive Statistics for Predicting Integrated Mathematics 3 P-Values

			CC Measure Frequency				
	Mean	SD	Low	Mod	High	*r*	R^2
Math Content (MC)	1.34	0.48	44	23	0	-.060	.004
Math Practices (MP)	1.22	0.45	53	13	1	-.147	.022
Stimulus Materials (SM)	1.28	0.62	54	7	6	.068	.005
Response Mode (RM)	1.06	0.34	65	0	2	-.015	.000
Processing Demands (PD)	1.57	0.58	32	32	3	.083	.007
Judgmental CC Measure (ordinal)	1.25	0.47	51	15	1	.012	.000
Judgmental CC Measure (numeric)	1.28	0.28				-.097	.009
Empirical CC Measure (ordinal)	1.22	0.45	53	13	1	-.147	.022
Empirical CC Measure (numeric)	1.22	0.45				-.147	.022

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.13

Descriptive Statistics for Predicting Grade 3 Mathematics Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.78	0.62	107	182	35	.012		.000
Math Practices (MP)	1.28	0.52	242	72	10	.171	**	.029
Stimulus Materials (SM)	1.56	0.82	212	44	68	-.314	***	.098
Response Mode (RM)	1.15	0.52	296	6	22	.203	***	.041
Processing Demands (PD)	1.52	0.56	164	150	10	.236	***	.056
Judgmental CC Measure (ordinal)	1.51	0.62	181	122	21	.041		.002
Judgmental CC Measure (numeric)	1.47	0.38				.093		.009
Empirical CC Measure (ordinal)	1.52	0.70	194	91	39	-.199	***	.040
Empirical CC Measure (numeric)	1.51	0.51				-.177	**	.031

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.14

Descriptive Statistics for Predicting Grade 4 Mathematics Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.60	0.60	149	160	19	.086		.007
Math Practices (MP)	1.35	0.55	227	88	13	.094		.009
Stimulus Materials (SM)	1.29	0.62	262	37	29	.137	*	.019
Response Mode (RM)	1.19	0.56	292	10	26	.271	***	.074
Processing Demands (PD)	1.53	0.54	161	161	6	.197	***	.039
Judgmental CC Measure (ordinal)	1.41	0.58	210	102	16	.127	*	.016
Judgmental CC Measure (numeric)	1.42	0.40				.178	**	.032
Empirical CC Measure (ordinal)	1.19	0.56	292	10	26	.271	***	.074
Empirical CC Measure (numeric)	1.26	0.44				.305	***	.093

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.15

Descriptive Statistics for Predicting Grade 5 Mathematics Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.52	0.57	146	122	11	-.054		.003
Math Practices (MP)	1.32	0.53	199	71	9	.127	*	.016
Stimulus Materials (SM)	1.27	0.58	224	35	20	-.058		.003
Response Mode (RM)	1.19	0.57	249	7	23	.328	***	.108
Processing Demands (PD)	1.57	0.55	129	142	8	.189	**	.036
Judgmental CC Measure (ordinal)	1.35	0.56	191	77	11	.146	*	.021
Judgmental CC Measure (numeric)	1.38	0.39				.112		.013
Empirical CC Measure (ordinal)	1.20	0.57	247	9	23	.304	***	.092
Empirical CC Measure (numeric)	1.29	0.45				.283	***	.080

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.16

Descriptive Statistics for Predicting Grade 6 Mathematics Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.47	0.59	140	89	12	.024		.001
Math Practices (MP)	1.18	0.45	203	32	6	.144	*	.021
Stimulus Materials (SM)	1.24	0.55	198	29	14	.120		.014
Response Mode (RM)	1.24	0.60	206	13	22	.364	***	.133
Processing Demands (PD)	1.68	0.59	92	134	15	.250	***	.062
Judgmental CC Measure (ordinal)	1.29	0.54	183	47	11	.158	*	.025
Judgmental CC Measure (numeric)	1.33	0.38				.192	**	.037
Empirical CC Measure (ordinal)	1.25	0.61	202	17	22	.338	***	.114
Empirical CC Measure (numeric)	1.34	0.51				.365	***	.133

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.17

Descriptive Statistics for Predicting Grade 7 Mathematics Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.37	0.49	162	93	1	.112		.012
Math Practices (MP)	1.19	0.39	208	48	0	.087		.008
Stimulus Materials (SM)	1.26	0.57	205	35	16	.023		.001
Response Mode (RM)	1.28	0.69	218	4	34	.412	***	.170
Processing Demands (PD)	1.66	0.54	95	153	8	.141	*	.020
Judgmental CC Measure (ordinal)	1.24	0.44	197	57	2	.220	***	.048
Judgmental CC Measure (numeric)	1.31	0.30				.221	***	.049
Empirical CC Measure (ordinal)	1.28	0.69	218	4	34	.412	***	.170
Empirical CC Measure (numeric)	1.29	0.66				.412	***	.170

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.18

Descriptive Statistics for Predicting Grade 8 Mathematics Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.50	0.59	141	104	13	.090		.008
Math Practices (MP)	1.21	0.45	210	43	5	.125	*	.016
Stimulus Materials (SM)	1.58	0.74	147	72	39	.140	*	.020
Response Mode (RM)	1.28	0.65	215	14	29	.369	***	.136
Processing Demands (PD)	1.61	0.60	117	125	16	.195	**	.038
Judgmental CC Measure (ordinal)	1.38	0.56	170	78	10	.217	***	.047
Judgmental CC Measure (numeric)	1.38	0.38				.223	***	.050
Empirical CC Measure (ordinal)	1.28	0.65	215	14	29	.369	***	.136
Empirical CC Measure (numeric)	1.32	0.58				.376	***	.142

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.19

Descriptive Statistics for Predicting Algebra I Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.35	0.51	149	70	4	.111		.012
Math Practices (MP)	1.28	0.51	167	50	6	-.077		.006
Stimulus Materials (SM)	1.31	0.62	171	34	18	.154	*	.024
Response Mode (RM)	1.22	0.59	195	8	20	.291	***	.085
Processing Demands (PD)	1.50	0.54	117	101	5	.054		.003
Judgmental CC Measure (ordinal)	1.33	0.53	156	60	7	.111		.012
Judgmental CC Measure (numeric)	1.32	0.33				.093		.009
Empirical CC Measure (ordinal)	1.23	0.61	193	9	21	.315	***	.099
Empirical CC Measure (numeric)	1.25	0.43				.293	***	.086

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.20

Descriptive Statistics for Predicting Algebra II Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.31	0.49	168	68	3	.084		.007
Math Practices (MP)	1.28	0.52	179	52	8	.028		.001
Stimulus Materials (SM)	1.22	0.51	196	33	10	-.056		.003
Response Mode (RM)	1.20	0.60	214	2	23	.285	***	.081
Processing Demands (PD)	1.43	0.55	143	89	7	.047		.002
Judgmental CC Measure (ordinal)	1.28	0.53	181	49	9	.096		.009
Judgmental CC Measure (numeric)	1.29	0.36				.098		.010
Empirical CC Measure (ordinal)	1.20	0.60	214	2	23	.285	***	.081
Empirical CC Measure (numeric)	1.22	0.54				.277	***	.077

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.21

Descriptive Statistics for Predicting Geometry Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.37	0.54	149	70	7	.023		.001
Math Practices (MP)	1.38	0.58	152	63	11	.106		.011
Stimulus Materials (SM)	1.56	0.73	132	62	32	.113		.013
Response Mode (RM)	1.26	0.67	195	3	28	.335	***	.112
Processing Demands (PD)	1.60	0.57	100	117	9	.046		.002
Judgmental CC Measure (ordinal)	1.41	0.60	147	66	13	.162	*	.026
Judgmental CC Measure (numeric)	1.40	0.41				.148	*	.022
Empirical CC Measure (ordinal)	1.26	0.67	195	3	28	.335	***	.112
Empirical CC Measure (numeric)	1.30	0.58				.323	***	.105

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.22

Descriptive Statistics for Predicting Integrated Mathematics 1 Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.40	0.54	50	28	2	.285	*	.081
Math Practices (MP)	1.26	0.47	60	19	1	.153		.024
Stimulus Materials (SM)	1.34	0.64	60	13	7	.097		.010
Response Mode (RM)	1.28	0.66	67	4	9	.505	***	.255
Processing Demands (PD)	1.59	0.52	34	45	1	-.003		.000
Judgmental CC Measure (ordinal)	1.34	0.57	57	19	4	.308	**	.095
Judgmental CC Measure (numeric)	1.35	0.35				.324	**	.105
Empirical CC Measure (ordinal)	1.28	0.66	67	4	9	.505	***	.255
Empirical CC Measure (numeric)	1.29	0.62				.513	***	.263

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.23

Descriptive Statistics for Predicting Integrated Mathematics 4 Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.25	0.44	51	17	0	.178		.032
Math Practices (MP)	1.28	0.48	50	17	1	-.001		.000
Stimulus Materials (SM)	1.35	0.75	55	2	11	.279	*	.078
Response Mode (RM)	1.13	0.49	63	1	4	.428	***	.183
Processing Demands (PD)	1.46	0.53	38	29	1	.055		.003
Judgmental CC Measure (ordinal)	1.24	0.49	54	12	2	.283	*	.080
Judgmental CC Measure (numeric)	1.28	0.31				.220		.048
Empirical CC Measure (ordinal)	1.26	0.51	52	14	2	.420	***	.176
Empirical CC Measure (numeric)	1.25	0.48				.436	***	.190

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B.24

Descriptive Statistics for Predicting Integrated Mathematics 3 Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Math Content (MC)	1.34	0.48	44	23	0	.000		.000
Math Practices (MP)	1.22	0.45	53	13	1	-.051		.003
Stimulus Materials (SM)	1.28	0.62	54	7	6	.135		.018
Response Mode (RM)	1.06	0.34	65	0	2	.176		.031
Processing Demands (PD)	1.57	0.58	32	32	3	.038		.002
Judgmental CC Measure (ordinal)	1.25	0.47	51	15	1	.067		.005
Judgmental CC Measure (numeric)	1.28	0.28				.027		.001
Empirical CC Measure (ordinal)								
Empirical CC Measure (numeric)								

* $p < .05$, ** $p < .01$, *** $p < .001$

Appendix C: ELA Cognitive Complexity Descriptive Statistics

Table C.1

Descriptive Statistics for Predicting Grade 3 ELA P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	1.61	0.62	72	73	11	-.152		.023
Command of Textual Evidence (CTE)	1.78	0.49	40	111	5	-.173	*	.030
Response Mode (RM)	1.29	0.48	113	41	2	-.082		.007
Processing Demands (PD)	1.72	0.49	46	107	3	-.242	**	.058
Judgmental CC Measure (ordinal)	1.50	0.62	88	58	10	-.210	**	.044
Judgmental CC Measure (numeric)	1.59	0.37				-.226	**	.051
Empirical CC Measure (ordinal)	1.72	0.49	46	107	3	-.242	**	.058
Empirical CC Measure (numeric)	1.72	0.49				-.242	**	.058

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.2

Descriptive Statistics for Predicting Grade 4 ELA P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	1.64	0.48	79	143	0	.006		.000
Command of Textual Evidence (CTE)	1.85	0.55	53	150	19	-.220	***	.048
Response Mode (RM)	1.48	0.63	132	74	16	-.132	*	.018
Processing Demands (PD)	1.64	0.56	89	124	9	-.226	***	.051
Judgmental CC Measure (ordinal)	1.56	0.59	109	102	11	-.143	*	.021
Judgmental CC Measure (numeric)	1.65	0.33				-.159	*	.025
Empirical CC Measure (ordinal)	1.62	0.60	99	109	14	-.251	***	.063
Empirical CC Measure (numeric)	1.72	0.43				-.290	***	.084

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.3

Descriptive Statistics for Predicting Grade 5 ELA P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.05	0.61	27	106	36	-.145		.021
Command of Textual Evidence (CTE)	1.91	0.54	33	118	18	-.294	***	.086
Response Mode (RM)	1.60	0.63	80	76	13	-.333	***	.111
Processing Demands (PD)	1.85	0.63	48	99	22	-.073		.005
Judgmental CC Measure (ordinal)	1.92	0.62	39	104	26	-.284	***	.081
Judgmental CC Measure (numeric)	1.91	0.38				-.320	***	.102
Empirical CC Measure (ordinal)	1.63	0.67	80	71	18	-.324	***	.105
Empirical CC Measure (numeric)	1.71	0.47				-.365	***	.133

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.4

Descriptive Statistics for Predicting Grade 6 ELA P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.14	0.70	37	96	64	-.151	*	.023
Command of Textual Evidence (CTE)	1.94	0.43	24	160	13	-.327	***	.107
Response Mode (RM)	1.53	0.63	106	77	14	-.366	***	.134
Processing Demands (PD)	1.70	0.59	72	112	13	-.196	**	.039
Judgmental CC Measure (ordinal)	2.05	0.71	44	99	54	-.221	**	.049
Judgmental CC Measure (numeric)	1.93	0.42				-.333	***	.111
Empirical CC Measure (ordinal)	1.53	0.64	108	74	15	-.361	***	.130
Empirical CC Measure (numeric)	1.66	0.43				-.433	***	.187

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.5

Descriptive Statistics for Predicting Grade 7 ELA P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.02	0.72	40	78	44	-.151		.023
Command of Textual Evidence (CTE)	2.09	0.74	38	72	52	-.139		.019
Response Mode (RM)	1.96	0.45	20	129	13	-.356	***	.127
Processing Demands (PD)	1.59	0.63	79	71	12	-.490	***	.241
Judgmental CC Measure (ordinal)	2.06	0.75	40	72	50	-.234	**	.055
Judgmental CC Measure (numeric)	1.98	0.52				-.263	***	.069
Empirical CC Measure (ordinal)	1.59	0.63	79	71	12	-.490	***	.241
Empirical CC Measure (numeric)	1.67	0.53				-.513	***	.263

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.6

Descriptive Statistics for Predicting Grade 8 ELA P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	1.88	0.80	64	59	44	-.302	***	.091
Command of Textual Evidence (CTE)	1.98	0.44	18	135	14	-.205	**	.042
Response Mode (RM)	1.58	0.63	83	71	13	-.255	***	.065
Processing Demands (PD)	1.69	0.56	59	100	8	-.172	*	.030
Judgmental CC Measure (ordinal)	1.84	0.79	68	58	41	-.349	***	.122
Judgmental CC Measure (numeric)	1.82	0.45				-.402	***	.162
Empirical CC Measure (ordinal)	1.84	0.81	69	55	43	-.352	***	.124
Empirical CC Measure (numeric)	1.81	0.48				-.393	***	.154

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.7

Descriptive Statistics for Predicting Grade 9 ELA P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.06	0.64	19	65	25	-.232	*	.054
Command of Textual Evidence (CTE)	1.95	0.48	15	84	10	-.301	**	.091
Response Mode (RM)	1.77	0.63	37	60	12	-.327	***	.107
Processing Demands (PD)	1.76	0.56	33	69	7	-.376	***	.141
Judgmental CC Measure (ordinal)	2.04	0.68	23	59	27	-.302	**	.091
Judgmental CC Measure (numeric)	1.95	0.42				-.392	***	.154
Empirical CC Measure (ordinal)	1.86	0.58	27	70	12	-.504	***	.254
Empirical CC Measure (numeric)	1.79	0.41				-.494	***	.244

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.8

Descriptive Statistics for Predicting Grade 10 ELA P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.07	0.78	19	28	24	.148		.022
Command of Textual Evidence (CTE)	2.01	0.27	2	66	3	-.345	**	.119
Response Mode (RM)	1.63	0.57	29	39	3	-.126		.016
Processing Demands (PD)	1.90	0.72	22	34	15	-.098		.010
Judgmental CC Measure (ordinal)	2.03	0.76	19	31	21	-.016		.000
Judgmental CC Measure (numeric)	1.95	0.44				.042		.002
Empirical CC Measure (ordinal)	2.07	0.78	19	28	24	.052		.003
Empirical CC Measure (numeric)	1.92	0.54				.080		.006

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.9
Descriptive Statistics for Predicting Grade 11 ELA P-Values

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.31	0.59	13	106	73	-.025		.001
Command of Textual Evidence (CTE)	1.95	0.41	21	160	11	-.249	***	.062
Response Mode (RM)	1.72	0.59	68	110	14	-.216	**	.046
Processing Demands (PD)	1.91	0.59	43	123	26	-.045		.002
Judgmental CC Measure (ordinal)	2.22	0.64	22	105	65	-.133		.018
Judgmental CC Measure (numeric)	2.08	0.37				-.150	*	.022
Empirical CC Measure (ordinal)	1.99	0.53	27	139	26	-.302	***	.091
Empirical CC Measure (numeric)	1.96	0.33				-.261	***	.068

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.10
Descriptive Statistics for Predicting Grade 3 ELA Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	1.61	0.62	72	73	11	-.168	*	.028
Command of Textual Evidence (CTE)	1.78	0.49	40	111	5	.009		.000
Response Mode (RM)	1.29	0.48	113	41	2	.067		.005
Processing Demands (PD)	1.72	0.49	46	107	3	-.149		.022
Judgmental CC Measure (ordinal)	1.50	0.62	88	58	10	-.158	*	.025
Judgmental CC Measure (numeric)	1.59	0.37				-.140		.020
Empirical CC Measure (ordinal)	1.48	0.64	93	51	12	-.204	*	.042
Empirical CC Measure (numeric)	1.67	0.41				-.215	**	.046

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.11

Descriptive Statistics for Predicting Grade 4 ELA Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	1.64	0.48	79	143	0	.045		.002
Command of Textual Evidence (CTE)	1.85	0.55	53	150	19	.263	***	.069
Response Mode (RM)	1.48	0.63	132	74	16	.530	***	.280
Processing Demands (PD)	1.64	0.56	89	124	9	-.161	*	.026
Judgmental CC Measure (ordinal)	1.56	0.59	109	102	11	.260	***	.067
Judgmental CC Measure (numeric)	1.65	0.33				.298	***	.089
Empirical CC Measure (ordinal)	1.48	0.63	132	74	16	.530	***	.280
Empirical CC Measure (numeric)	1.50	0.61				.529	***	.280

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.12

Descriptive Statistics for Predicting Grade 5 ELA Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.05	0.61	27	106	36	-.053		.003
Command of Textual Evidence (CTE)	1.91	0.54	33	118	18	.151	*	.023
Response Mode (RM)	1.60	0.63	80	76	13	.479	***	.229
Processing Demands (PD)	1.85	0.63	48	99	22	-.296	***	.088
Judgmental CC Measure (ordinal)	1.92	0.62	39	104	26	.084		.007
Judgmental CC Measure (numeric)	1.91	0.38				.110		.012
Empirical CC Measure (ordinal)	1.60	0.63	80	76	13	.479	***	.229
Empirical CC Measure (numeric)	1.63	0.57				.433	***	.187

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.13

Descriptive Statistics for Predicting Grade 6 ELA Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.14	0.70	37	96	64	-.025		.001
Command of Textual Evidence (CTE)	1.94	0.43	24	160	13	.259	***	.067
Response Mode (RM)	1.53	0.63	106	77	14	.500	***	.250
Processing Demands (PD)	1.70	0.59	72	112	13	-.214	**	.046
Judgmental CC Measure (ordinal)	2.05	0.71	44	99	54	.136		.019
Judgmental CC Measure (numeric)	1.93	0.42				.153	*	.023
Empirical CC Measure (ordinal)	1.53	0.63	106	77	14	.500	***	.250
Empirical CC Measure (numeric)	1.57	0.58				.493	***	.243

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.14

Descriptive Statistics for Predicting Grade 7 ELA Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.02	0.72	40	78	44	.039		.002
Command of Textual Evidence (CTE)	2.09	0.74	38	72	52	.039		.002
Response Mode (RM)	1.96	0.45	20	129	13	.189	*	.036
Processing Demands (PD)	1.59	0.63	79	71	12	.424	***	.180
Judgmental CC Measure (ordinal)	2.06	0.75	40	72	50	.150		.023
Judgmental CC Measure (numeric)	1.98	0.52				.121		.015
Empirical CC Measure (ordinal)	1.59	0.63	79	71	12	.424	***	.180
Empirical CC Measure (numeric)	1.62	0.59				.422	***	.178

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.15

Descriptive Statistics for Predicting Grade 8 ELA Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	1.88	0.80	64	59	44	-.067		.005
Command of Textual Evidence (CTE)	1.98	0.44	18	135	14	.269	***	.073
Response Mode (RM)	1.58	0.63	83	71	13	.485	***	.236
Processing Demands (PD)	1.69	0.56	59	100	8	-.279	***	.078
Judgmental CC Measure (ordinal)	1.84	0.79	68	58	41	.058		.003
Judgmental CC Measure (numeric)	1.82	0.45				.096		.009
Empirical CC Measure (ordinal)	1.58	0.63	83	71	13	.485	***	.236
Empirical CC Measure (numeric)	1.61	0.56				.468	***	.219

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.16

Descriptive Statistics for Predicting Grade 9 ELA Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.06	0.64	19	65	25	.088		.008
Command of Textual Evidence (CTE)	1.95	0.48	15	84	10	.349	***	.122
Response Mode (RM)	1.77	0.63	37	60	12	.554	***	.307
Processing Demands (PD)	1.76	0.56	33	69	7	-.247	**	.061
Judgmental CC Measure (ordinal)	2.04	0.68	23	59	27	.250	**	.062
Judgmental CC Measure (numeric)	1.95	0.42				.280	**	.078
Empirical CC Measure (ordinal)	1.77	0.63	37	60	12	.554	***	.307
Empirical CC Measure (numeric)	1.79	0.57				.546	***	.298

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.17

Descriptive Statistics for Predicting Grade 10 ELA Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.07	0.78	19	28	24	.158		.025
Command of Textual Evidence (CTE)	2.01	0.27	2	66	3	.155		.024
Response Mode (RM)	1.63	0.57	29	39	3	.308	**	.095
Processing Demands (PD)	1.90	0.72	22	34	15	-.073		.005
Judgmental CC Measure (ordinal)	2.03	0.76	19	31	21	.195		.038
Judgmental CC Measure (numeric)	1.95	0.44				.228		.052
Empirical CC Measure (ordinal)	1.73	0.70	29	32	10	.231		.053
Empirical CC Measure (numeric)	1.82	0.40				.275	*	.076

* $p < .05$, ** $p < .01$, *** $p < .001$

Table C.18

Descriptive Statistics for Predicting Grade 11 ELA Point-Biserial Correlations

			CC Measure Frequency					
	Mean	SD	Low	Mod	High	*r*		R^2
Text Complexity (TC)	2.31	0.59	13	106	73	.069		.005
Command of Textual Evidence (CTE)	1.95	0.41	21	160	11	.291	***	.084
Response Mode (RM)	1.72	0.59	68	110	14	.375	***	.141
Processing Demands (PD)	1.91	0.59	43	123	26	-.168	*	.028
Judgmental CC Measure (ordinal)	2.22	0.64	22	105	65	.105		.011
Judgmental CC Measure (numeric)	2.08	0.37				.211	**	.044
Empirical CC Measure (ordinal)	1.72	0.59	68	110	14	.375	***	.141
Empirical CC Measure (numeric)	1.77	0.51				.393	***	.154

* $p < .05$, ** $p < .01$, *** $p < .001$

Appendix D: Mathematics Conditional Trees using Cognitive Complexity Source Codes as Predictors

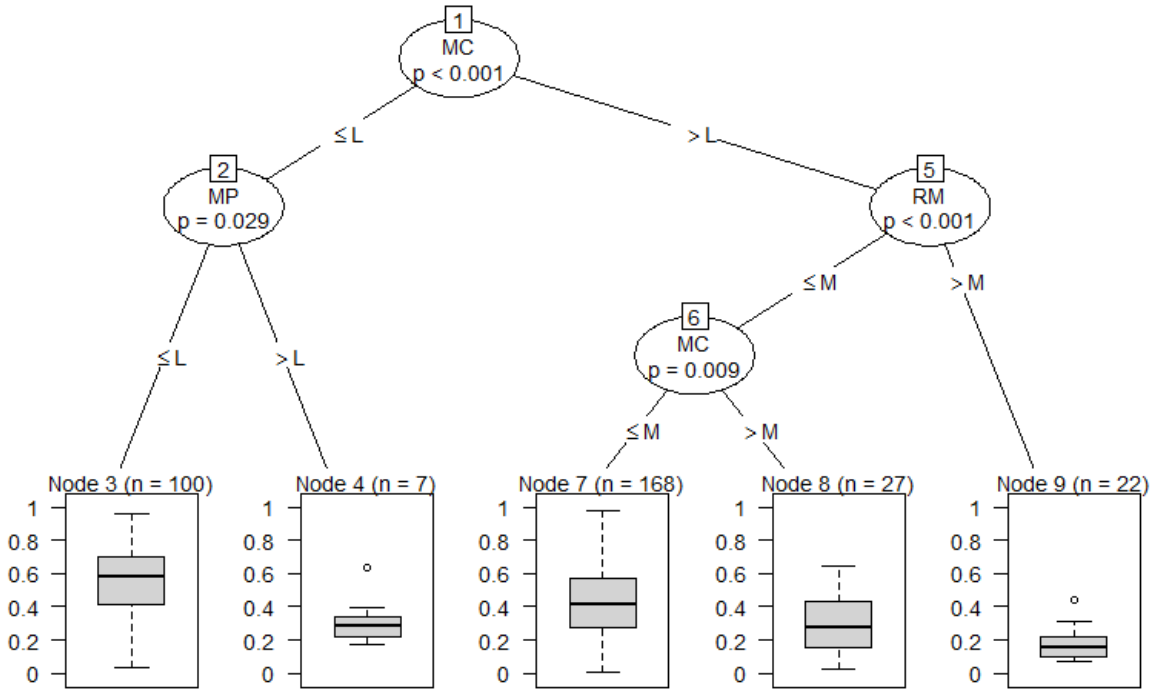


Figure D.1. Conditional Tree for Predicting Grade 3 Mathematics P-Values from Cognitive Complexity Source Codes.

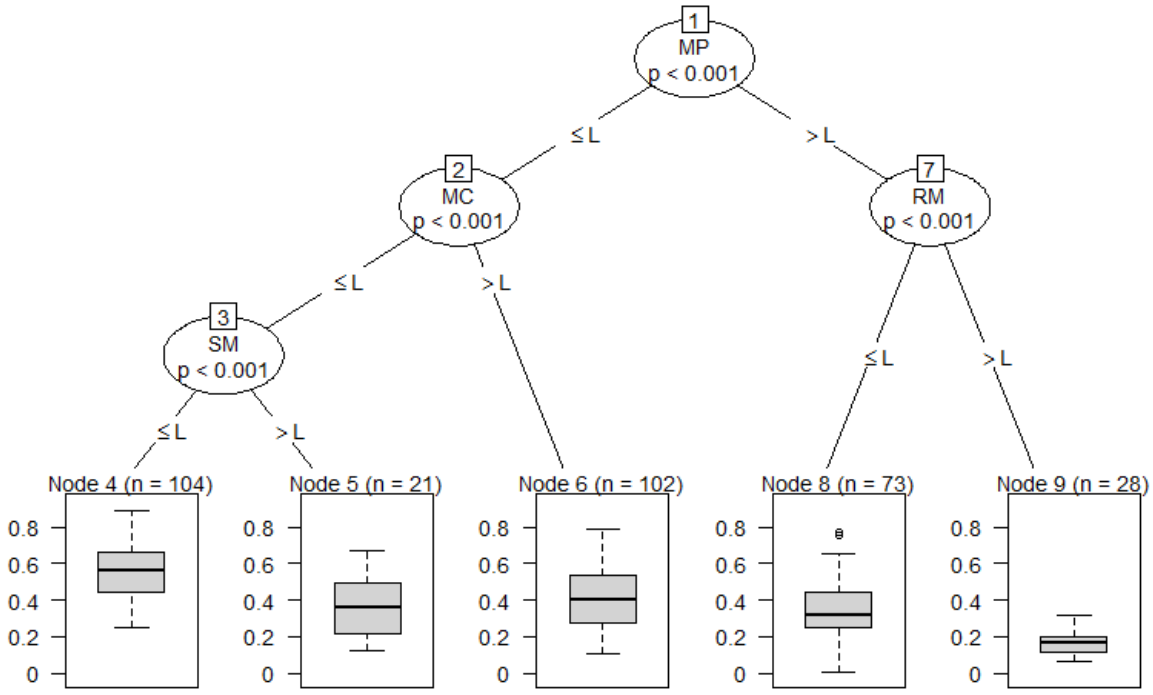


Figure D.2. Conditional Tree for Predicting Grade 4 Mathematics P-Values from Cognitive Complexity Source Codes.

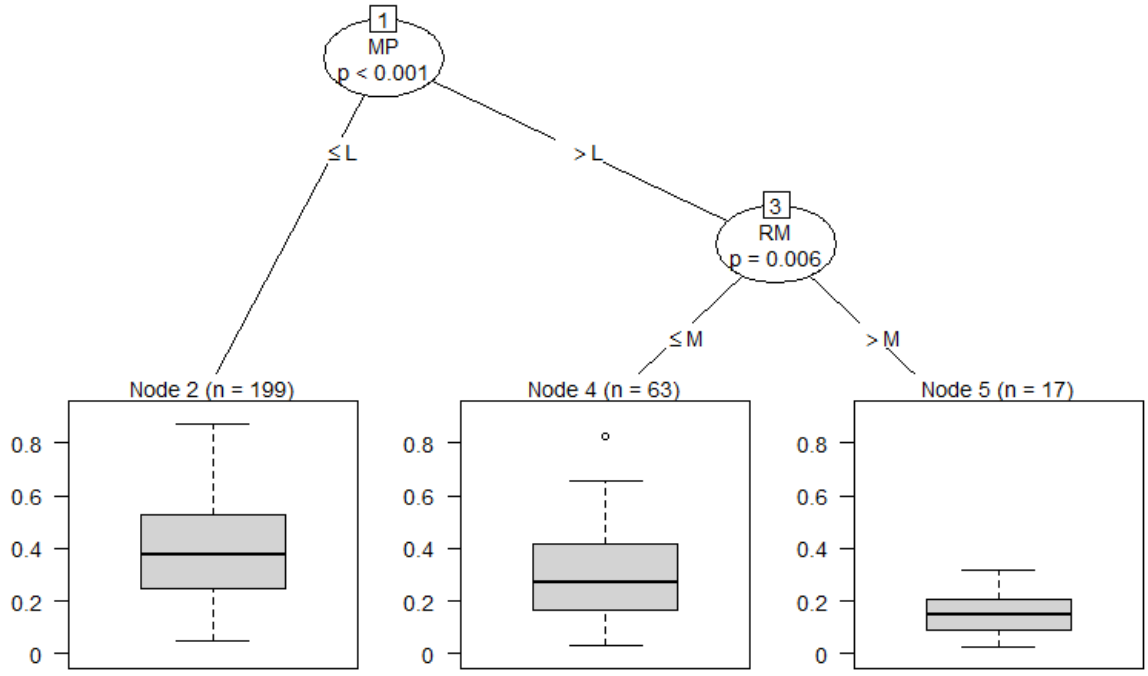
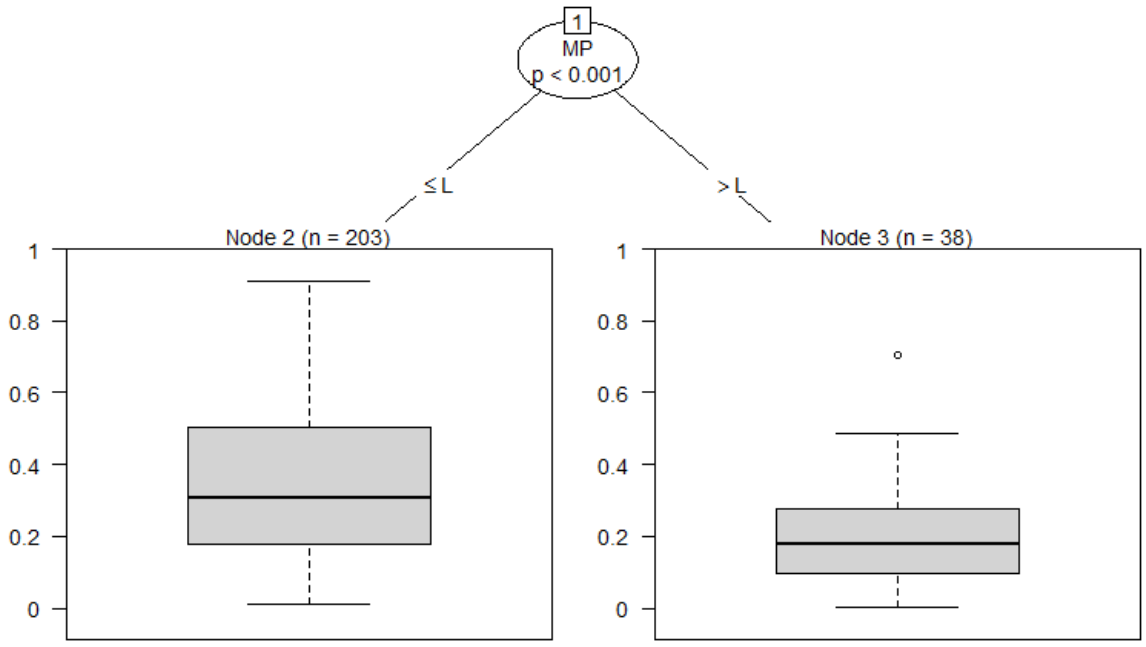


Figure D.3. Conditional Tree for Predicting Grade 5 Mathematics P-Values from Cognitive Complexity Source Codes.



4

Figure D.4. Conditional Tree for Predicting Grade 6 Mathematics P-Values from Cognitive Complexity Source Codes.

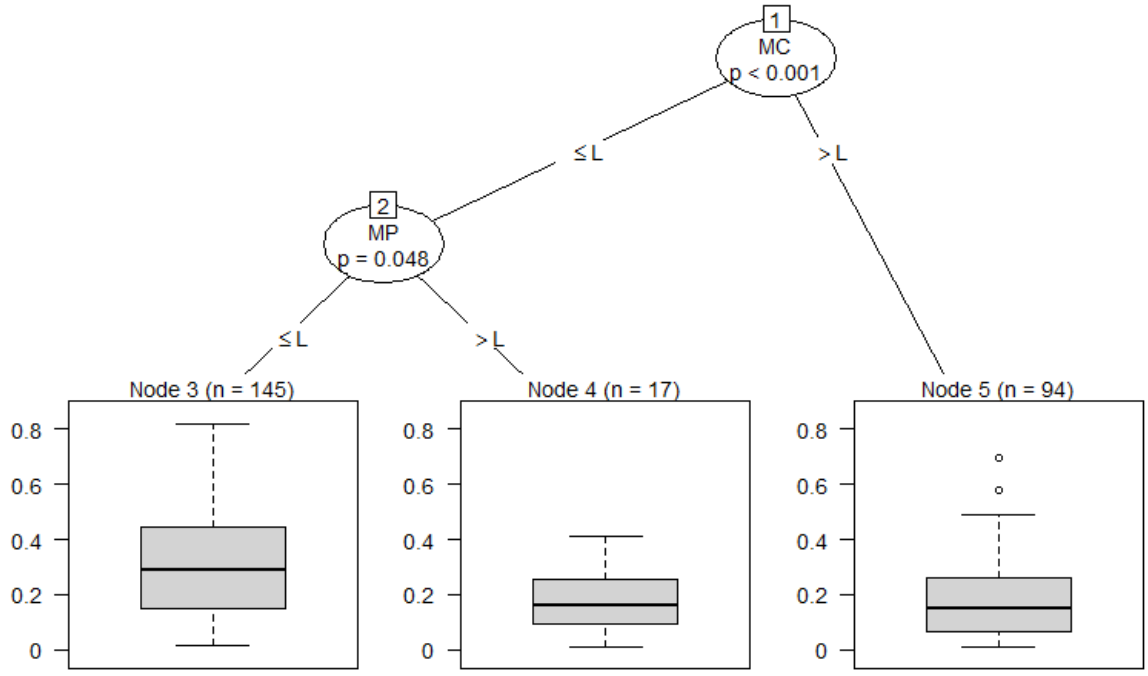


Figure D.5. Conditional Tree for Predicting Grade 7 Mathematics P-Values from Cognitive Complexity Source Codes.

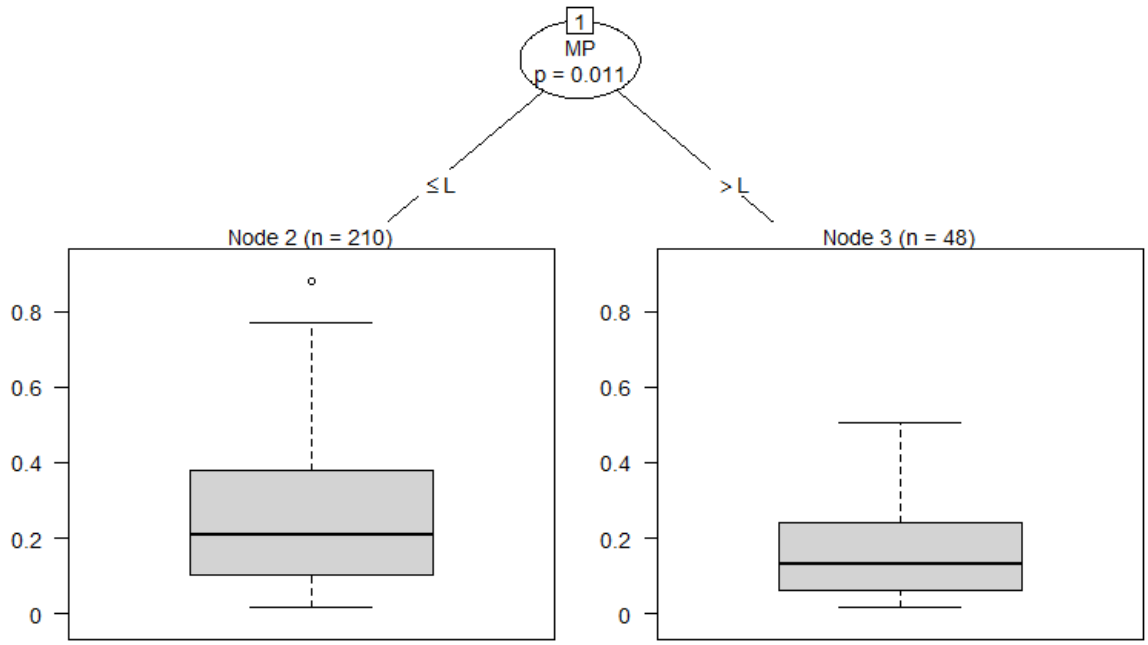


Figure D.6. Conditional Tree for Predicting Grade 8 Mathematics P-Values from Cognitive Complexity Source Codes.

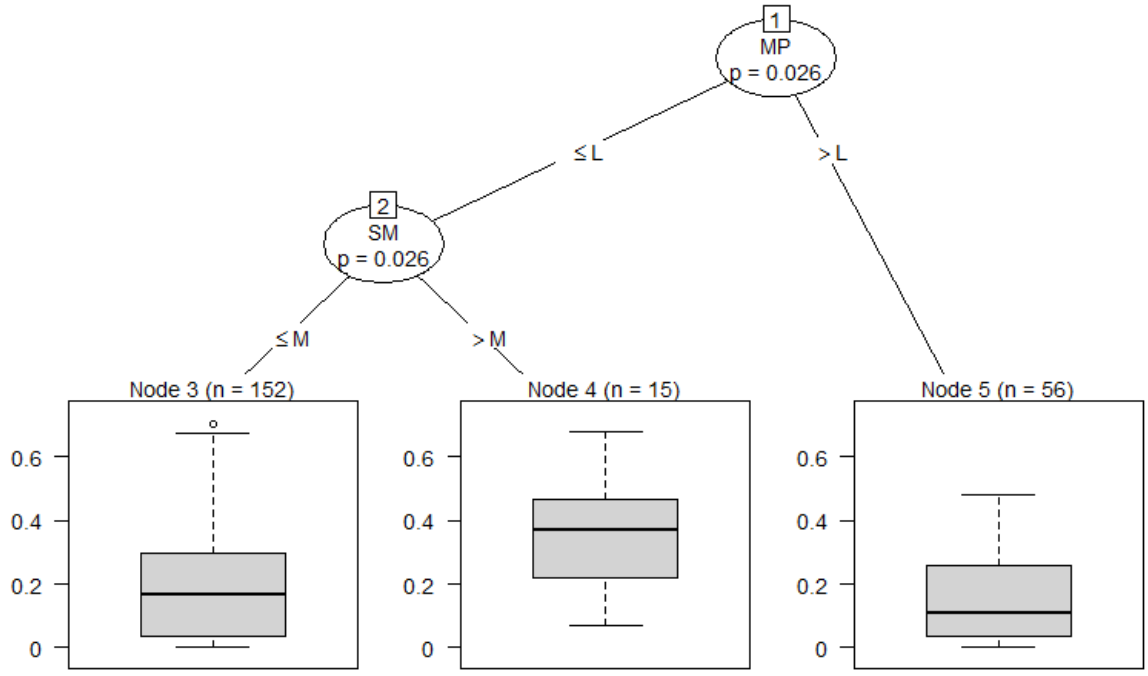


Figure D.7. Conditional Tree for Predicting Algebra I P-Values from Cognitive Complexity Source Codes.

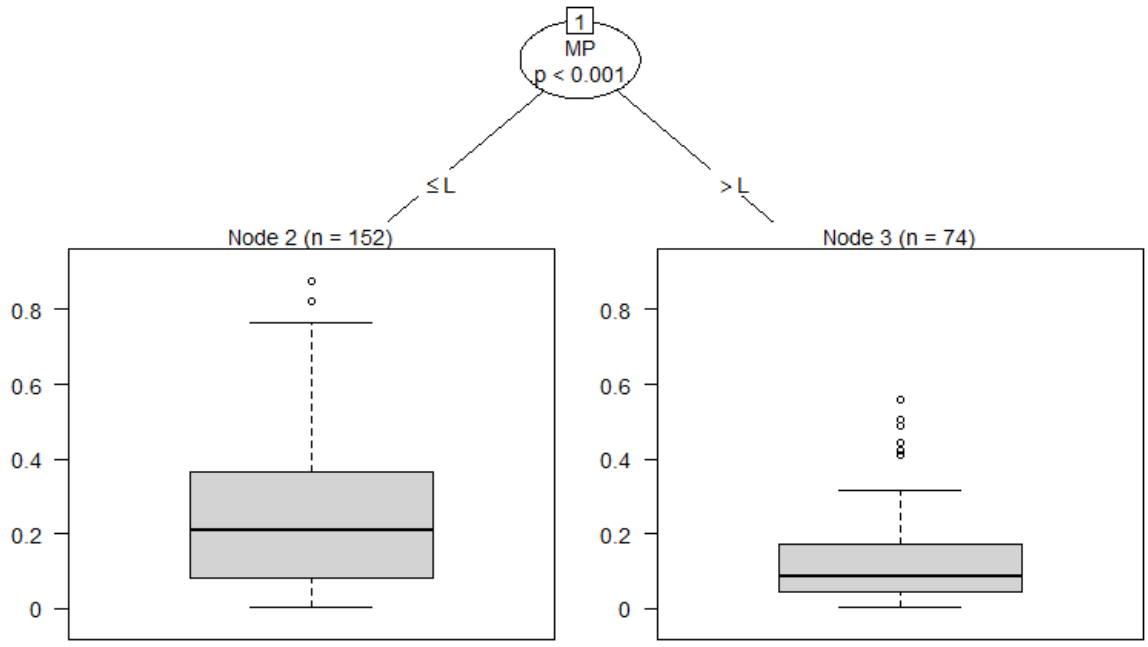


Figure D.8. Conditional Tree for Predicting Geometry P-Values from Cognitive Complexity Source Codes.

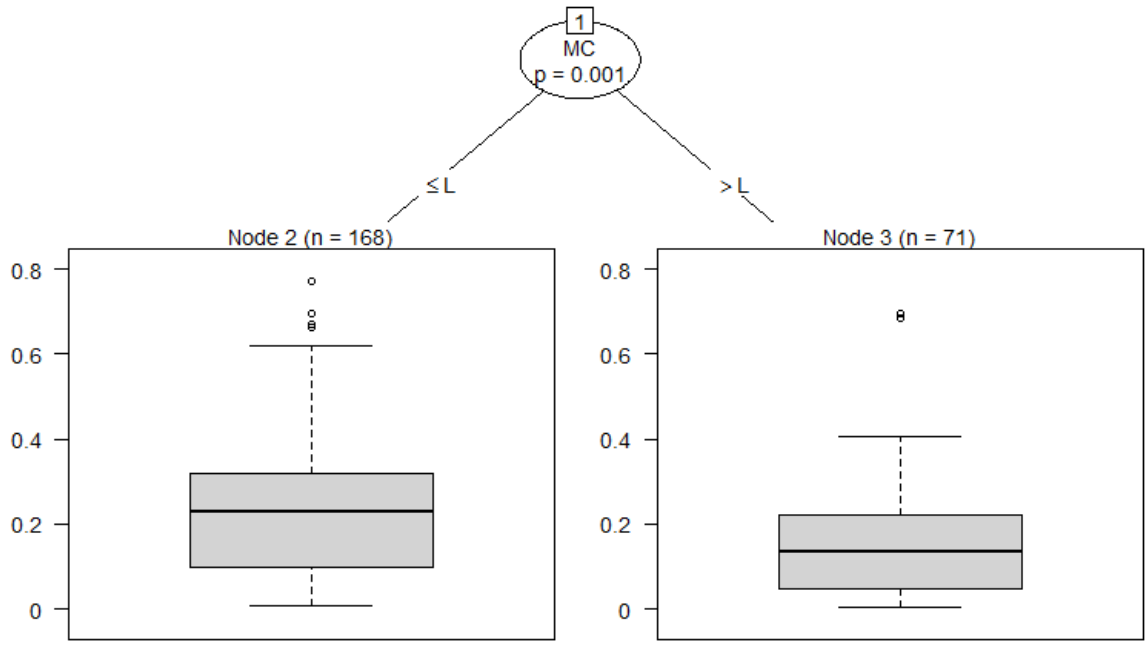


Figure D.9. Conditional Tree for Predicting Algebra II P-Values from Cognitive Complexity Source Codes.

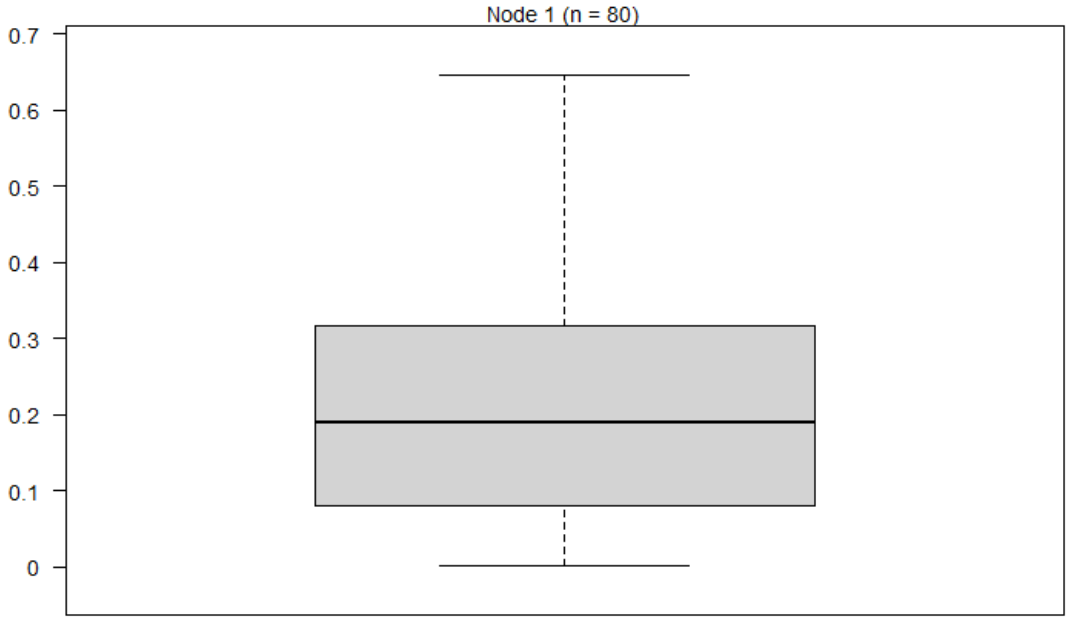


Figure D.10. Conditional Tree for Predicting Integrated Mathematics 1 P-Values from Cognitive Complexity Source Codes.

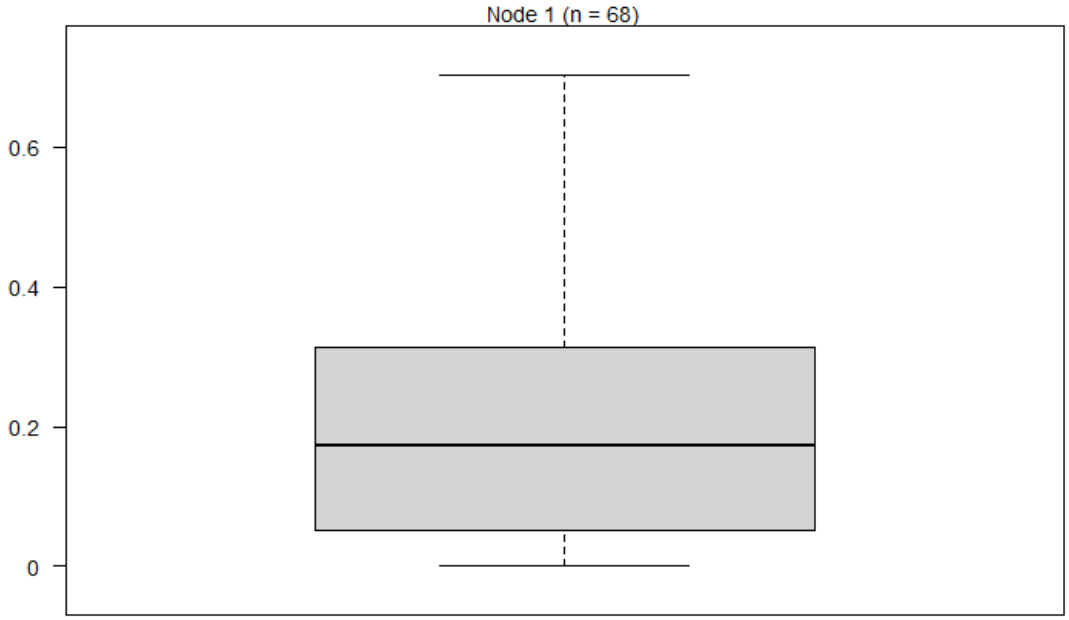


Figure D.11. Conditional Tree for Predicting Integrated Mathematics 2 P-Values from Cognitive Complexity Source Codes.

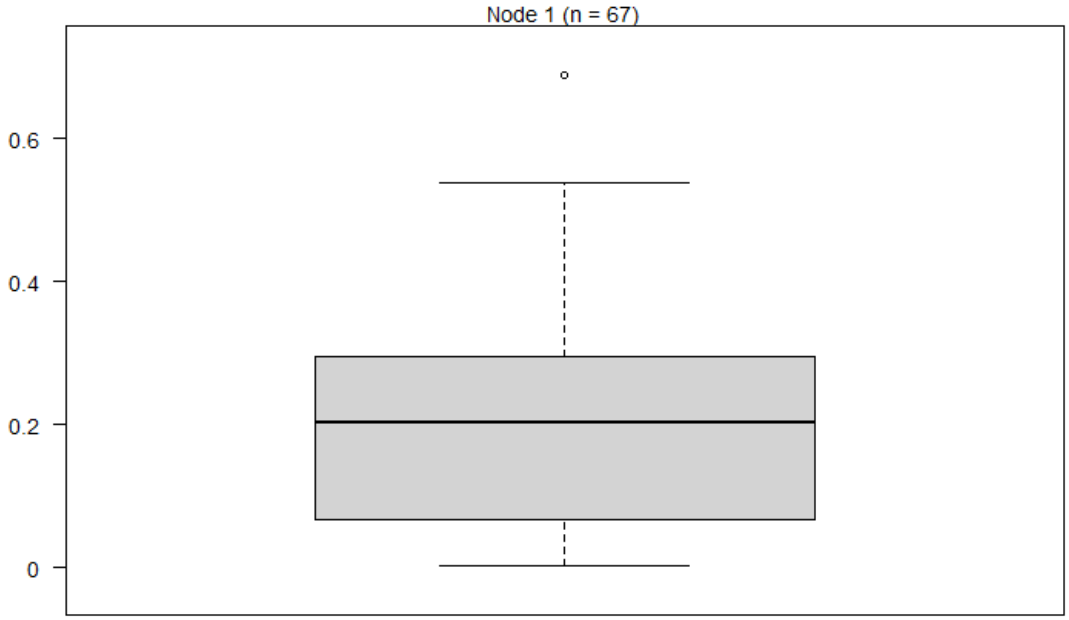


Figure D.12. Conditional Tree for Predicting Integrated Mathematics 3 P-Values from Cognitive Complexity Source Codes.

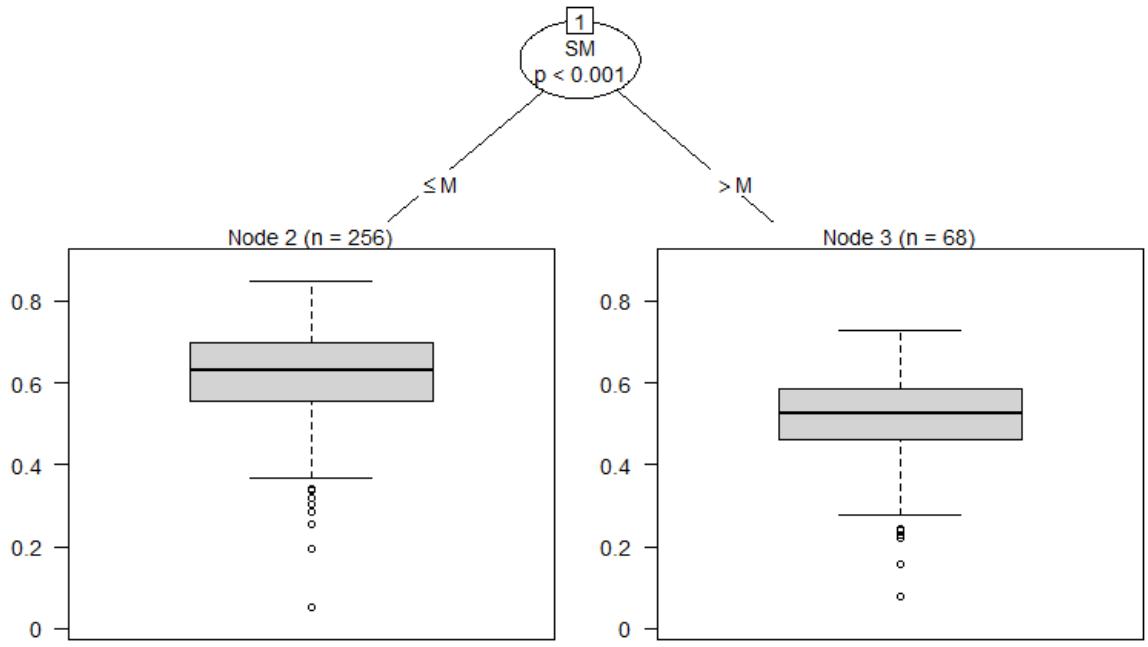


Figure D.13. Conditional Tree for Predicting Grade 3 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes.

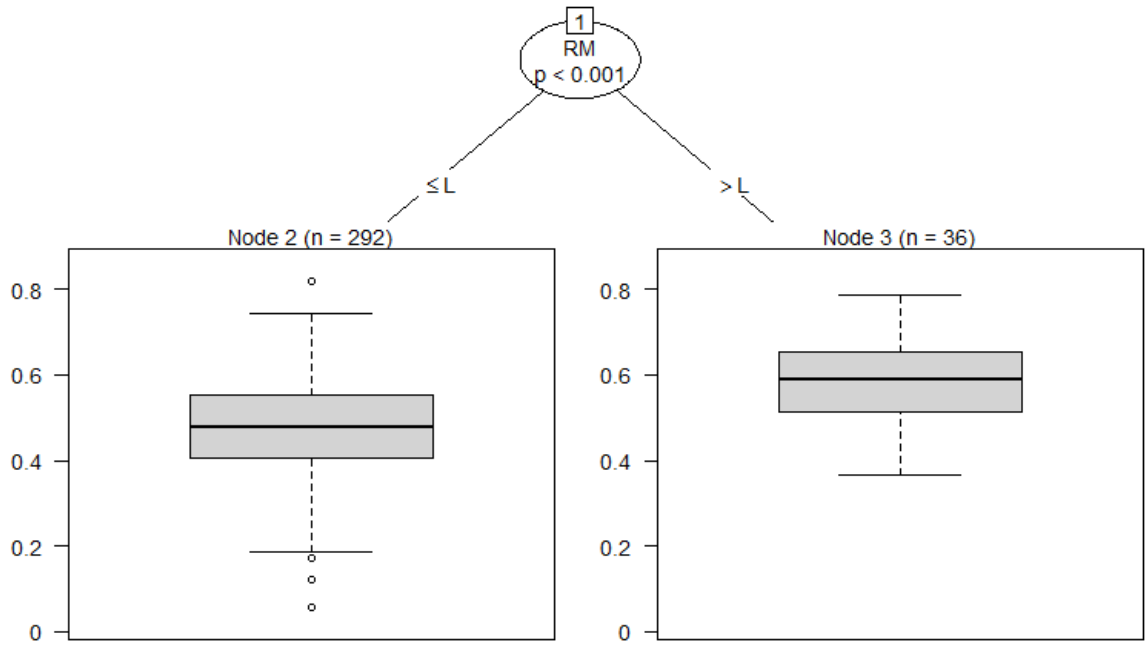


Figure D.14. Conditional Tree for Predicting Grade 4 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes.

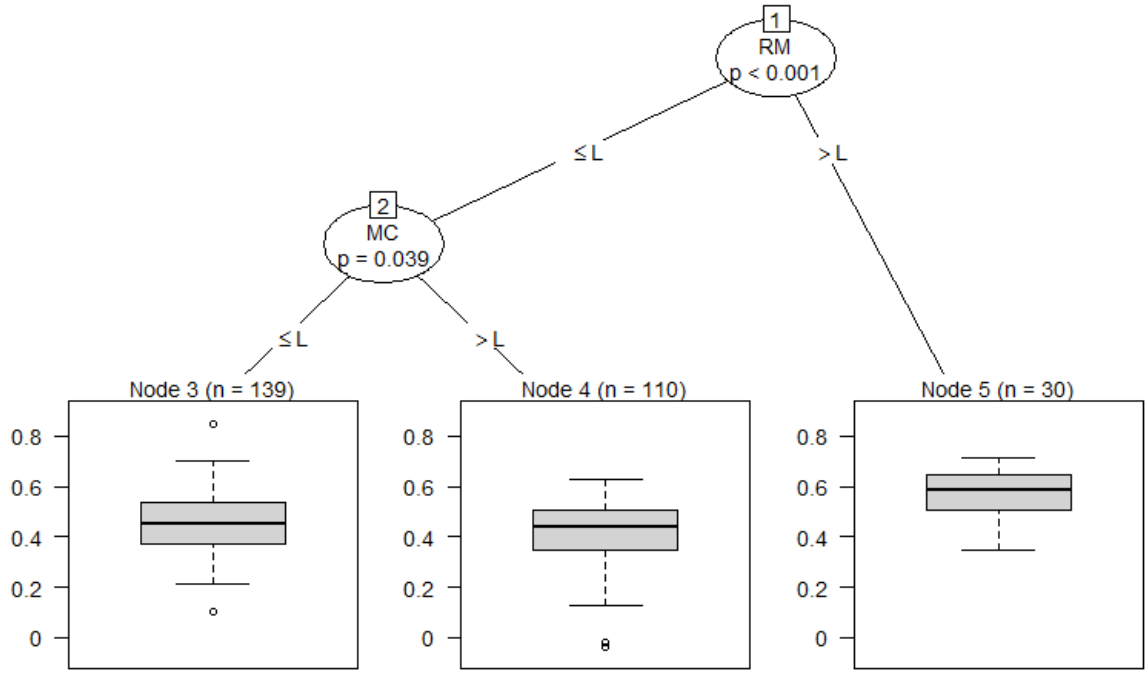


Figure D.15. Conditional Tree for Predicting Grade 5 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes.

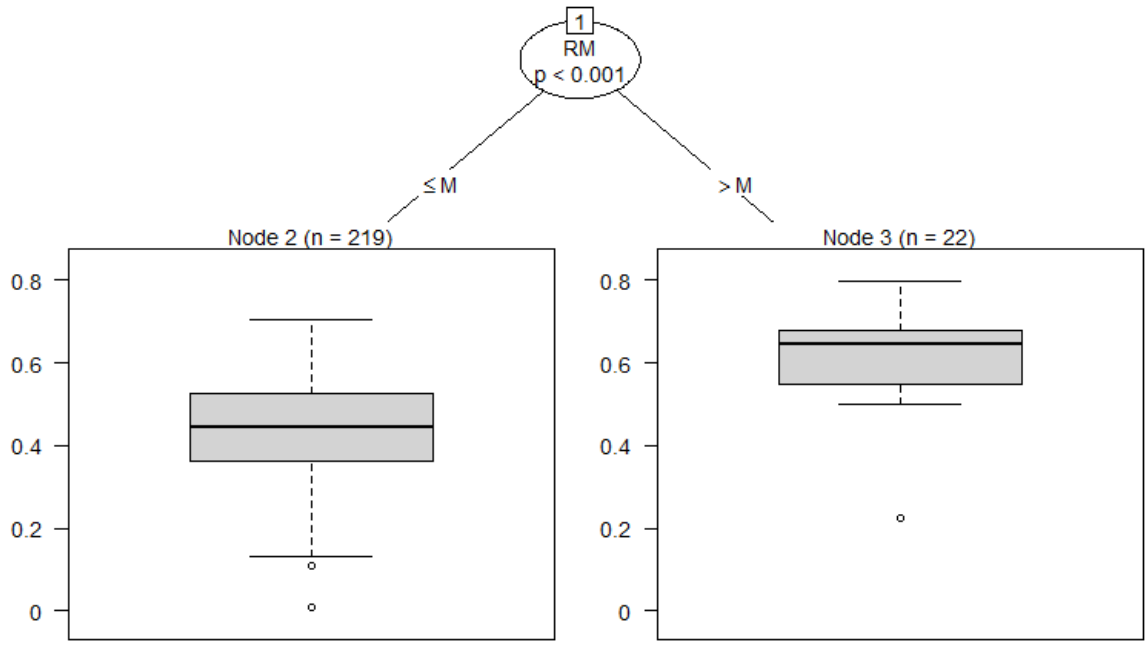


Figure D.16. Conditional Tree for Predicting Grade 6 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes.

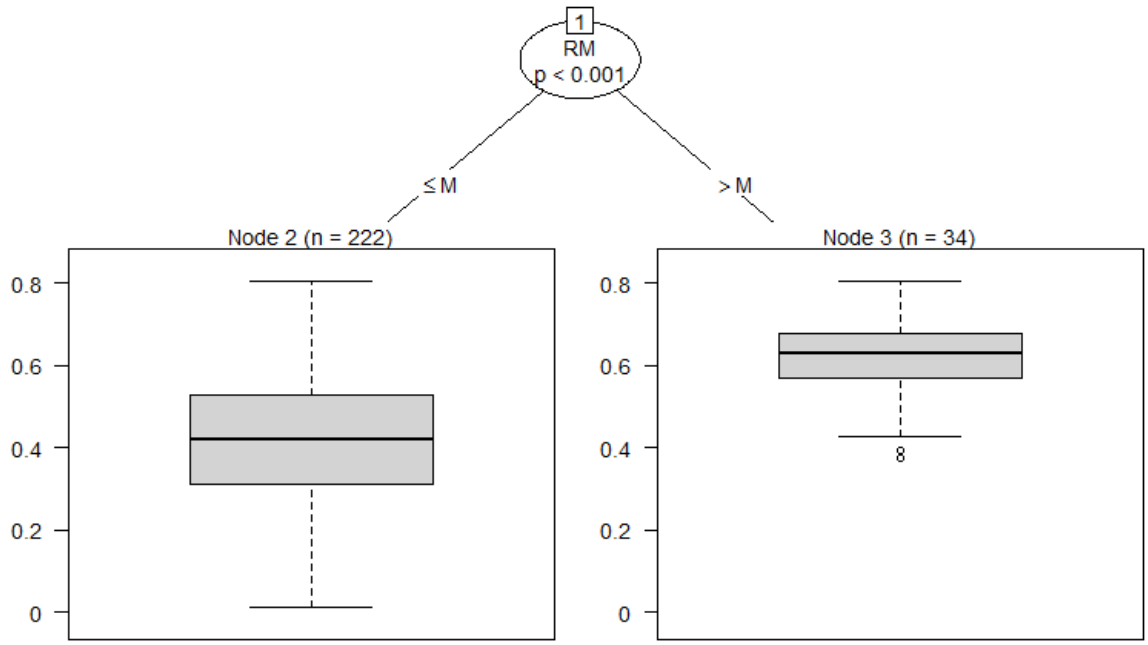


Figure D.17. Conditional Tree for Predicting Grade 7 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes.

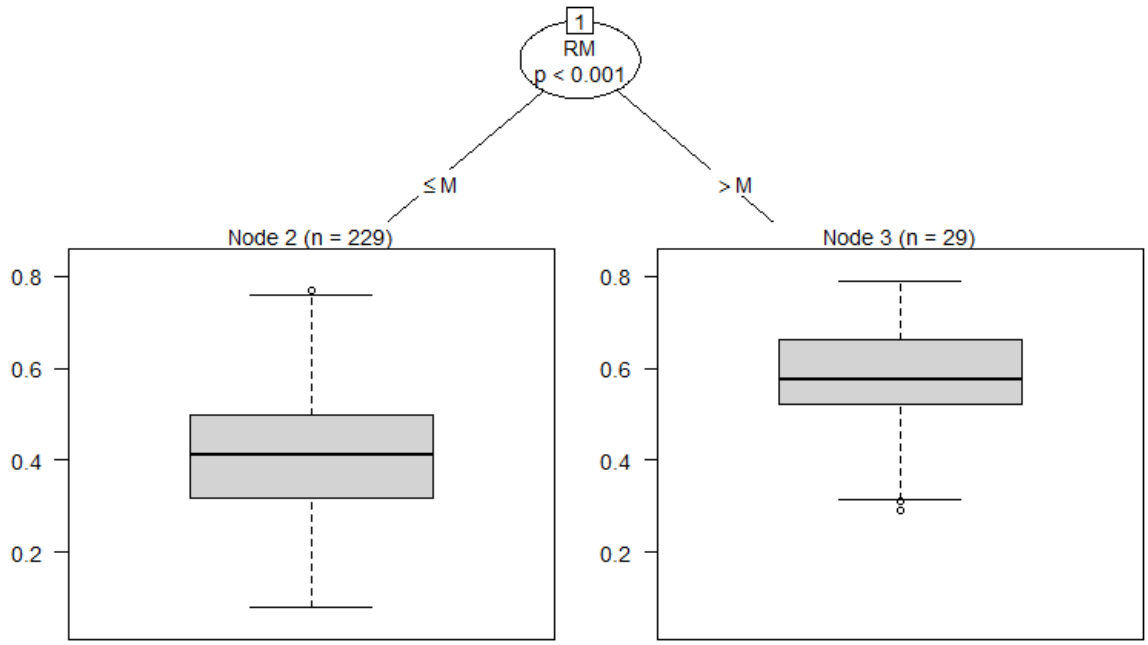


Figure D.18. Conditional Tree for Predicting Grade 8 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes.

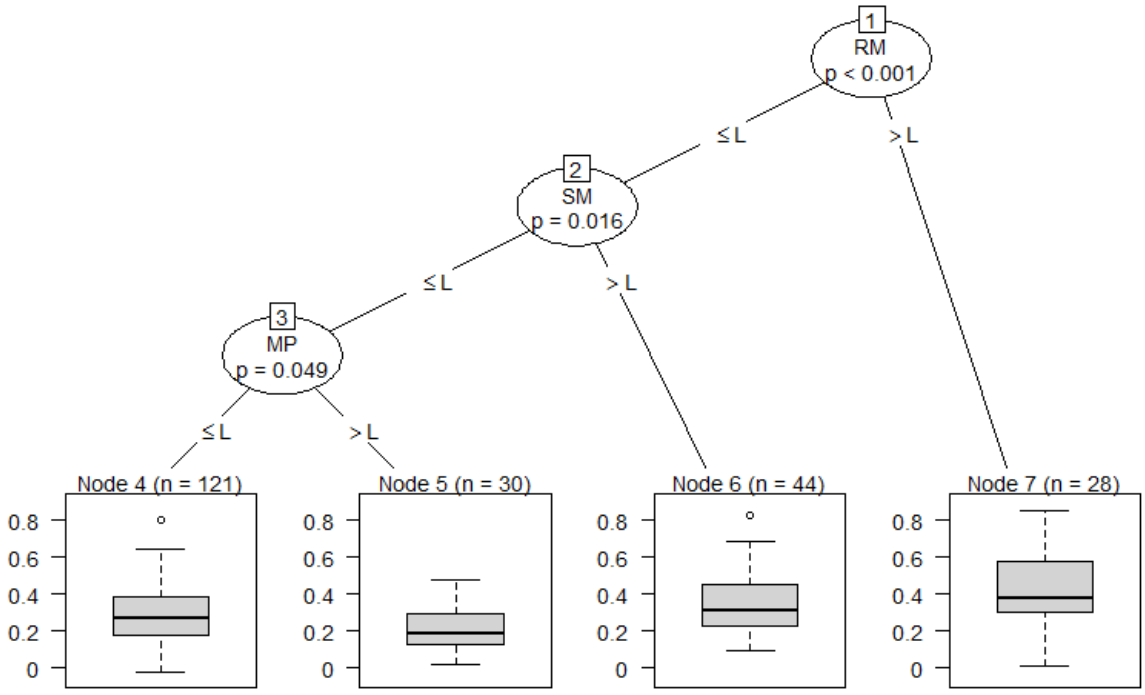


Figure D.19. Conditional Tree for Predicting Algebra I Point-Biserial Correlations from Cognitive Complexity Source Codes.

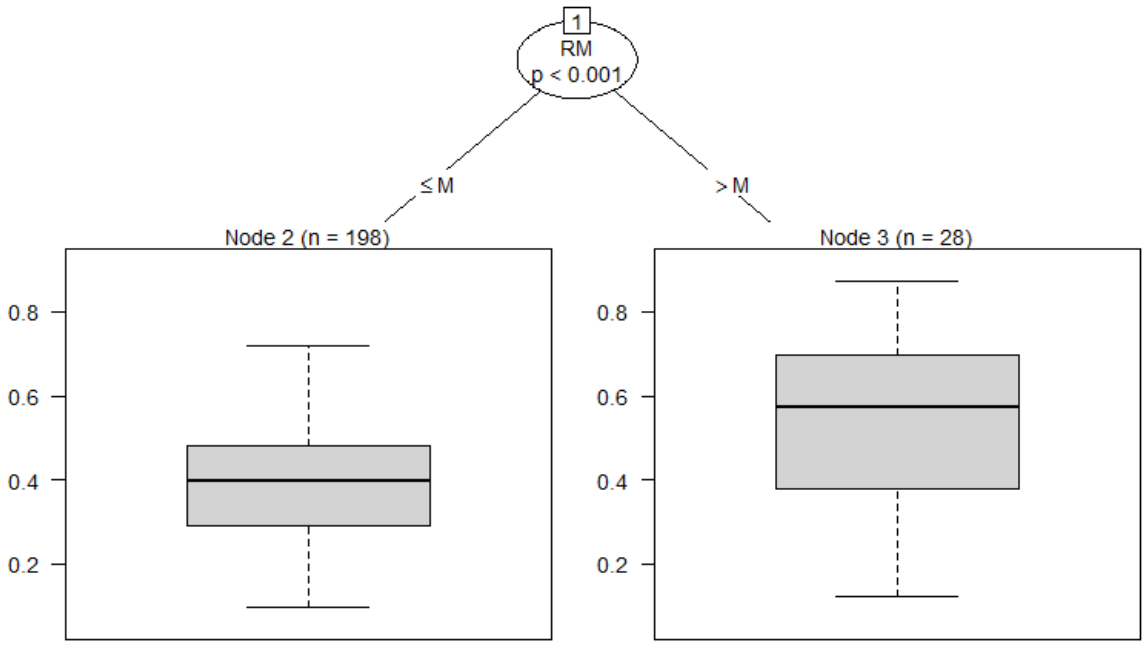


Figure D.20. Conditional Tree for Predicting Geometry Point-Biserial Correlations from Cognitive Complexity Source Codes.

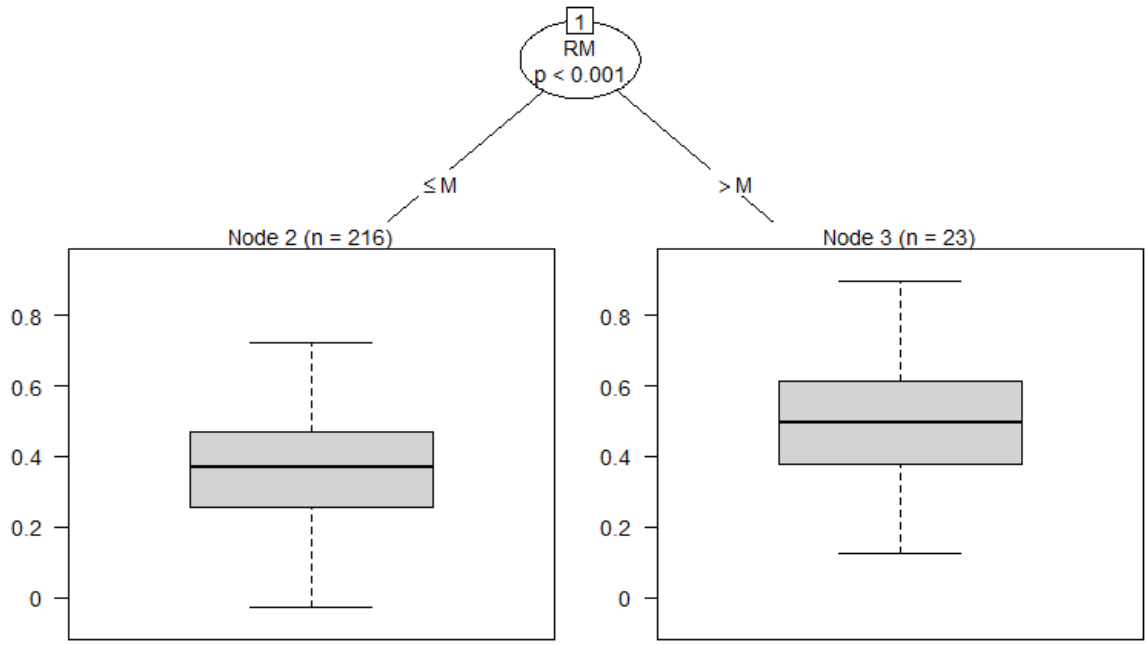


Figure D.21. Conditional Tree for Predicting Algebra II Point-Biserial Correlations from Cognitive Complexity Source Codes.

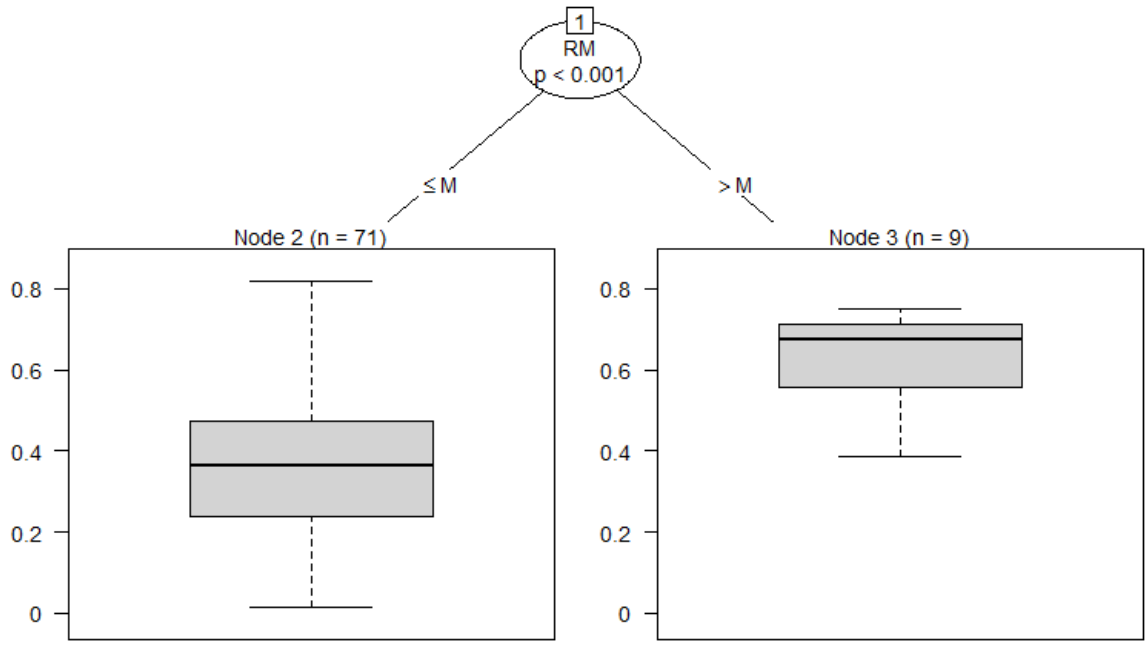


Figure D.22. Conditional Tree for Predicting Integrated Mathematics 1 Point-Biserial Correlations from Cognitive Complexity Source Codes.

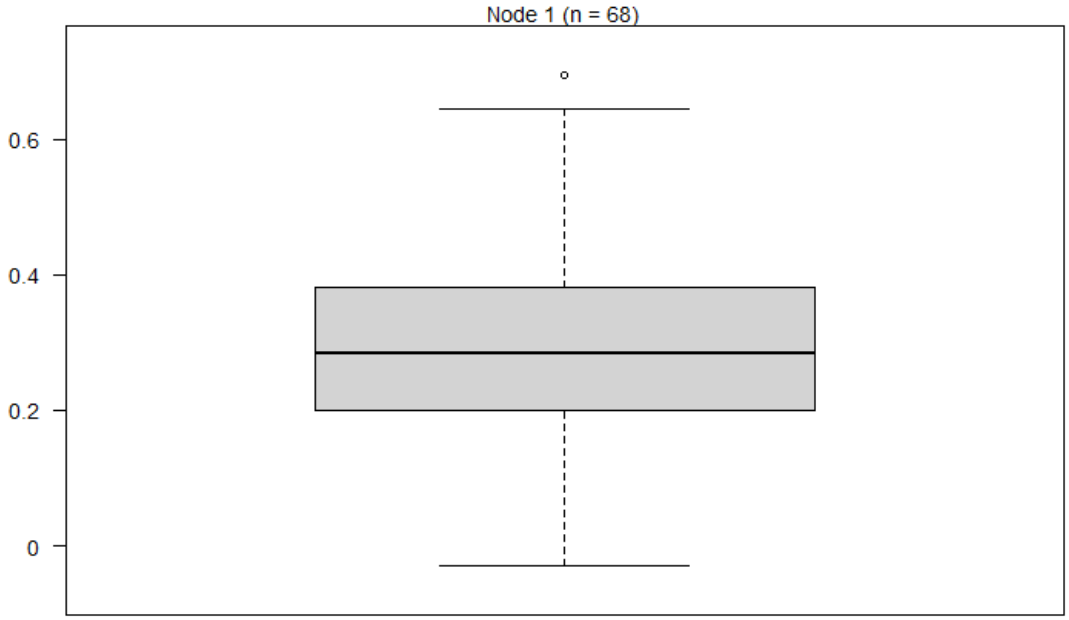


Figure D.23. Conditional Tree for Predicting Integrated Mathematics 2 Point-Biserial Correlations from Cognitive Complexity Source Codes.

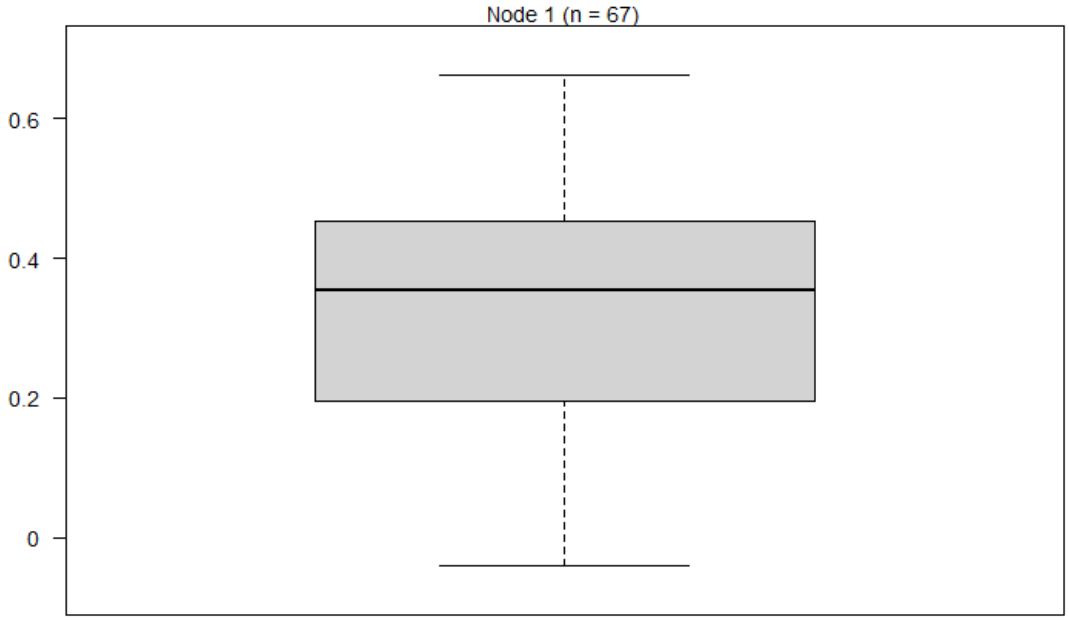


Figure D.24. Conditional Tree for Predicting Integrated Mathematics 3 Point-Biserial Correlations from Cognitive Complexity Source Codes.

Appendix E: ELA/L Conditional Trees using Cognitive Complexity Source Codes as Predictors

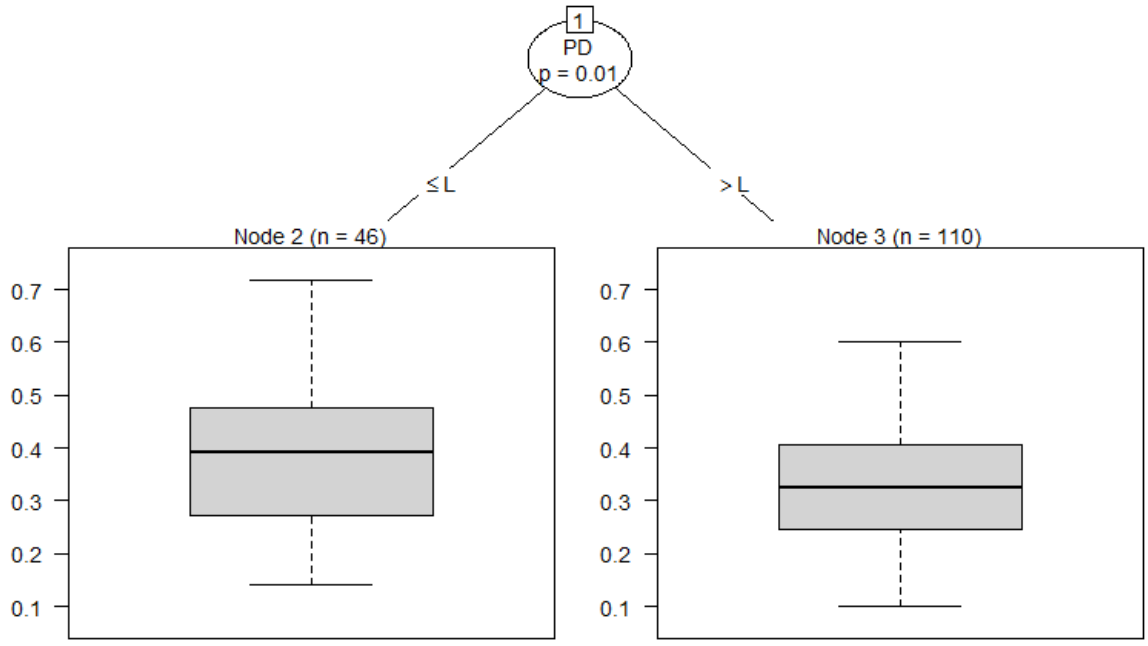


Figure E.1. Conditional Tree for Predicting Grade 3 ELA/L P-Values from Cognitive Complexity Source Codes.

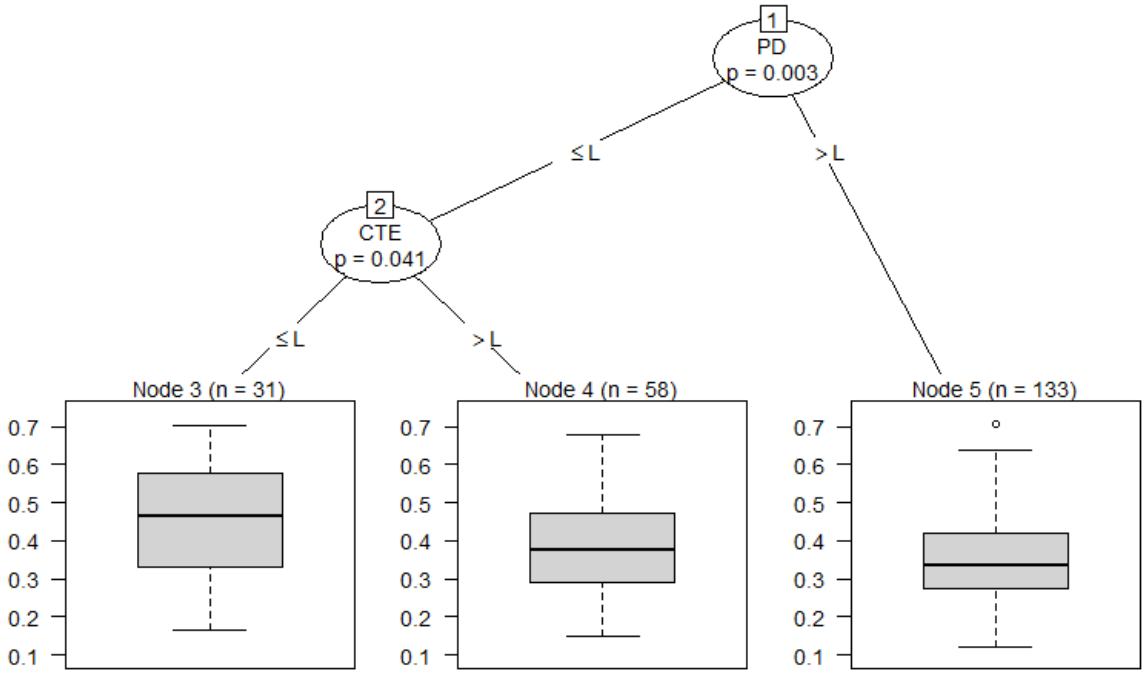


Figure E.2. Conditional Tree for Predicting Grade 4 ELA/L P-Values from Cognitive Complexity Source Codes.

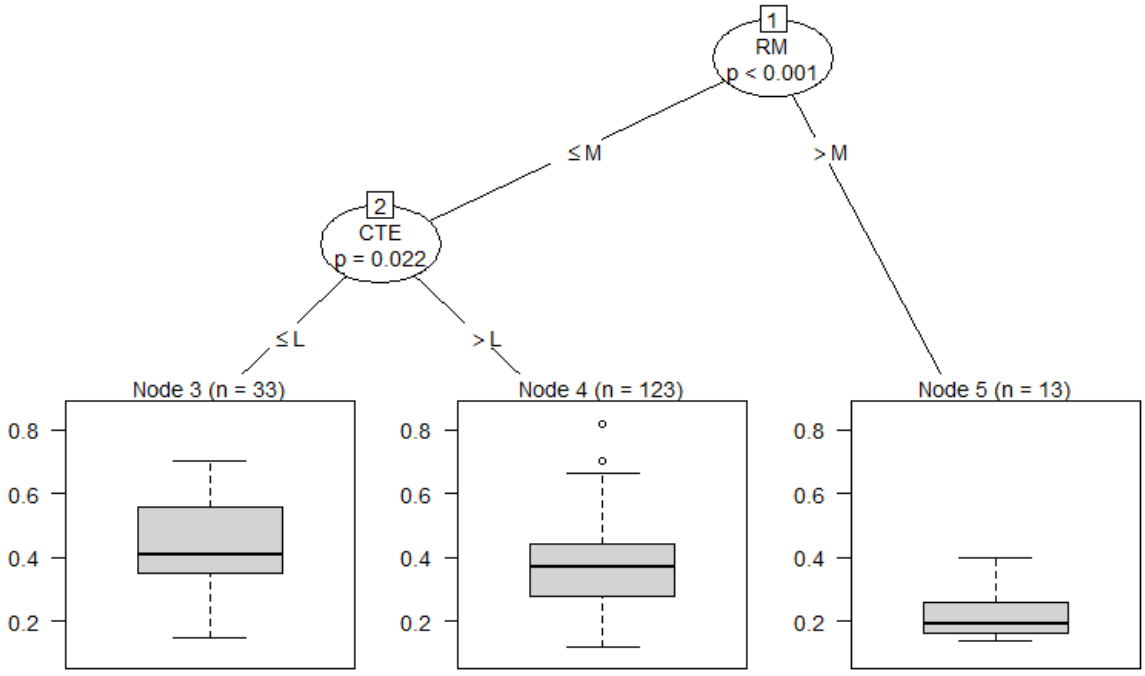


Figure E.3. Conditional Tree for Predicting Grade 5 ELA/L P-Values from Cognitive Complexity Source Codes.

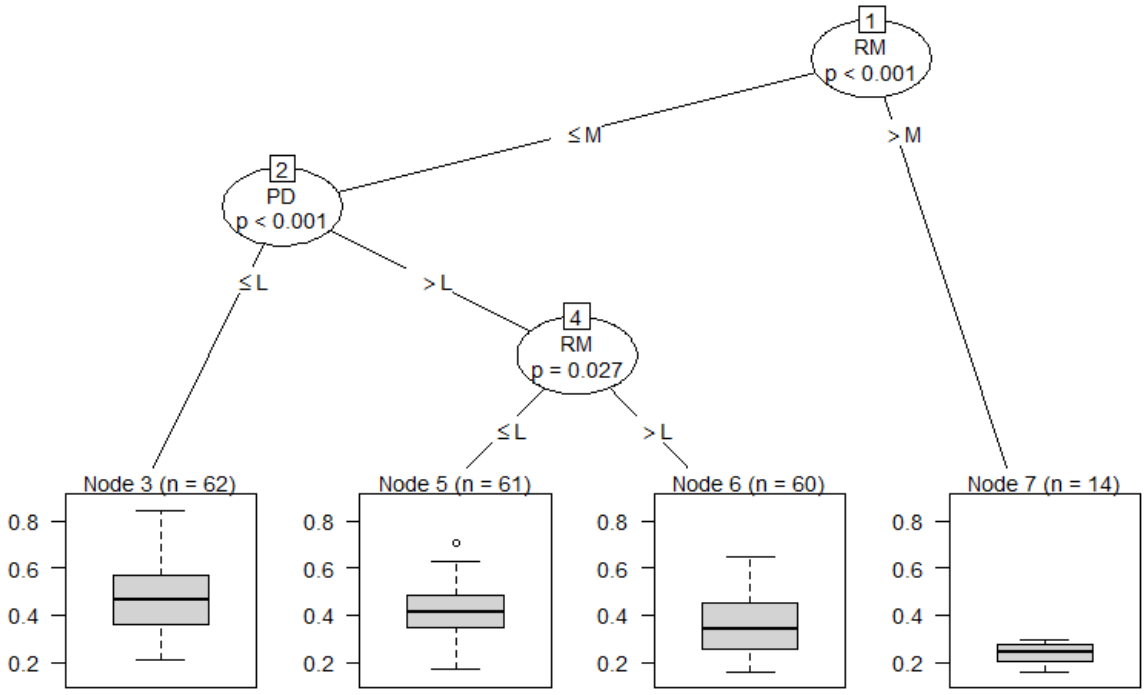


Figure E.4. Conditional Tree for Predicting Grade 6 ELA/L P-Values from Cognitive Complexity Source Codes.

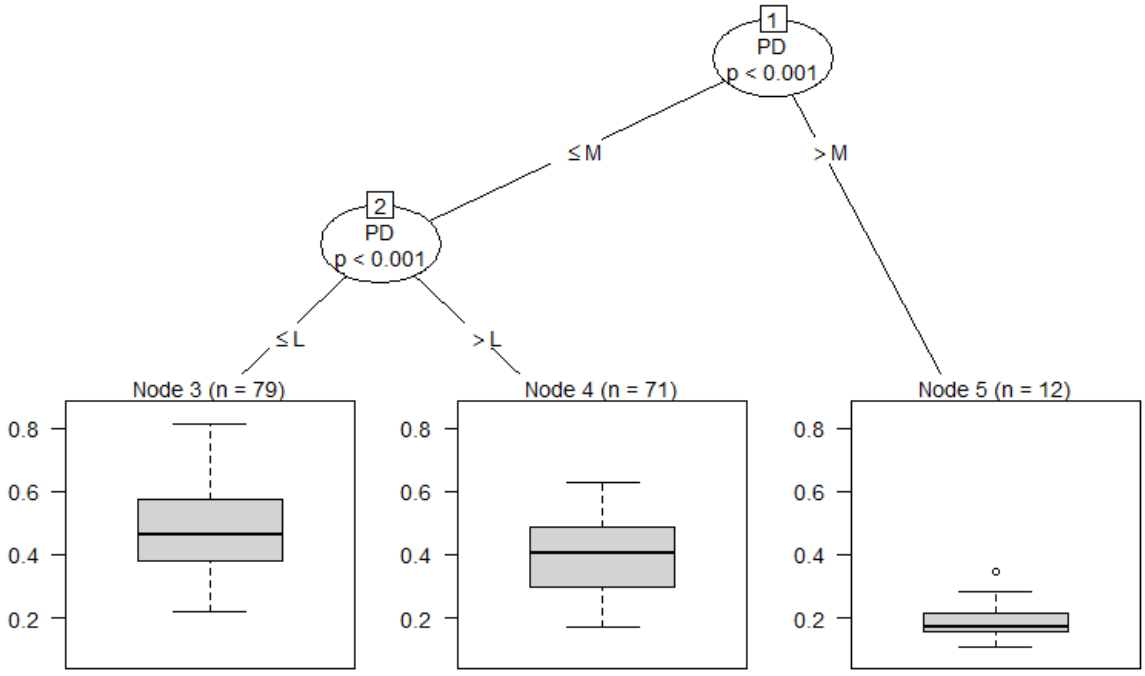


Figure E.5. Conditional Tree for Predicting Grade 7 ELA/L P-Values from Cognitive Complexity Source Codes.

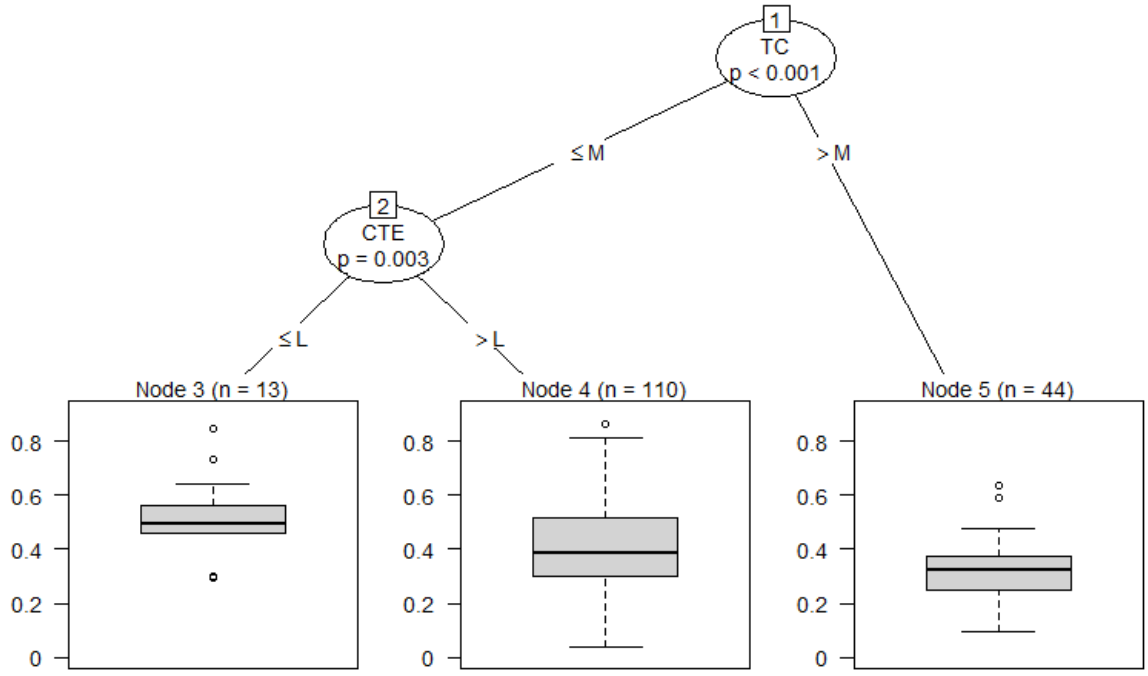


Figure E.6. Conditional Tree for Predicting Grade 8 ELA/L P-Values from Cognitive Complexity Source Codes.

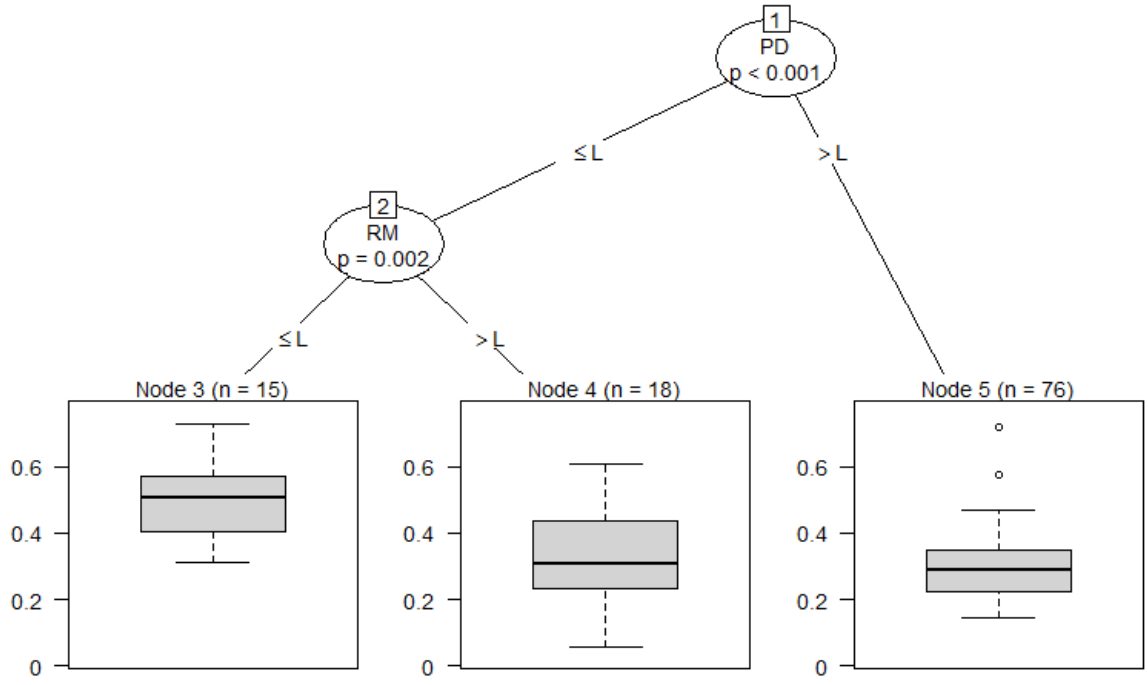


Figure E.7. Conditional Tree for Predicting Grade 9 ELA/LI P-Values from Cognitive Complexity Source Codes.

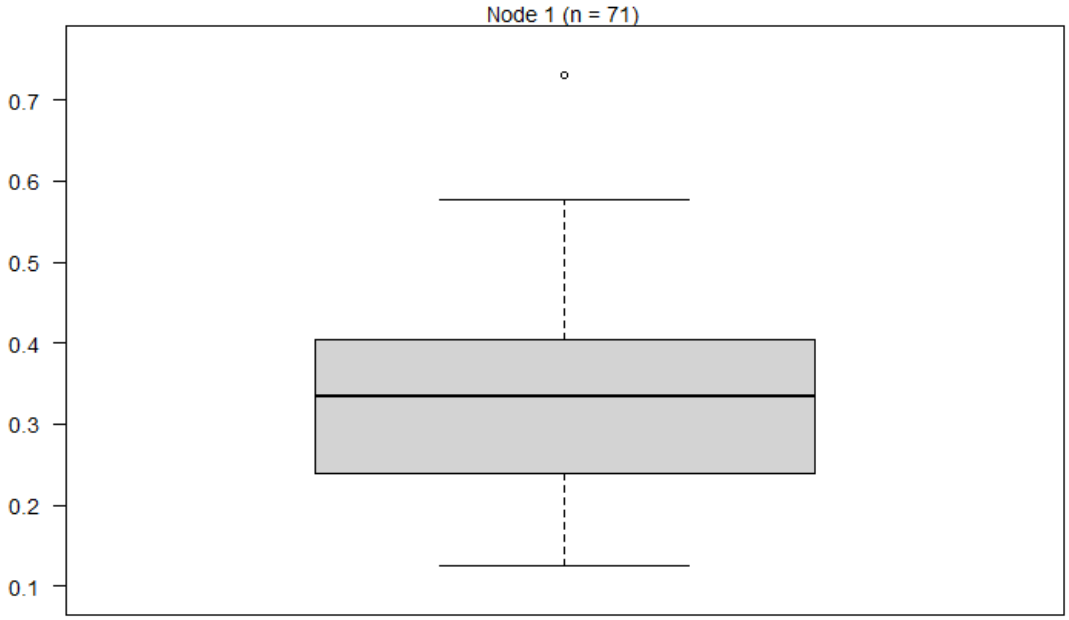


Figure E.8. Conditional Tree for Predicting Grade 10 ELA/L P-Values from Cognitive Complexity Source Codes.

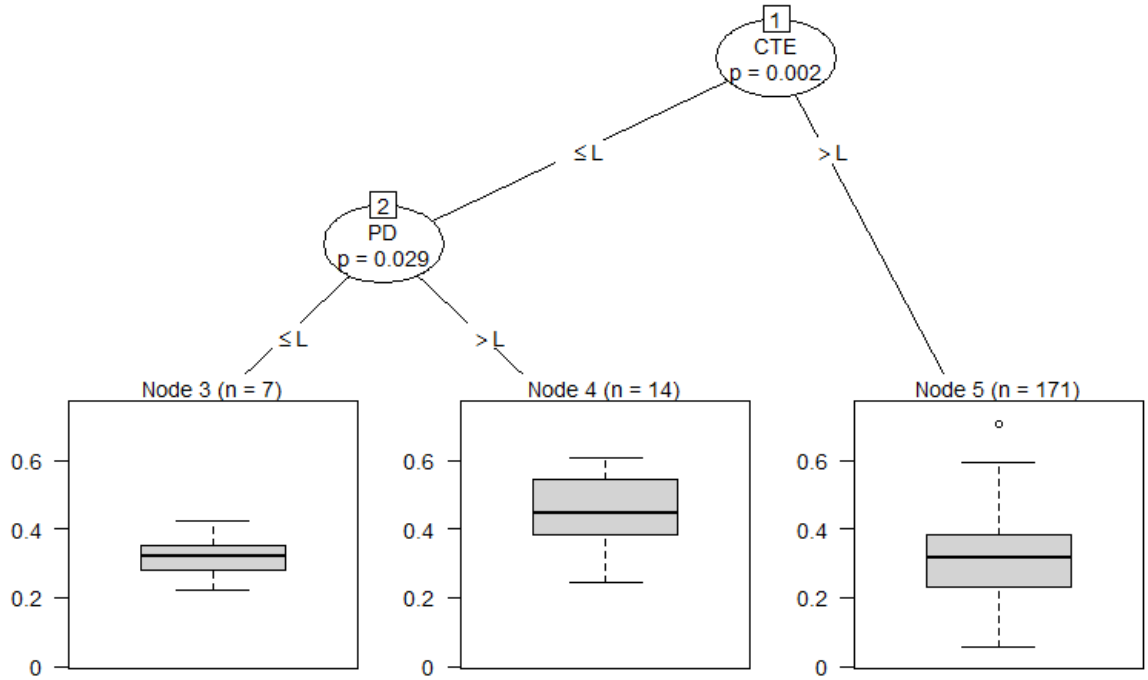


Figure E.9. Conditional Tree for Predicting Grade 11 ELA/L P-Values from Cognitive Complexity Source Codes.

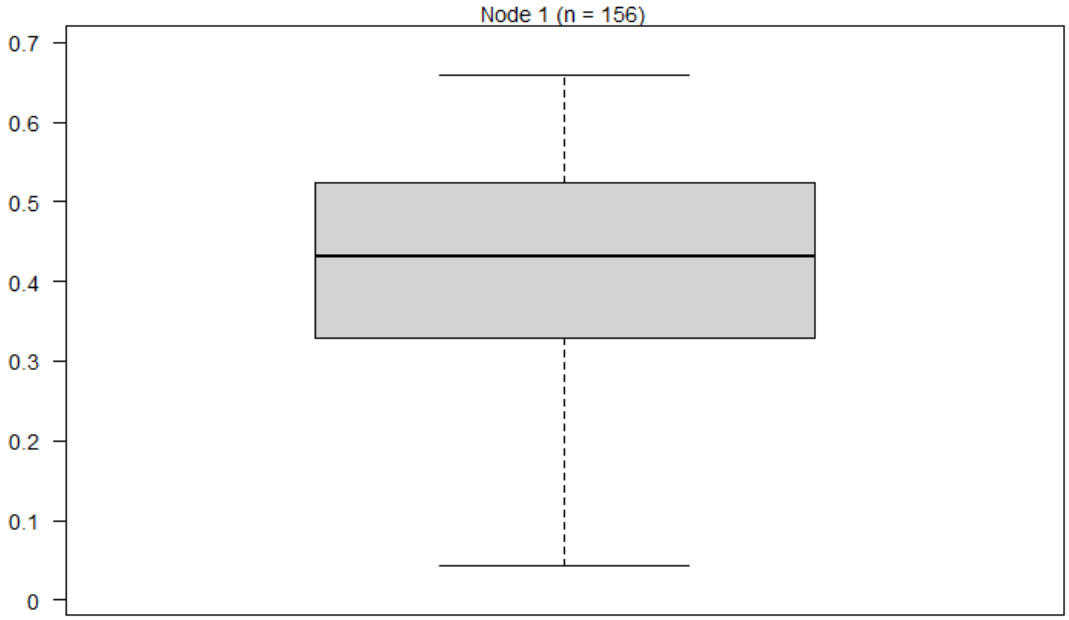


Figure E.10. Conditional Tree for Predicting Grade 3 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes.

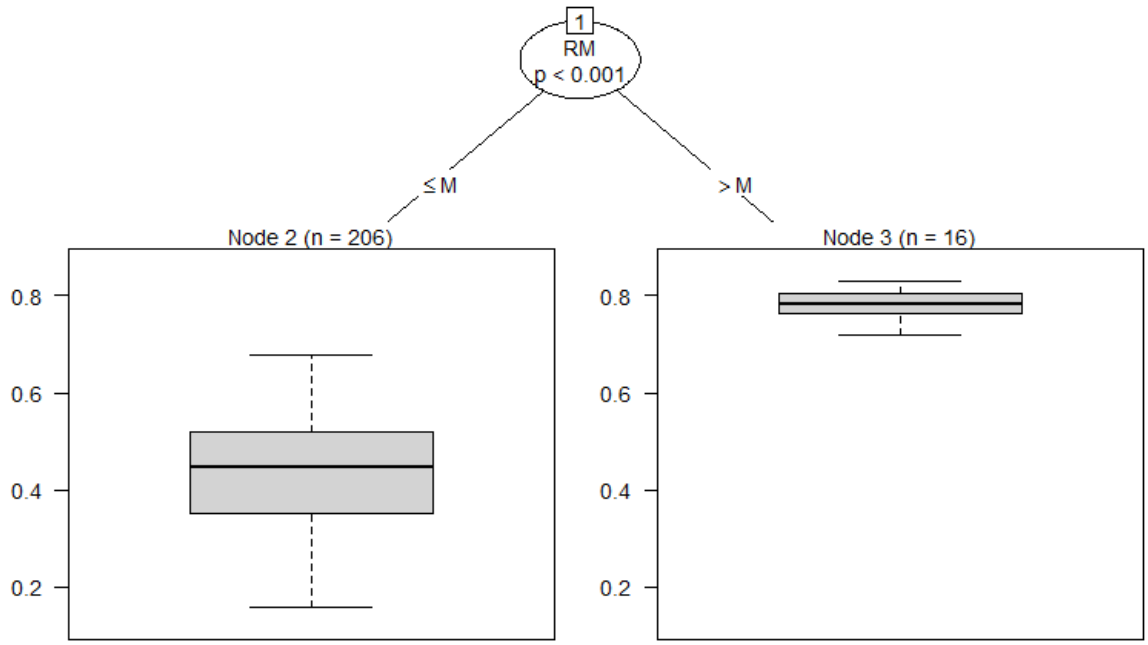


Figure E.11. Conditional Tree for Predicting Grade 4 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes.

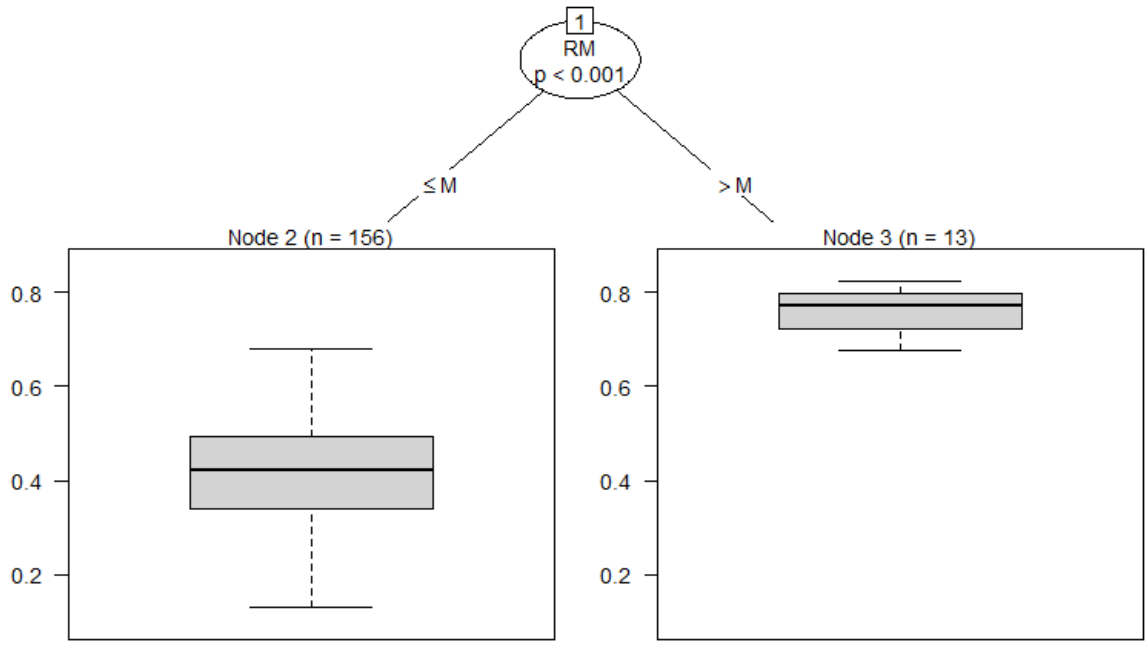


Figure E.12. Conditional Tree for Predicting Grade 5 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes.

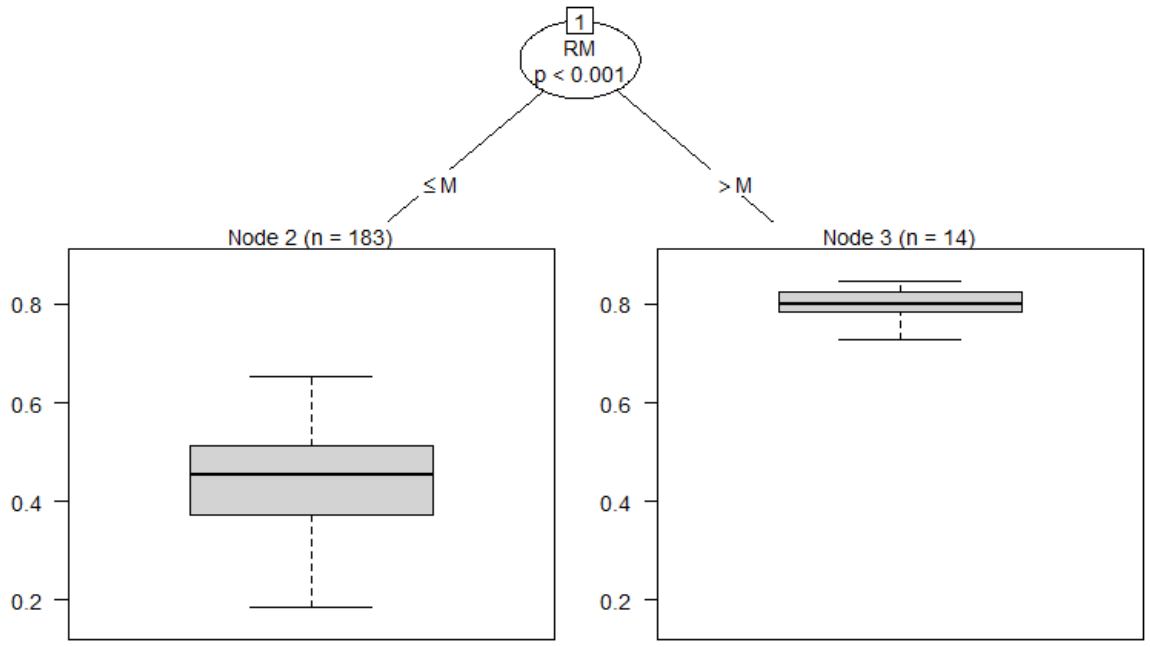


Figure E.13. Conditional Tree for Predicting Grade 6 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes.

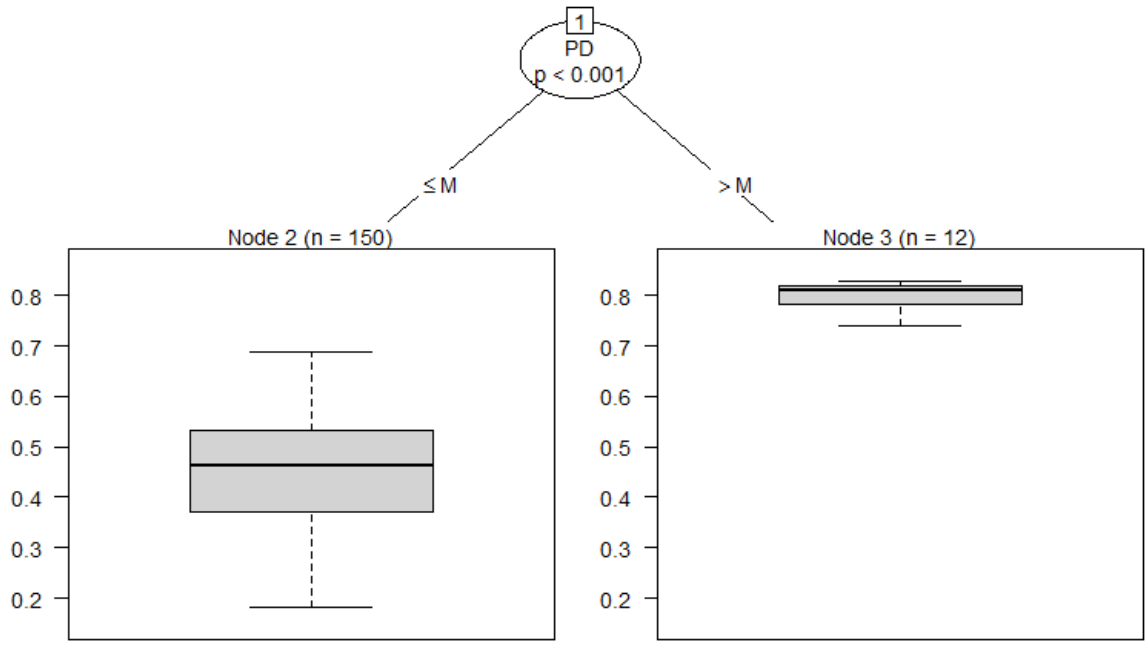


Figure E.14. Conditional Tree for Predicting Grade 7 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes.

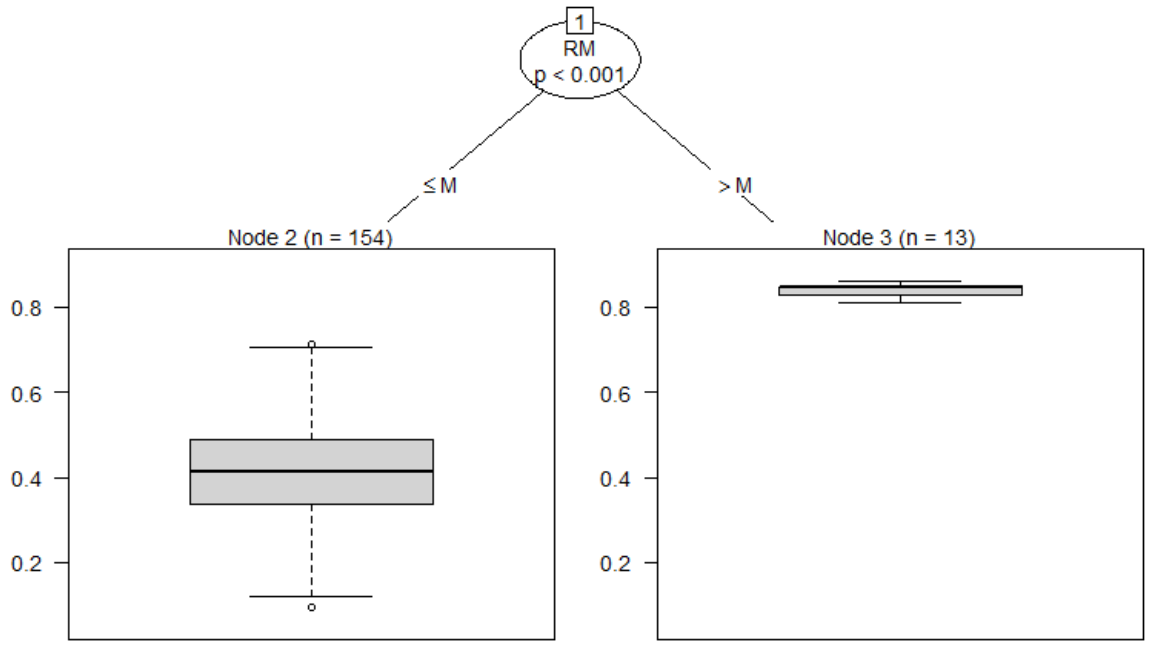


Figure E.15. Conditional Tree for Predicting Grade 8 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes.

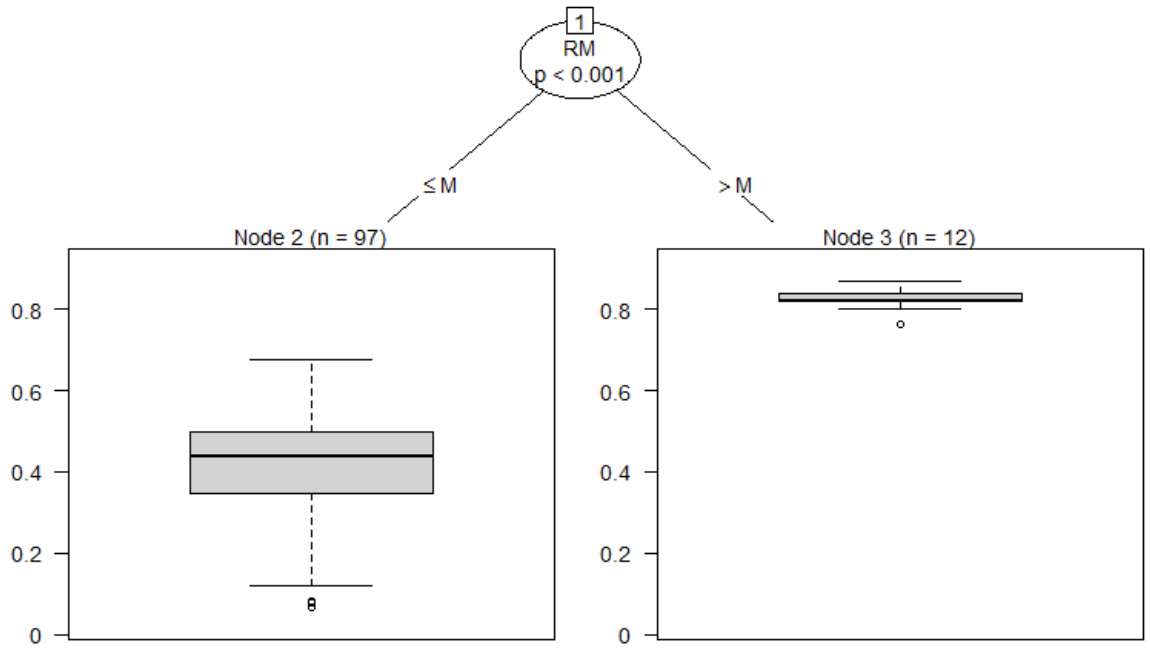


Figure E.16. Conditional Tree for Predicting Grade 9 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes.

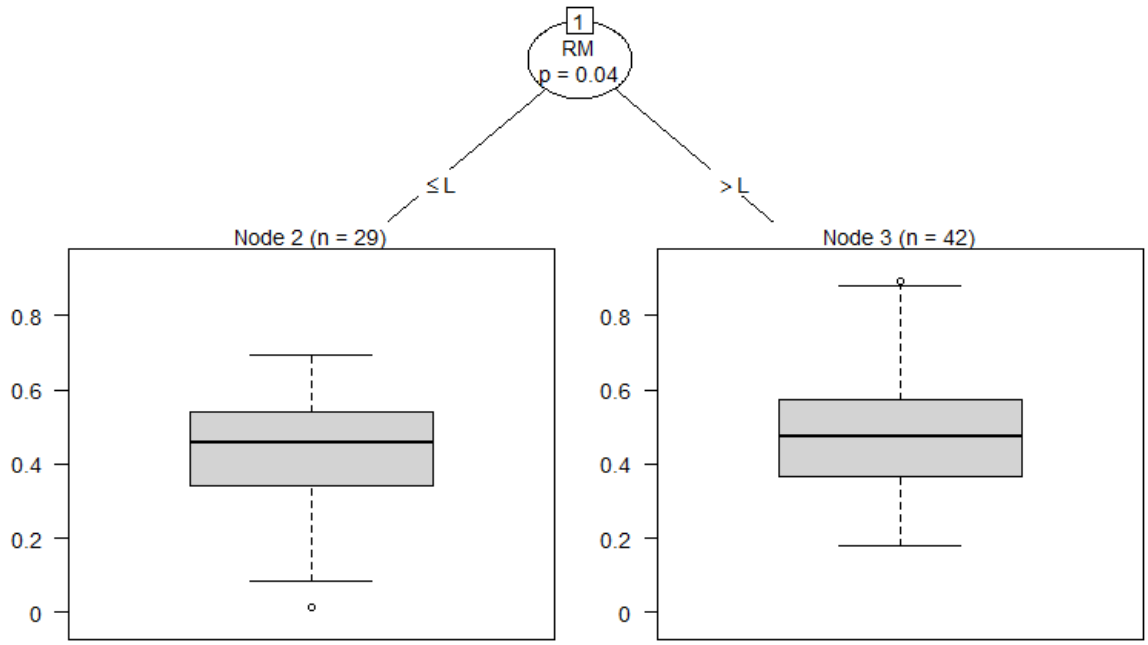


Figure E.17. Conditional Tree for Predicting Grade 10 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes.

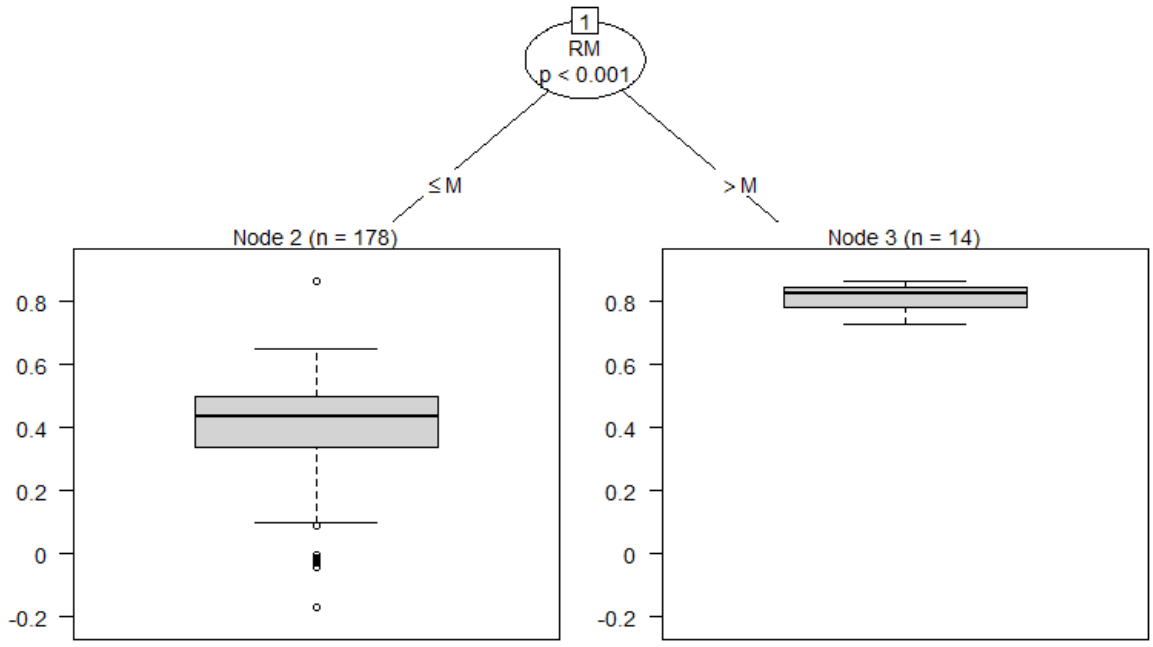


Figure E.18. Conditional Tree for Predicting Grade 11 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes.

Appendix F: Mathematics Conditional Trees using Cognitive Complexity Source Codes and Metadata as Predictors

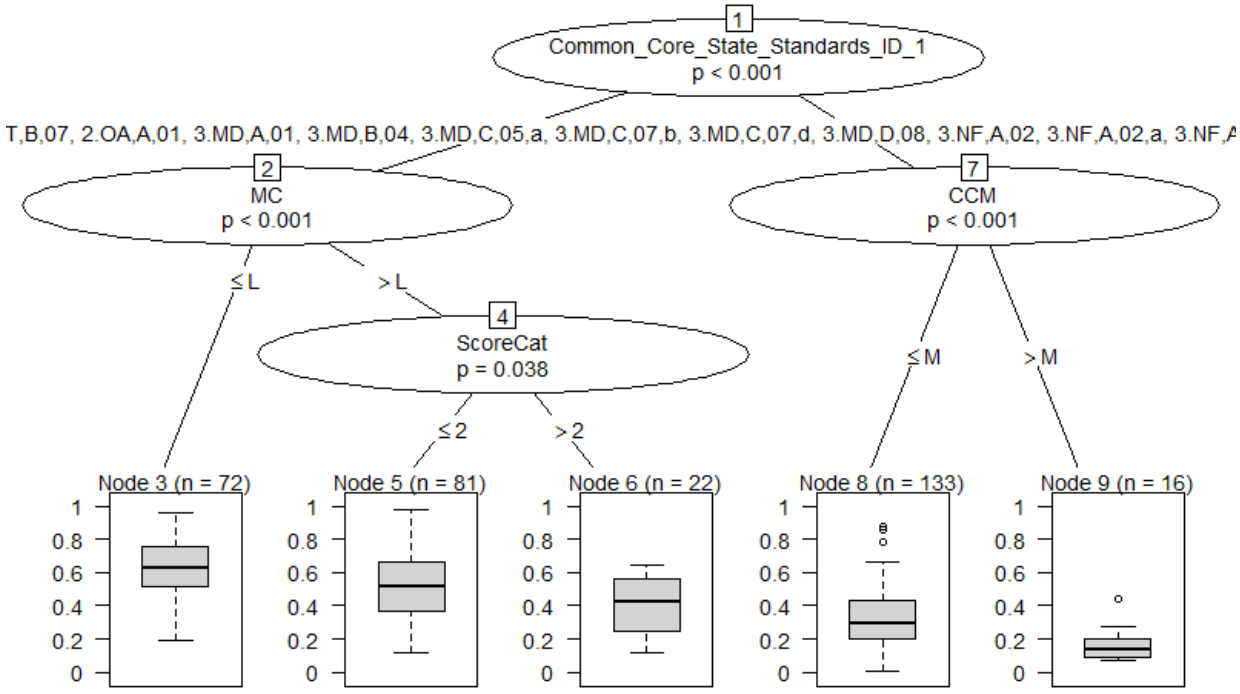


Figure F.1. Conditional Tree for Predicting Grade 3 Mathematics P-Values from Cognitive Complexity Source Codes and Metadata.

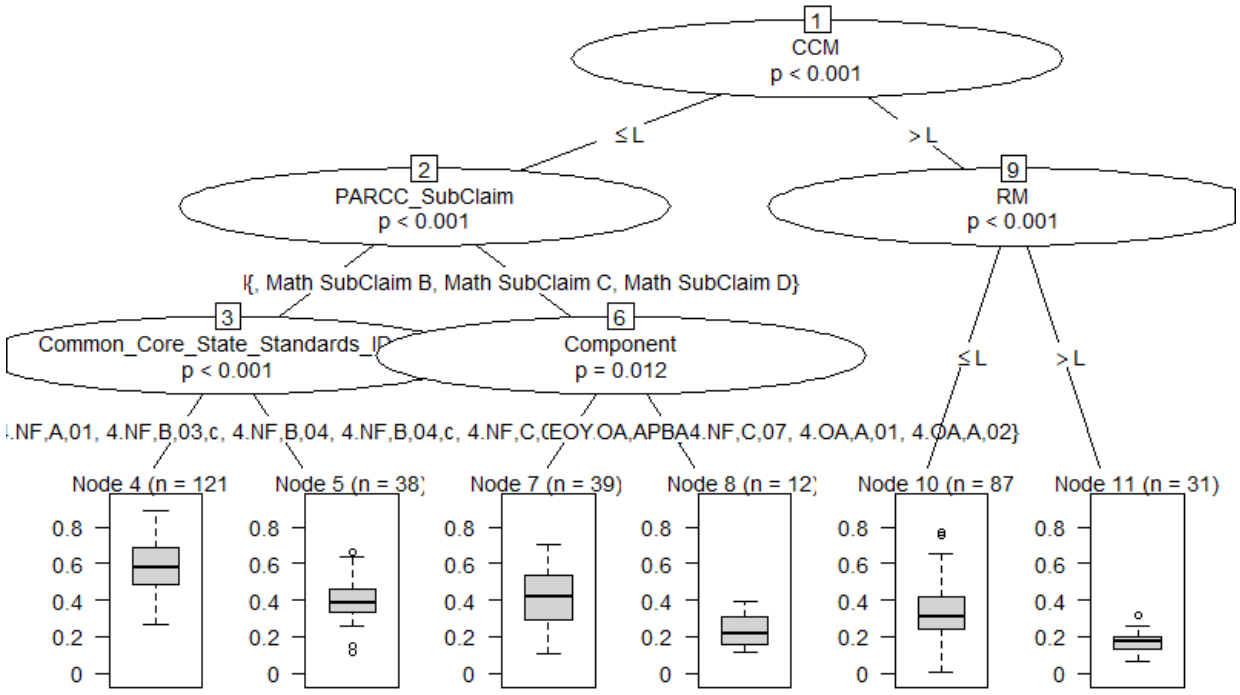


Figure F.2. Conditional Tree for Predicting Grade 4 Mathematics P-Values from Cognitive Complexity Source Codes and Metadata.

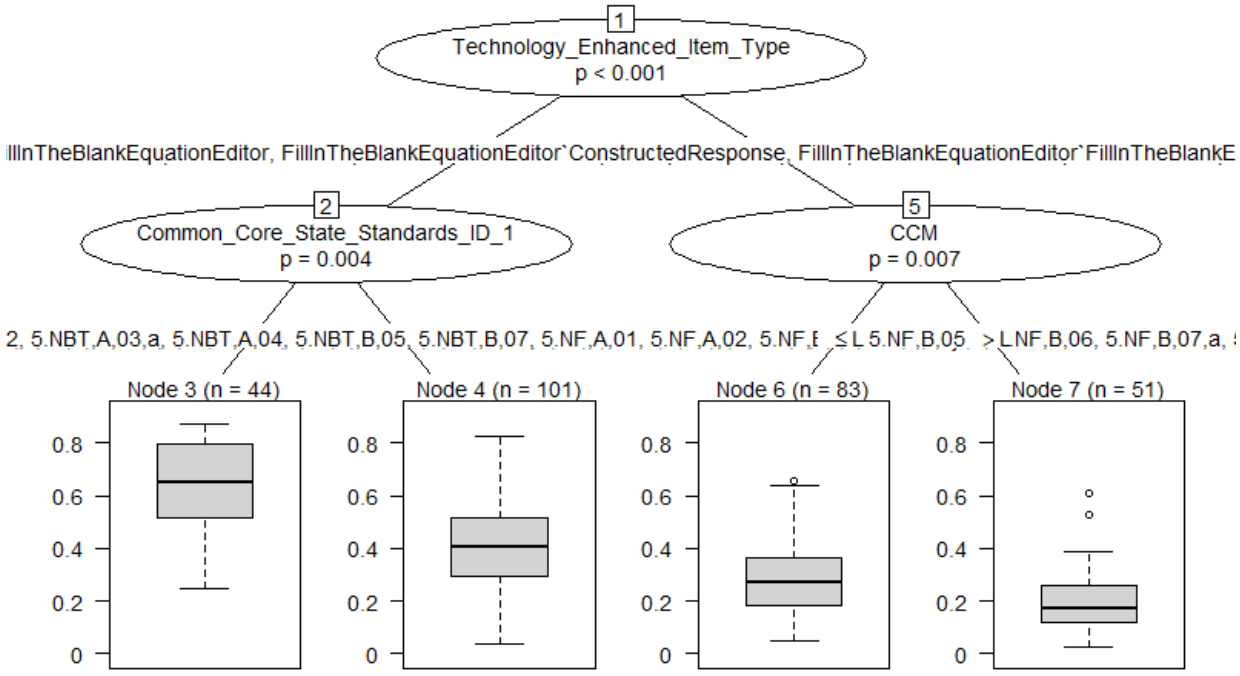


Figure F.3. Conditional Tree for Predicting Grade 5 Mathematics P-Values from Cognitive Complexity Source Codes and Metadata.

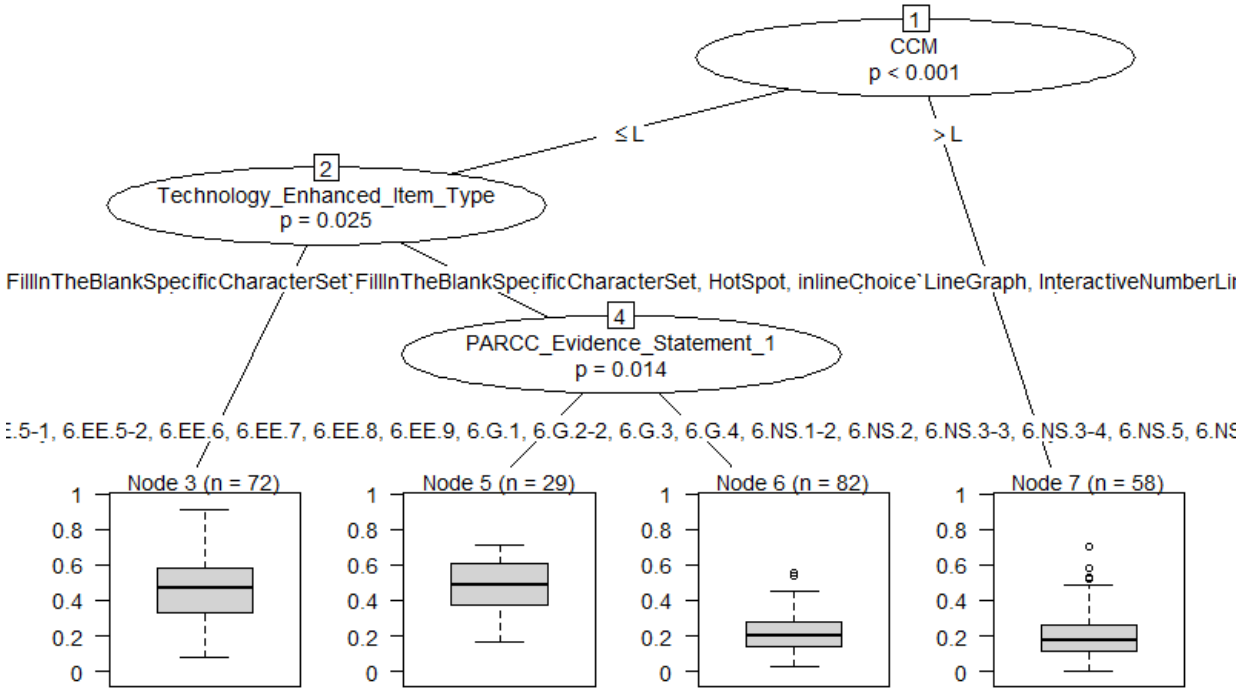


Figure F.4. Conditional Tree for Predicting Grade 6 Mathematics P-Values from Cognitive Complexity Source Codes and Metadata.

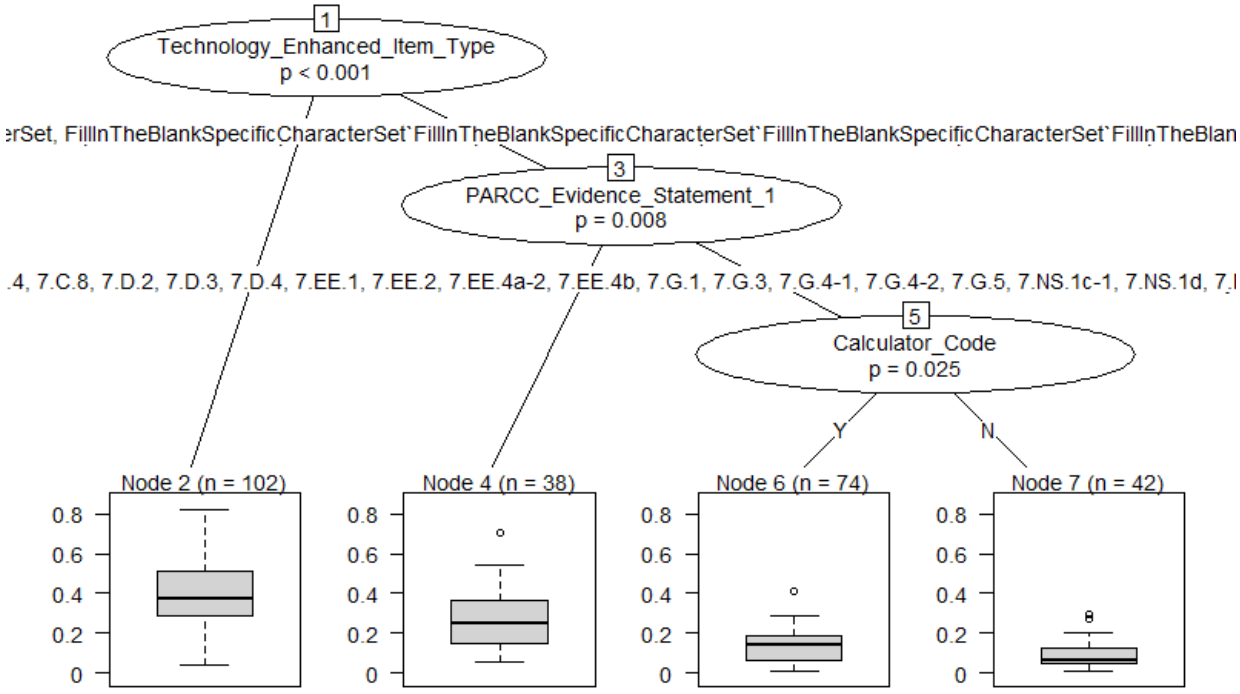


Figure F.5. Conditional Tree for Predicting Grade 7 Mathematics P-Values from Cognitive Complexity Source Codes and Metadata.

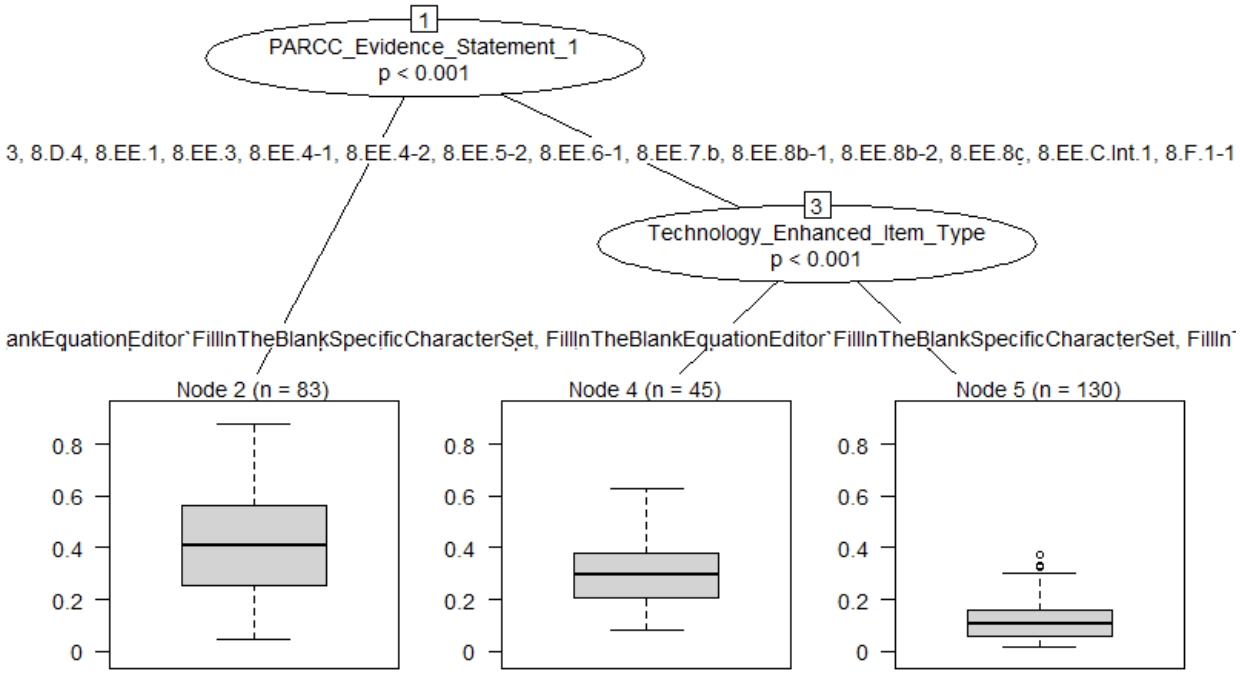


Figure F.6. Conditional Tree for Predicting Grade 8 Mathematics P-Values from Cognitive Complexity Source Codes and Metadata.

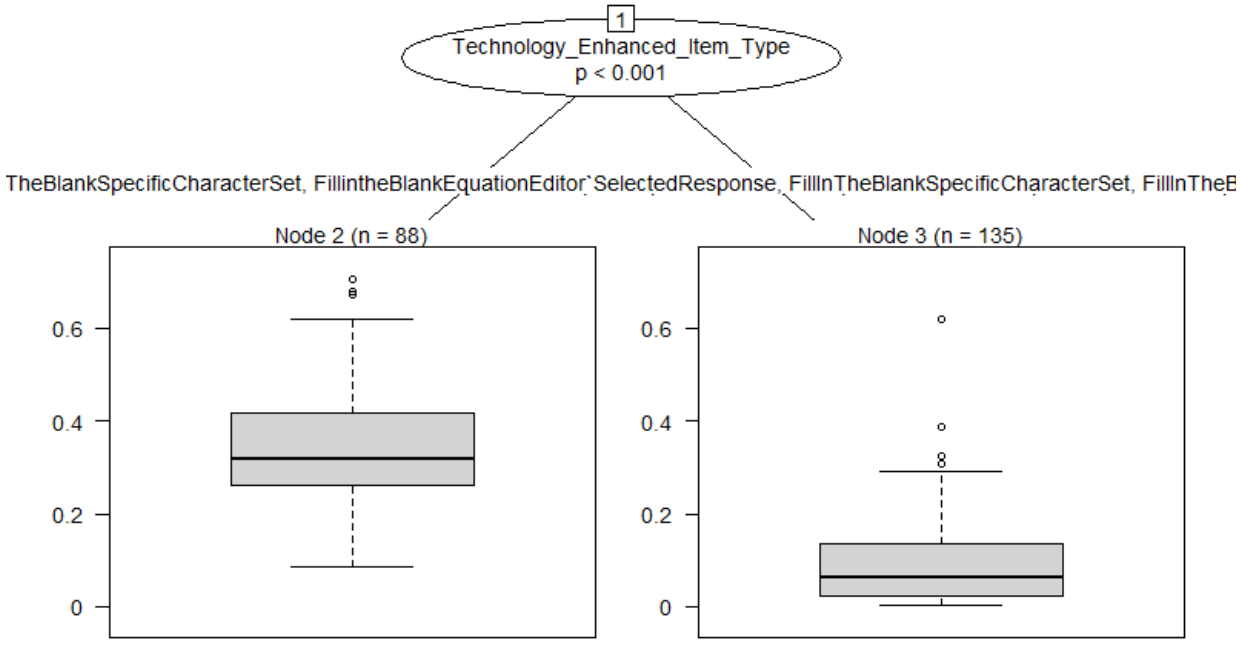


Figure F.7. Conditional Tree for Predicting Algebra I P-Values from Cognitive Complexity Source Codes and Metadata.

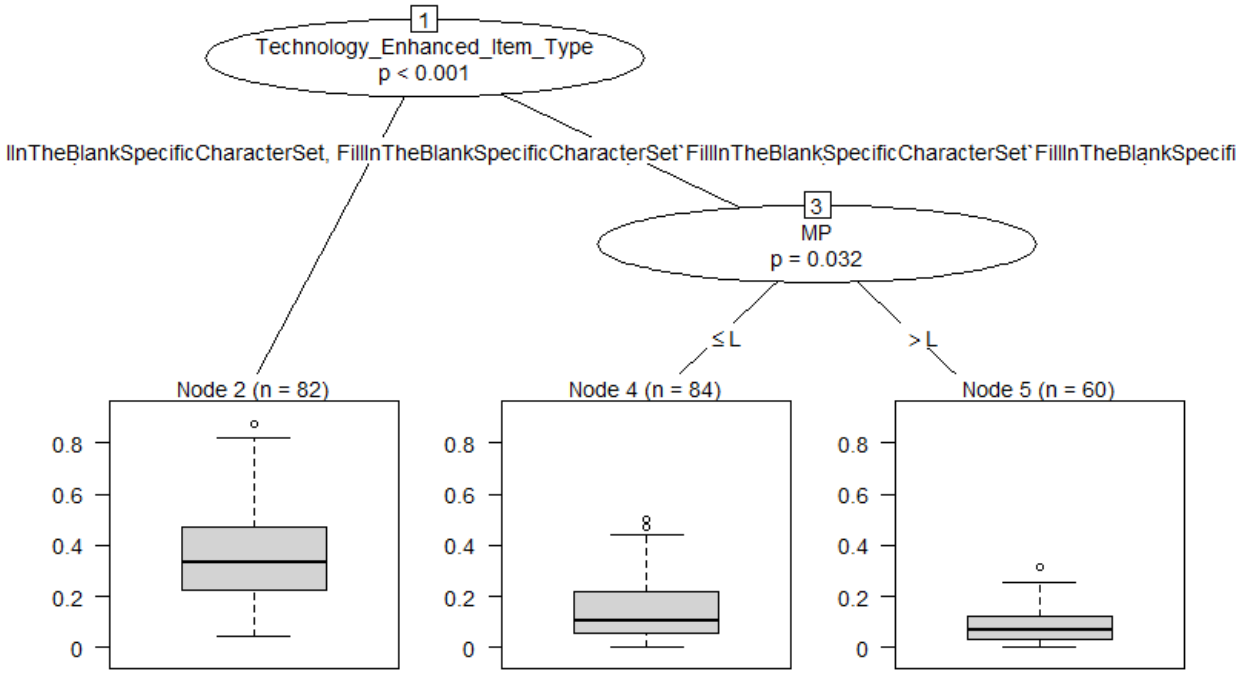


Figure F.8. Conditional Tree for Predicting Geometry P-Values from Cognitive Complexity Source Codes and Metadata.

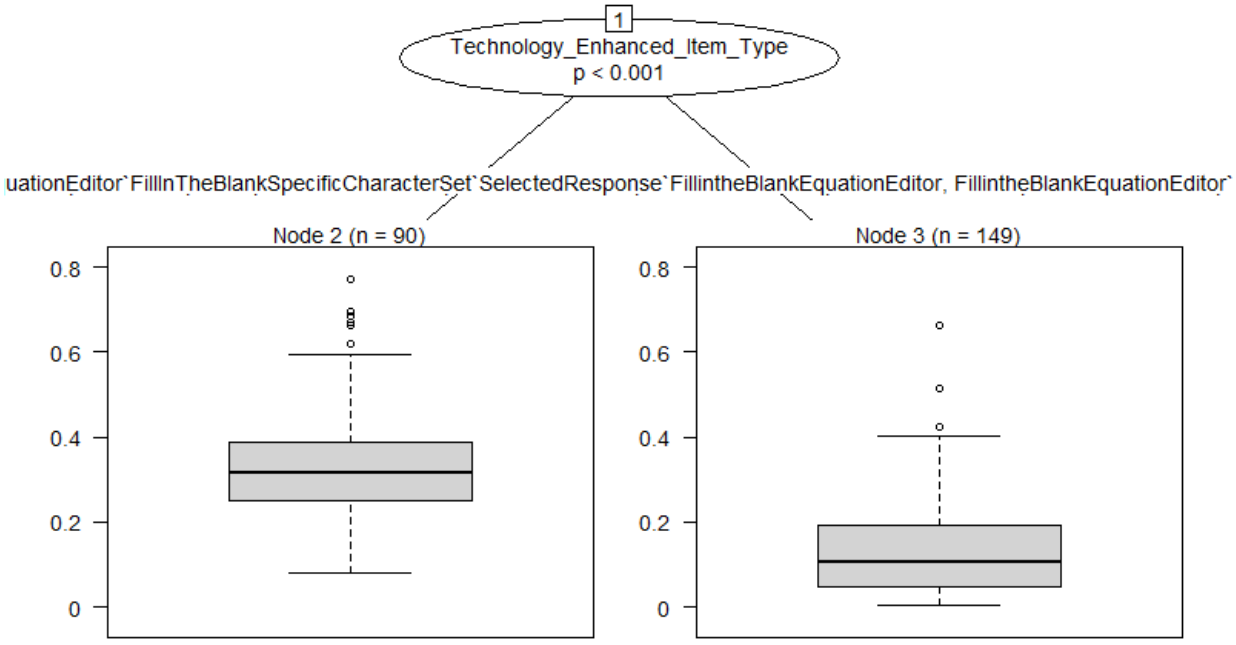


Figure F.9. Conditional Tree for Predicting Algebra II P-Values from Cognitive Complexity Source Codes and Metadata.

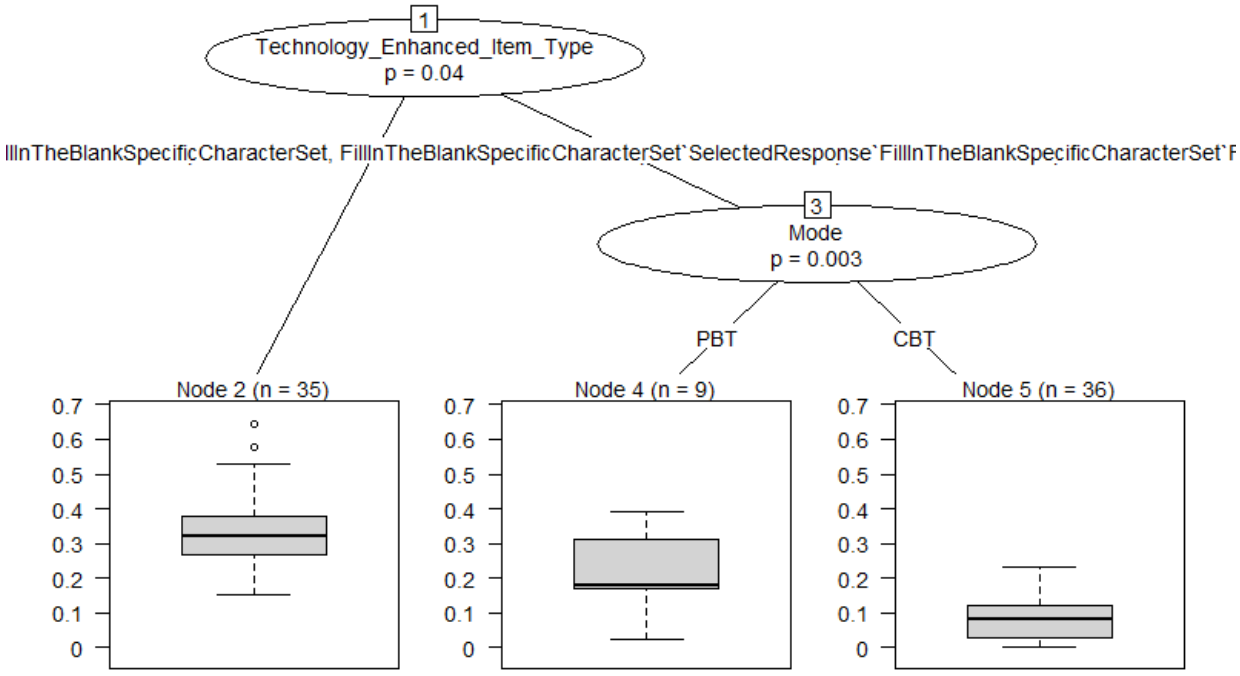


Figure F.10. Conditional Tree for Predicting Integrated Mathematics 1 P-Values from Cognitive Complexity Source Codes and Metadata.

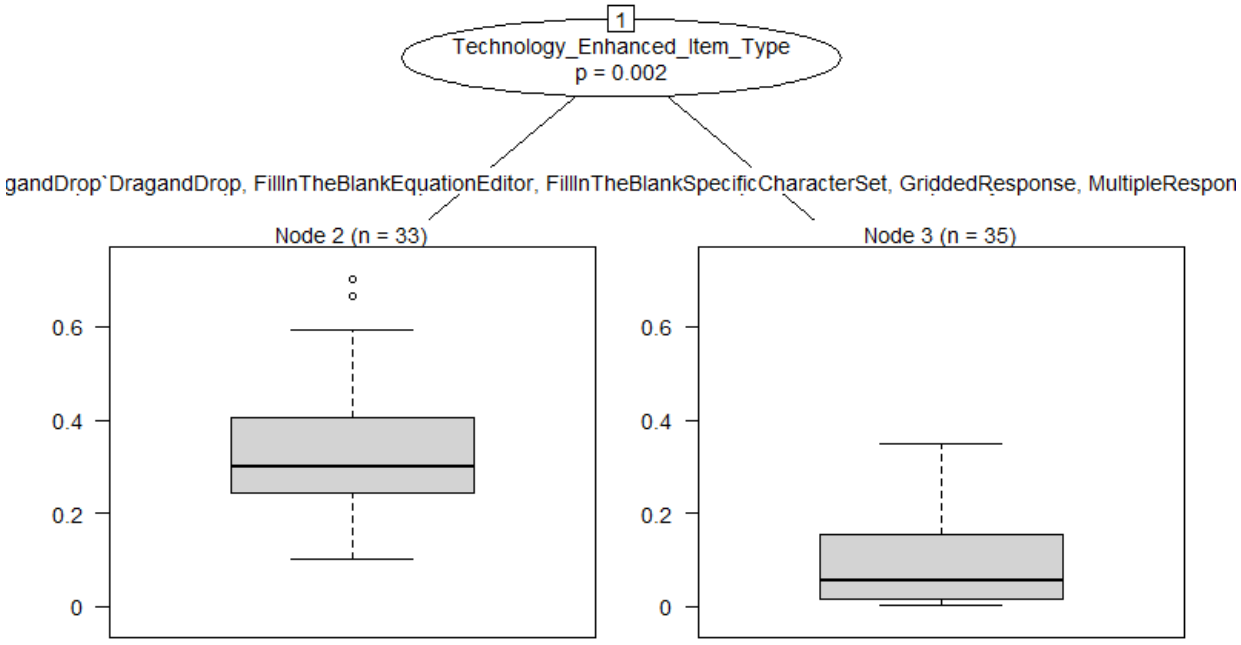


Figure F.11. Conditional Tree for Predicting Integrated Mathematics 2 P-Values from Cognitive Complexity Source Codes and Metadata.

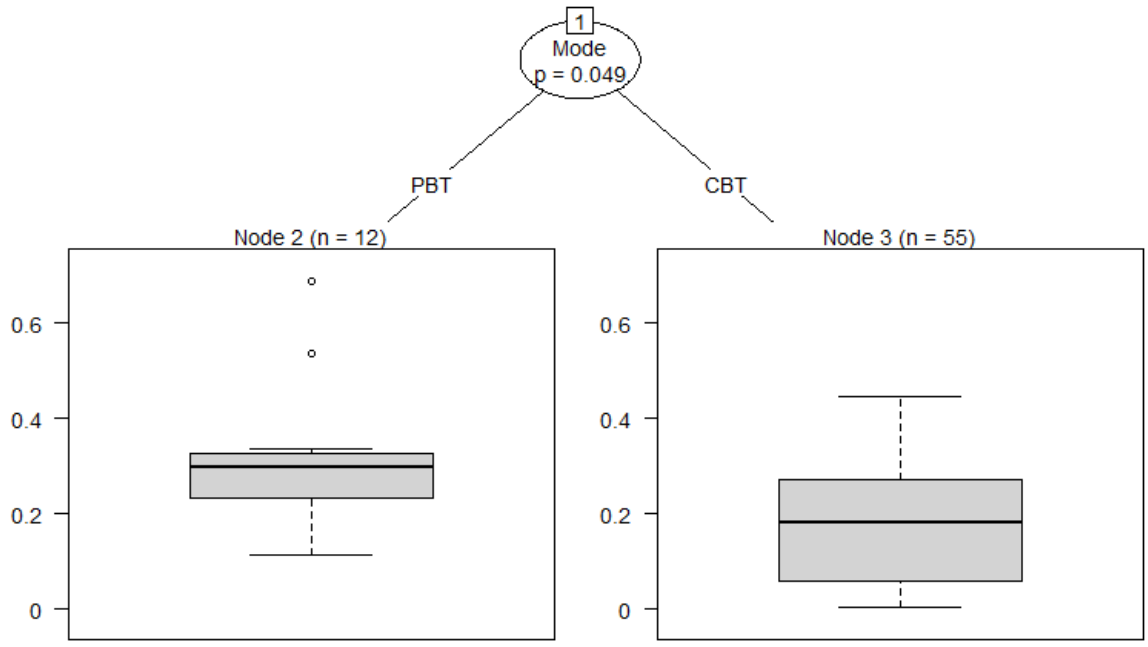


Figure F.12. Conditional Tree for Predicting Integrated Mathematics 3 P-Values from Cognitive Complexity Source Codes and Metadata.

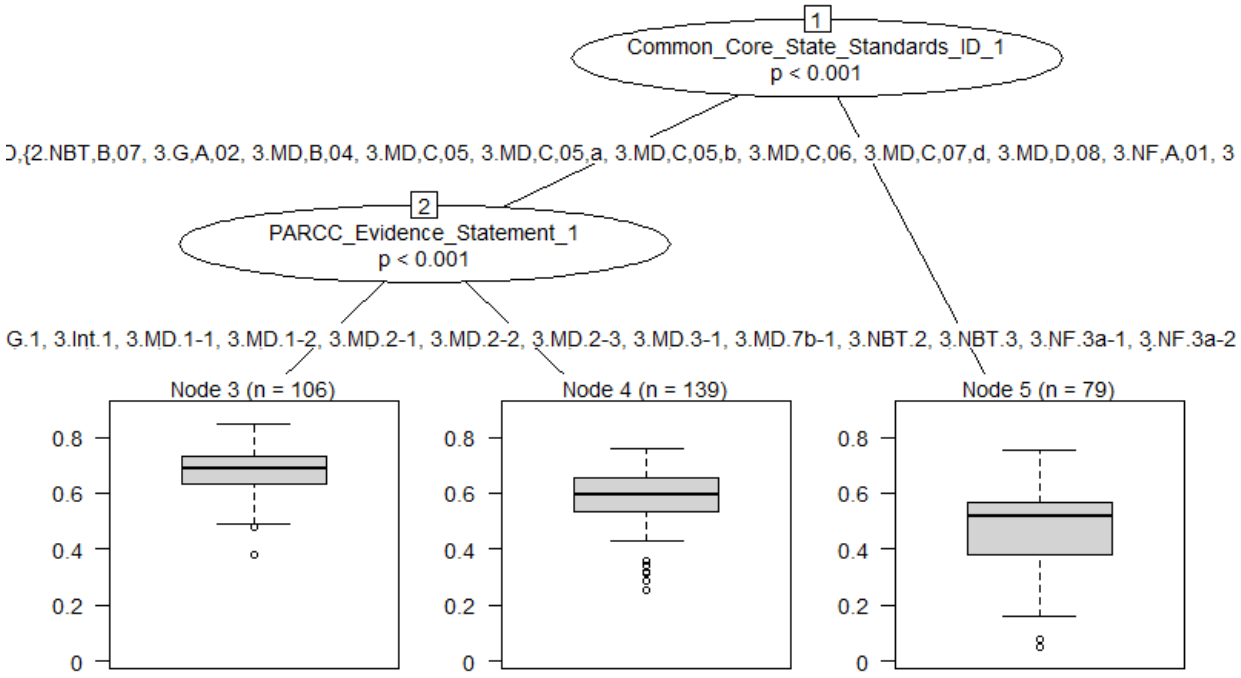


Figure F.13. Conditional Tree for Predicting Grade 3 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

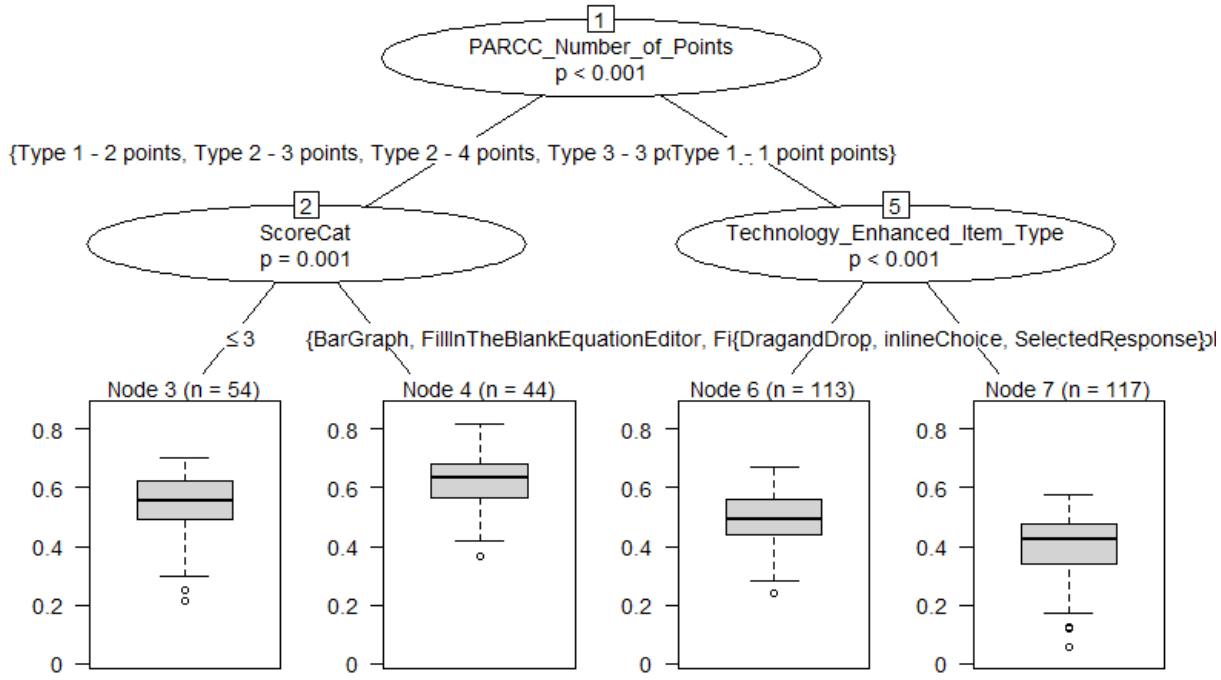


Figure F.14. Conditional Tree for Predicting Grade 4 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

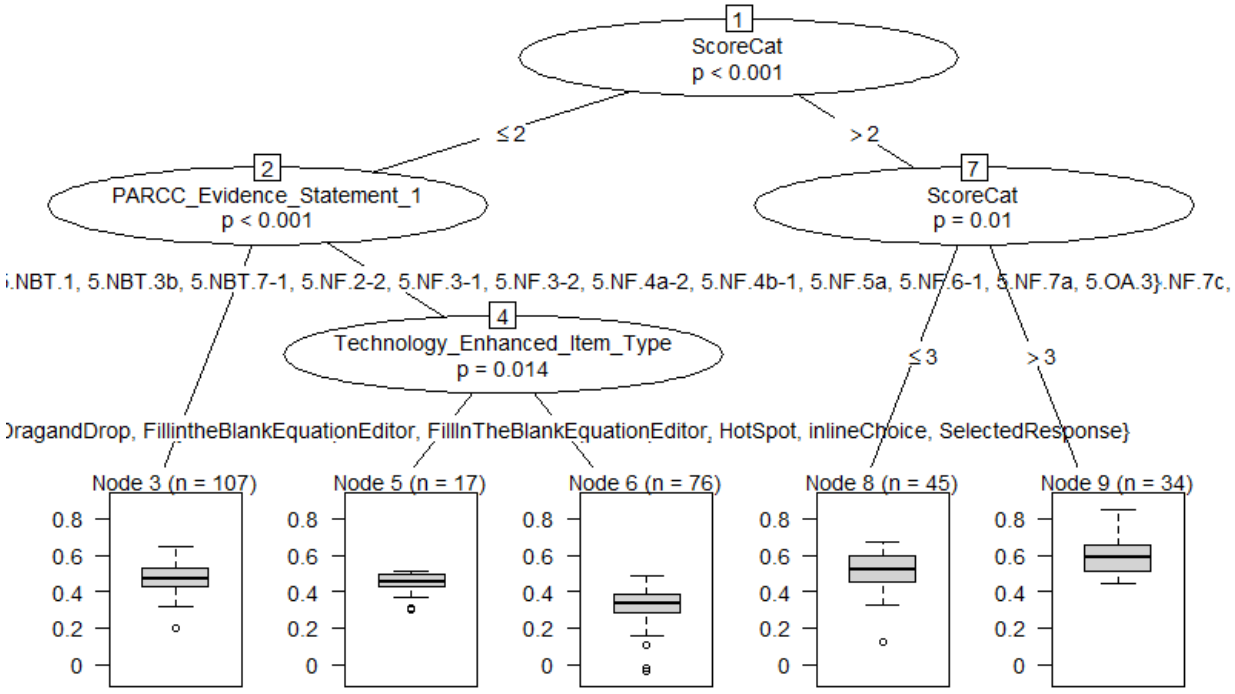


Figure F.15. Conditional Tree for Predicting Grade 5 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

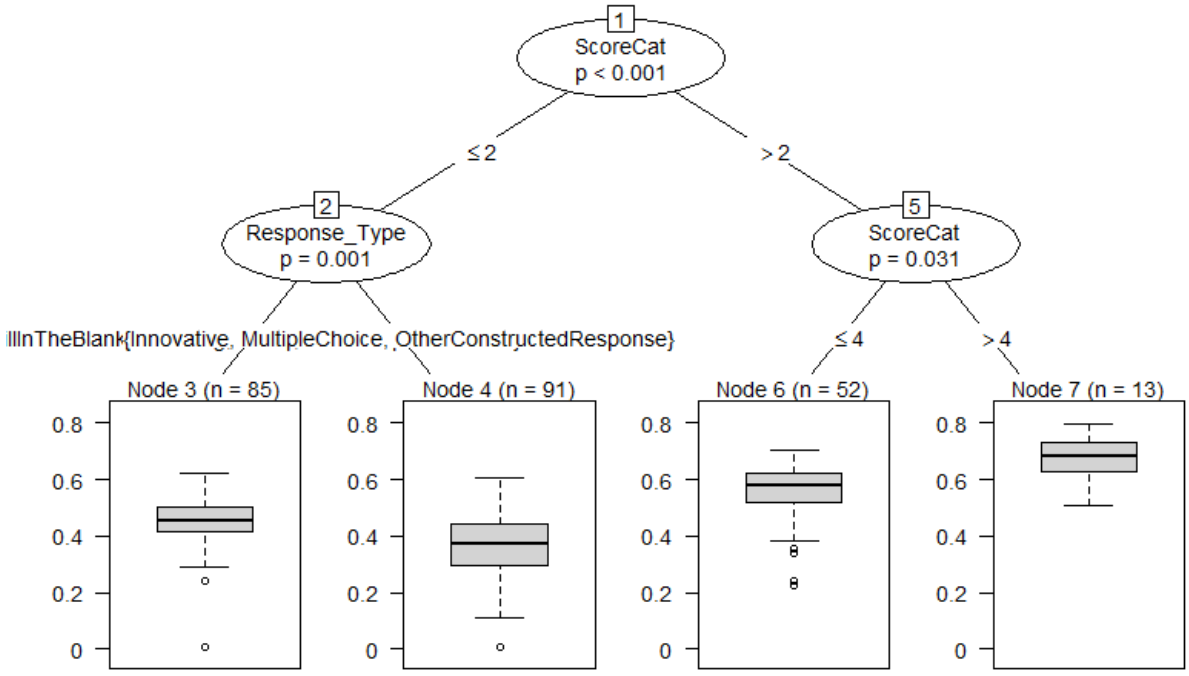


Figure F.16. Conditional Tree for Predicting Grade 6 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

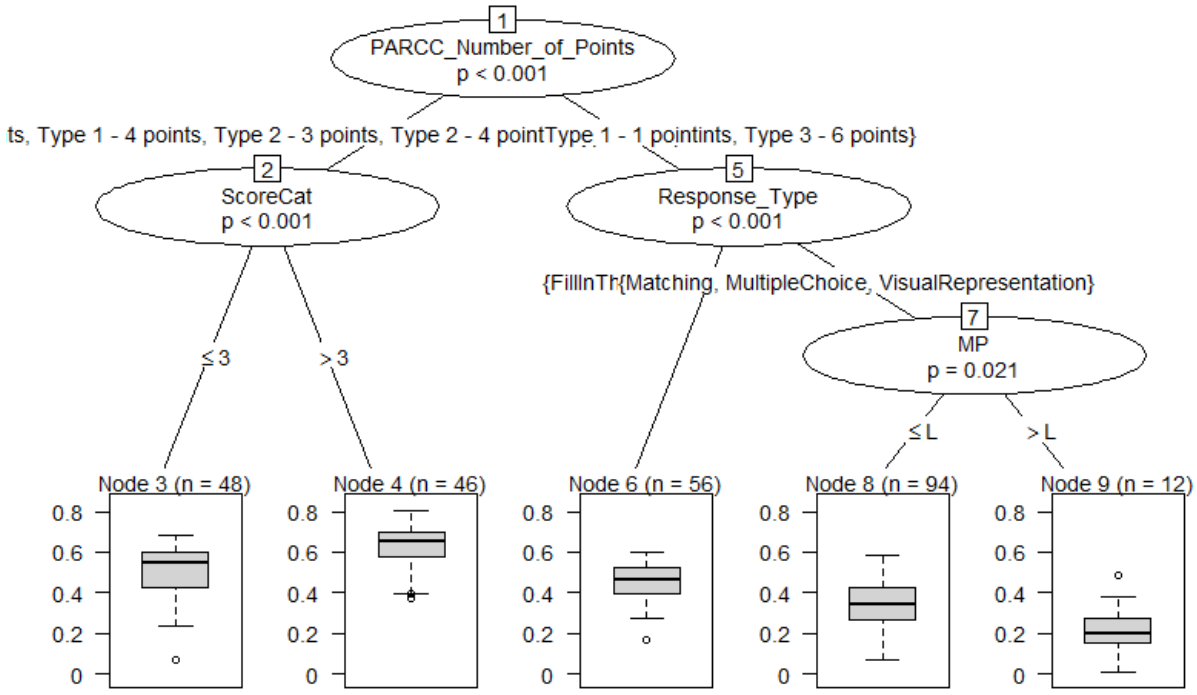


Figure F.17. Conditional Tree for Predicting Grade 7 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

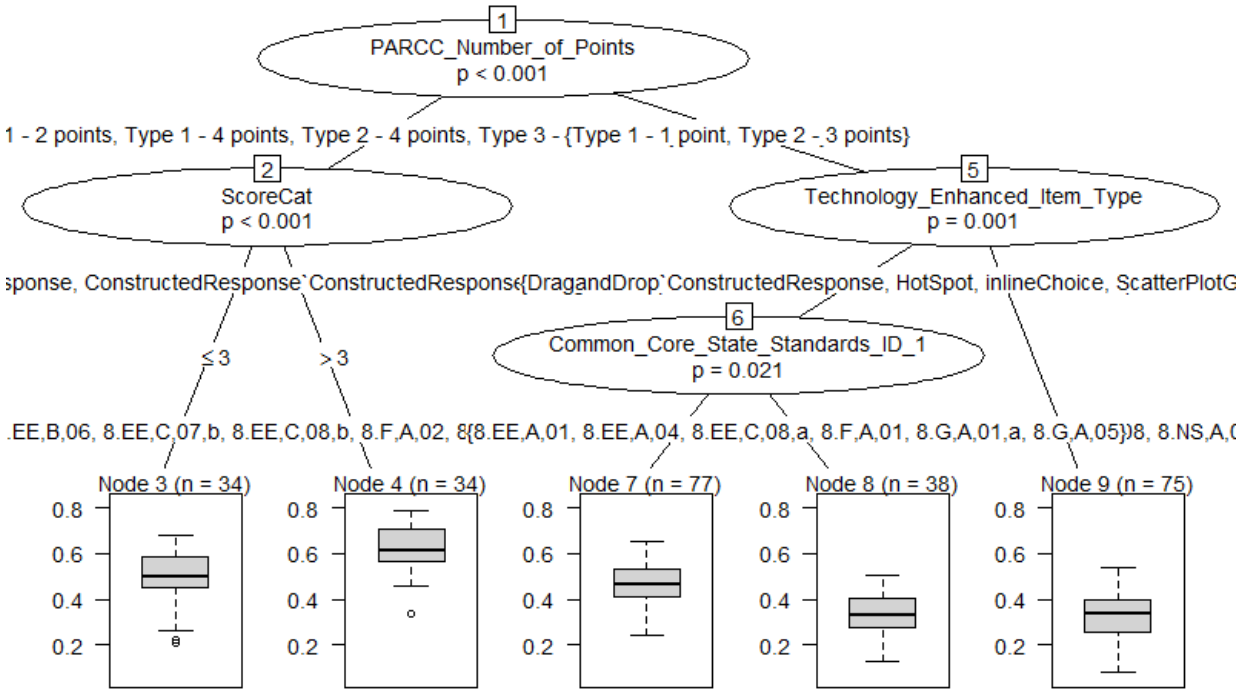


Figure F.18. Conditional Tree for Predicting Grade 8 Mathematics Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

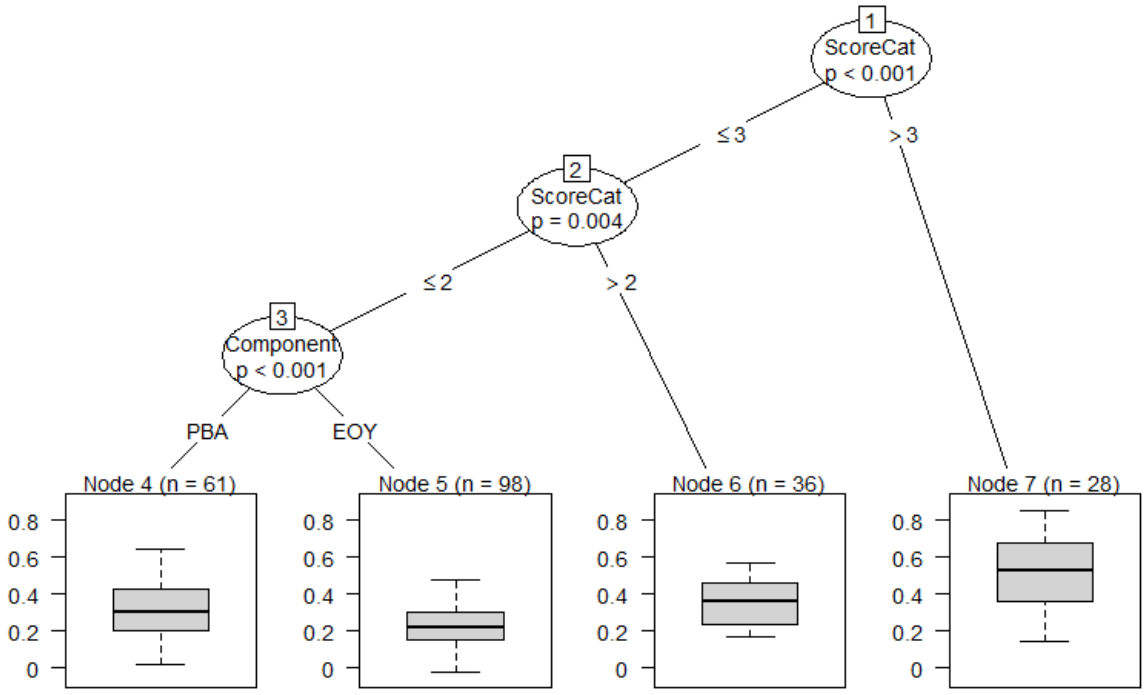


Figure F.19. Conditional Tree for Predicting Algebra I Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

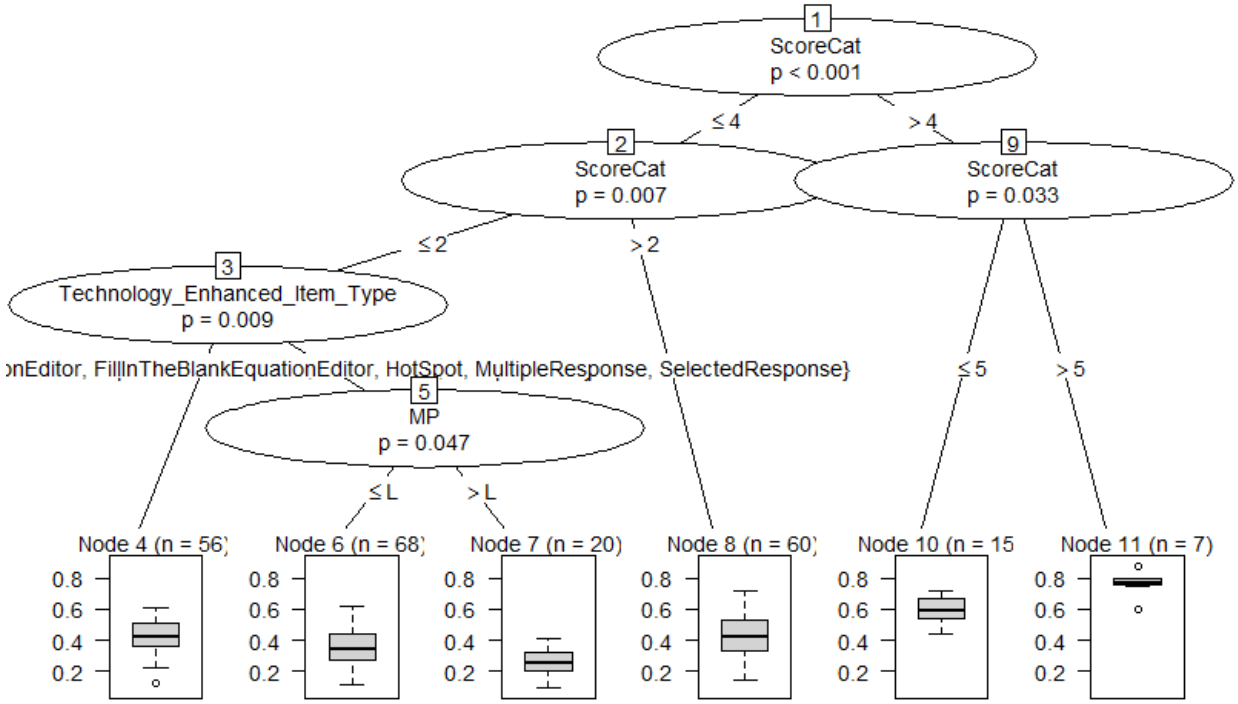


Figure F.20. Conditional Tree for Predicting Geometry Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

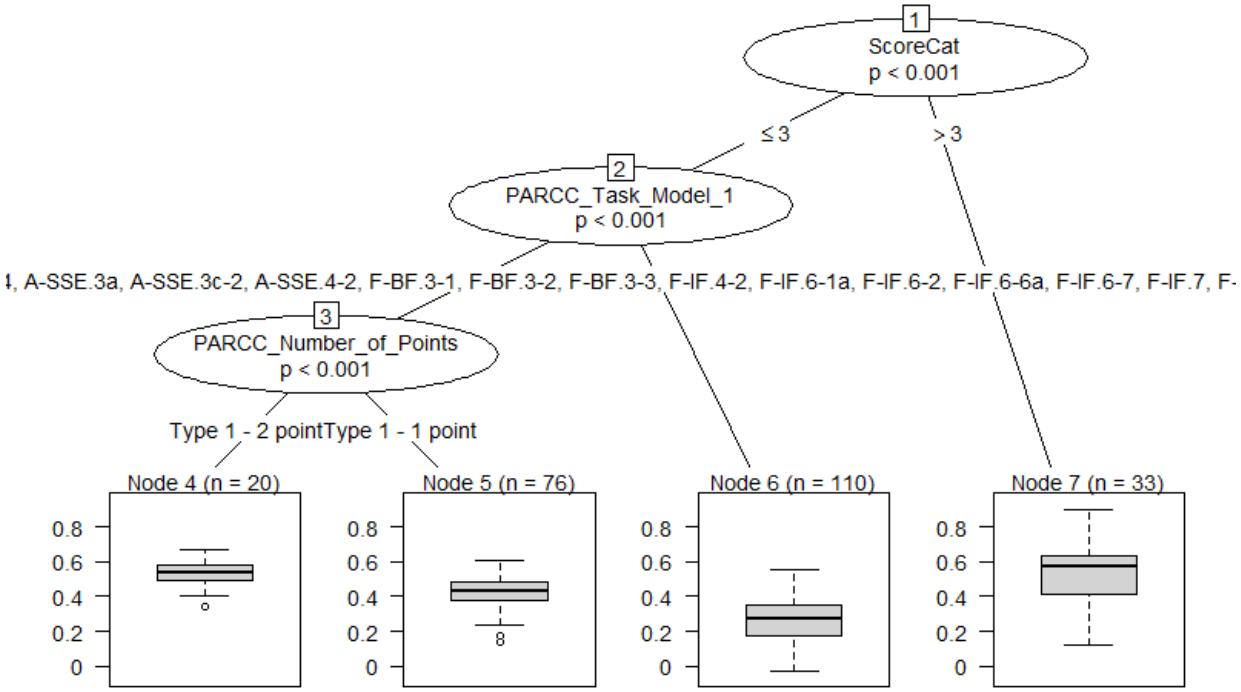


Figure F.21. Conditional Tree for Predicting Algebra II Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

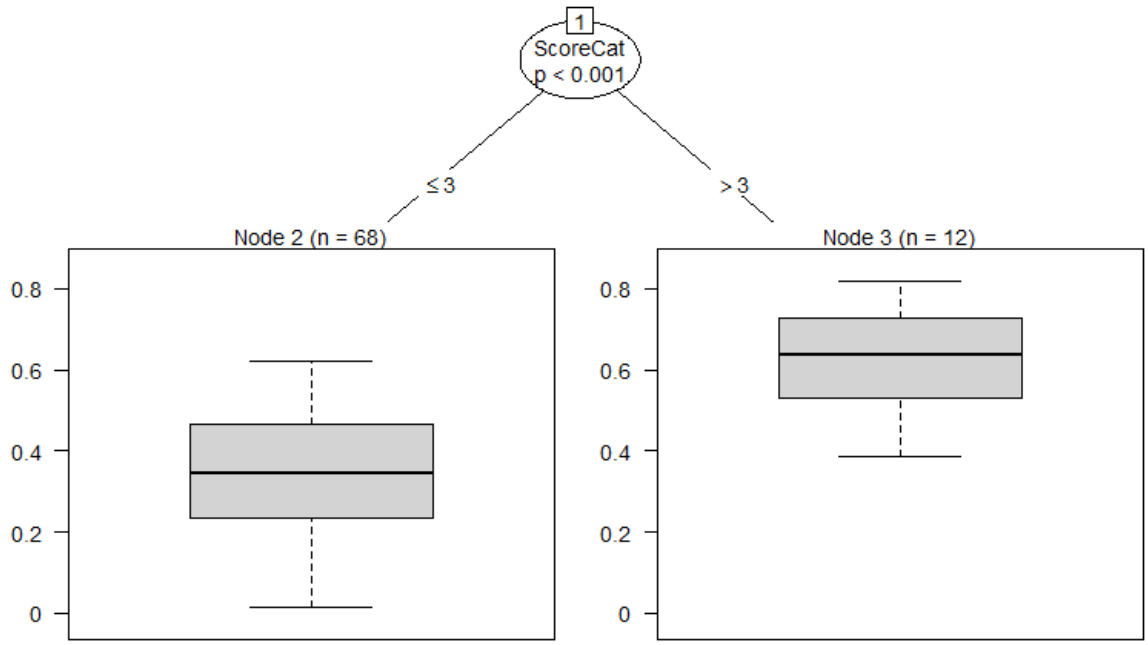


Figure F.22. Conditional Tree for Predicting Integrated Mathematics 1 Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

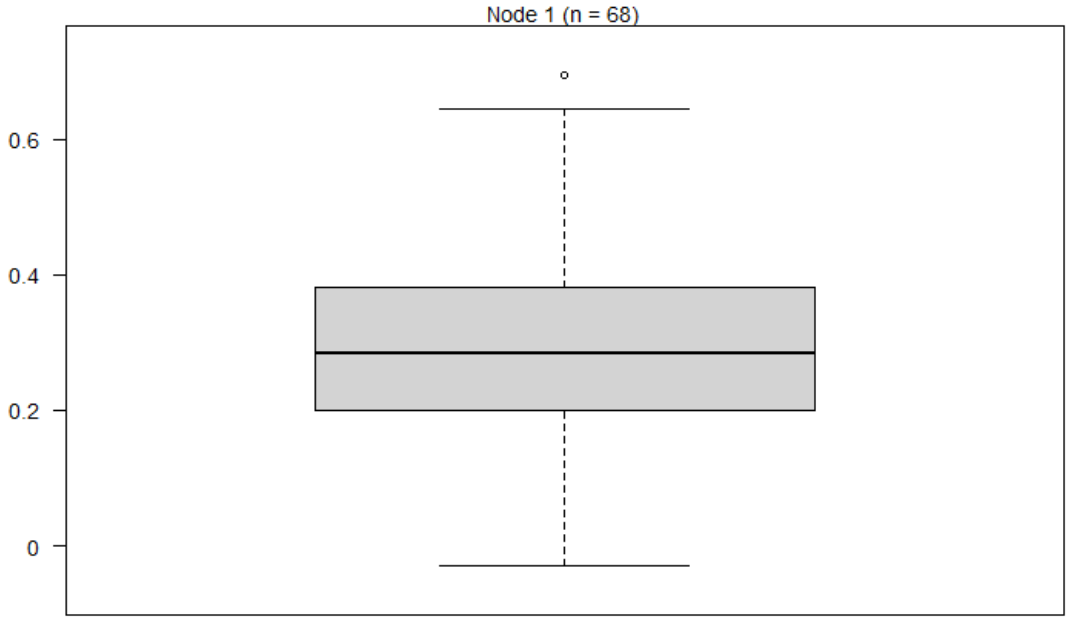


Figure F.23. Conditional Tree for Predicting Integrated Mathematics 2 Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

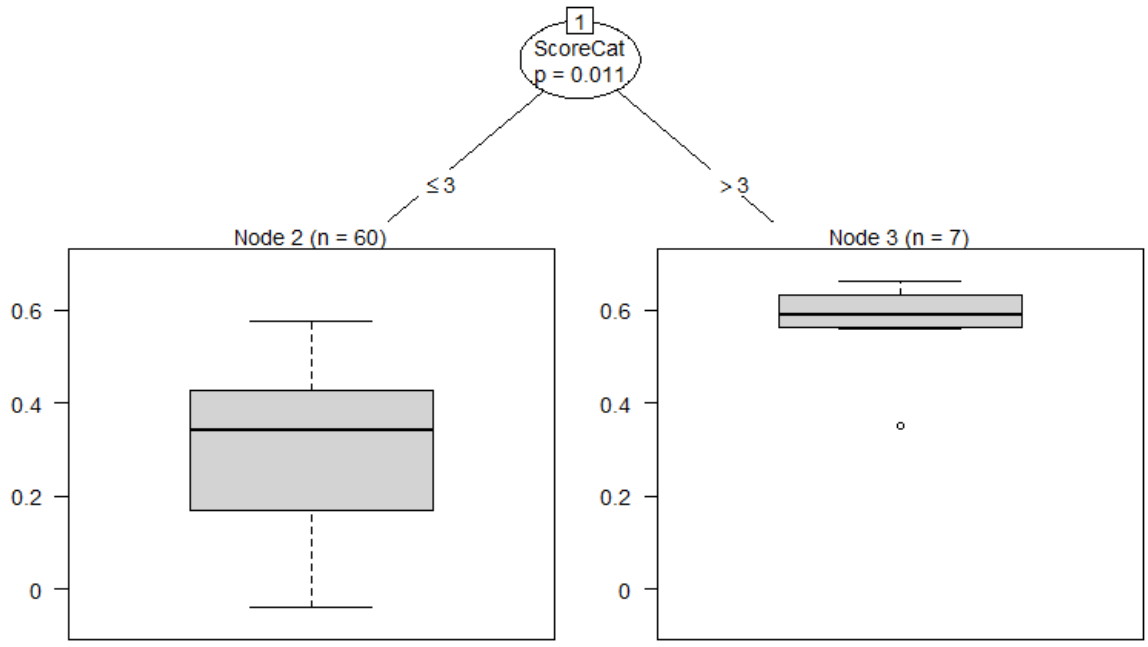


Figure F.24. Conditional Tree for Predicting Integrated Mathematics 3 Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

Appendix G: ELA/L Conditional Trees using Cognitive Complexity Source Codes and Metadata as Predictors

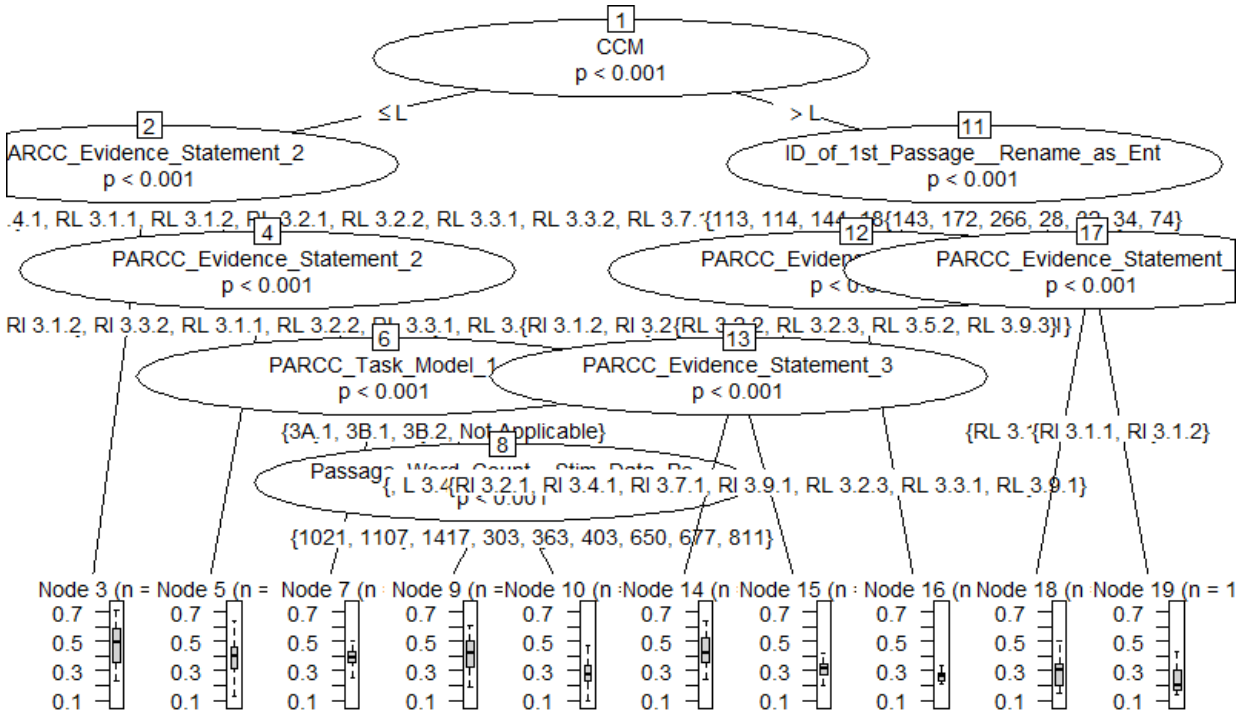


Figure G.1. Conditional Tree for Predicting Grade 3 ELA/L P-Values from Cognitive Complexity Source Codes and Metadata.

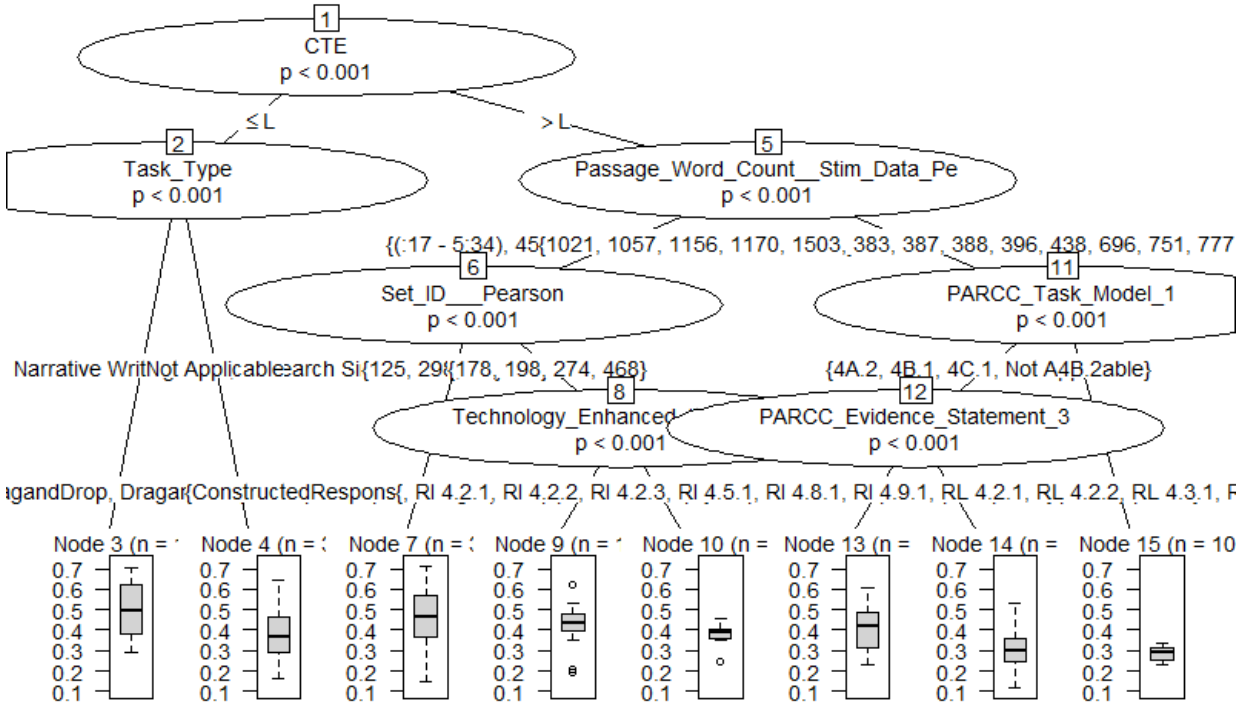


Figure G.2. Conditional Tree for Predicting Grade 4 ELA/L P-Values from Cognitive Complexity Source Codes and Metadata.

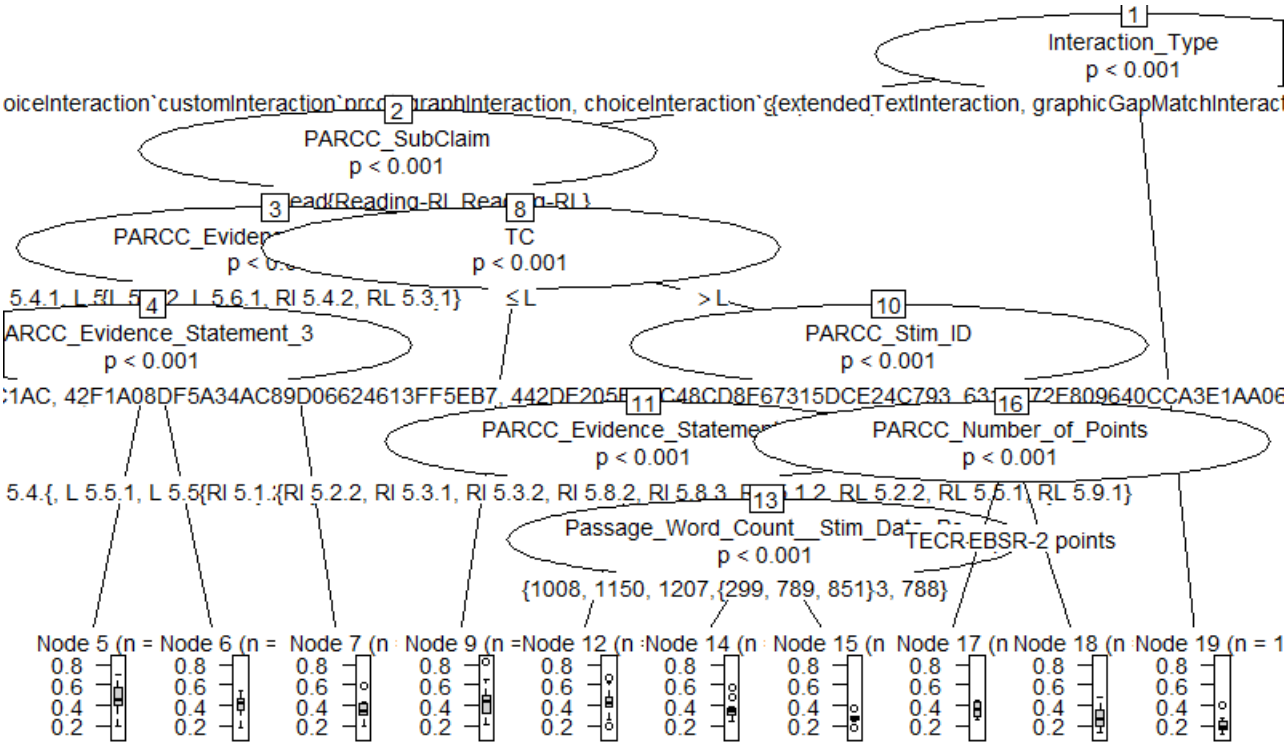


Figure G.3. Conditional Tree for Predicting Grade 5 ELA/L P-Values from Cognitive Complexity Source Codes and Metadata.

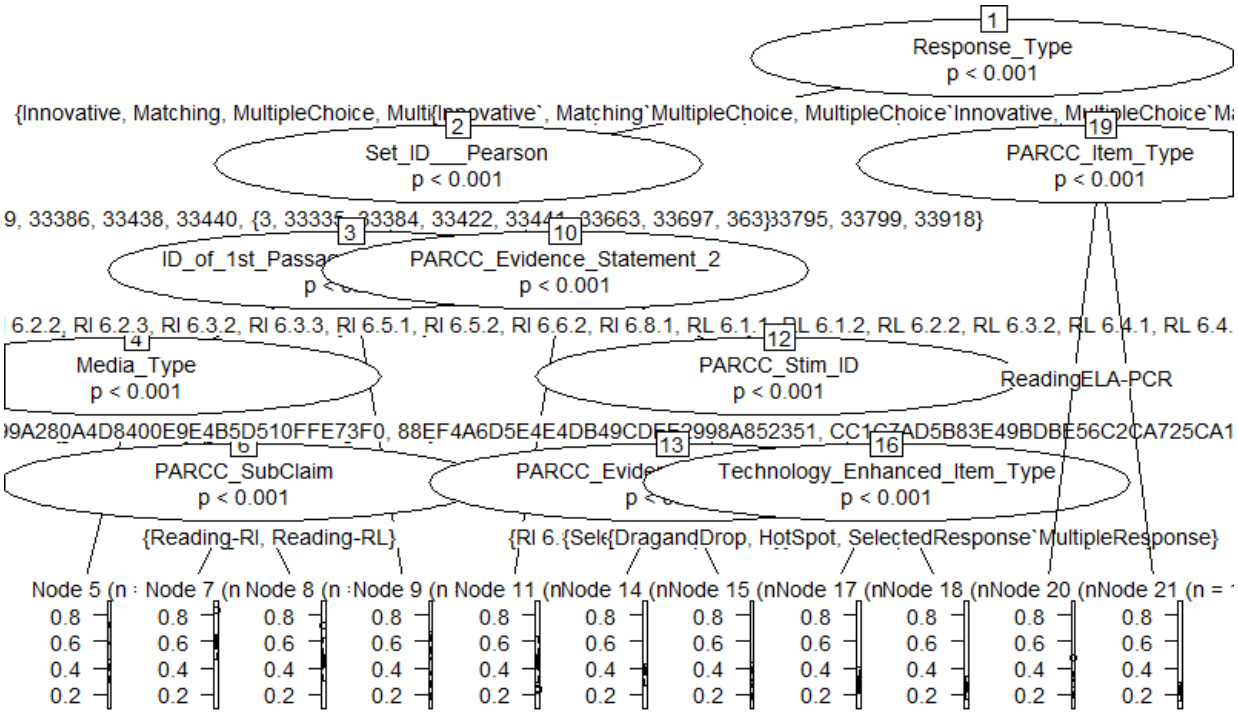


Figure G.4. Conditional Tree for Predicting Grade 6 ELA/L P-Values from Cognitive Complexity Source Codes and Metadata.

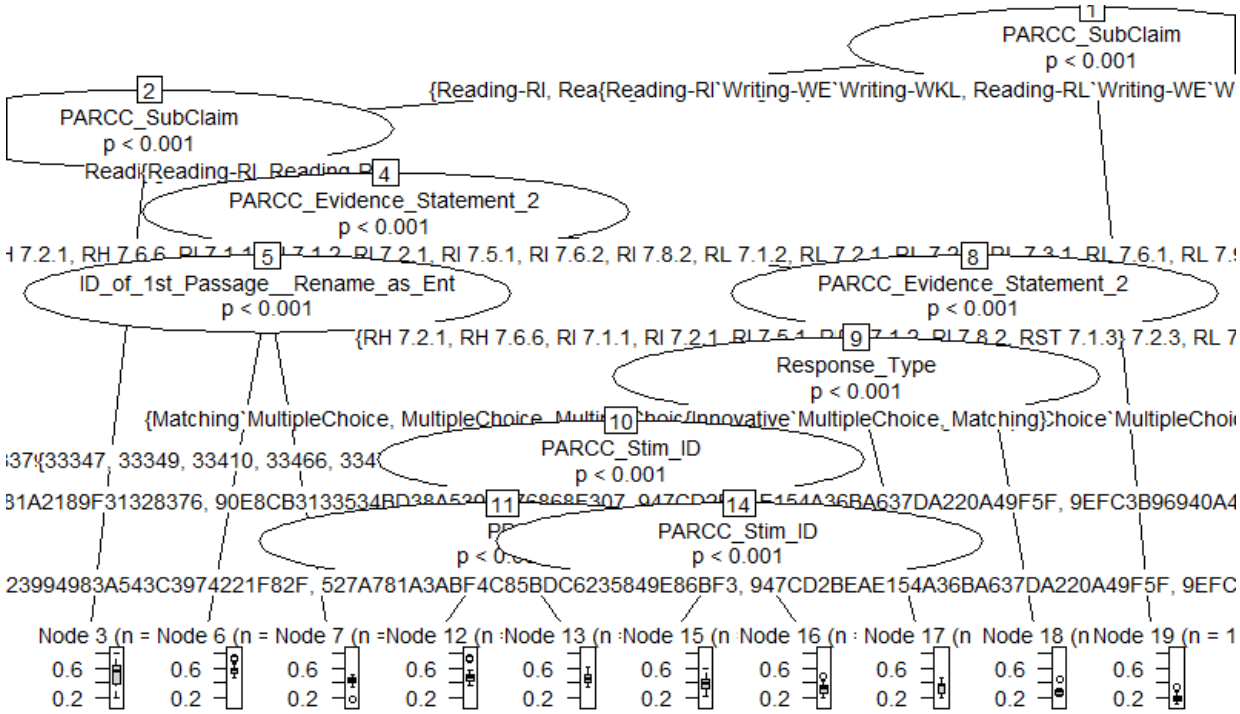


Figure G.5. Conditional Tree for Predicting Grade 7 ELA/L P-Values from Cognitive Complexity Source Codes and Metadata.

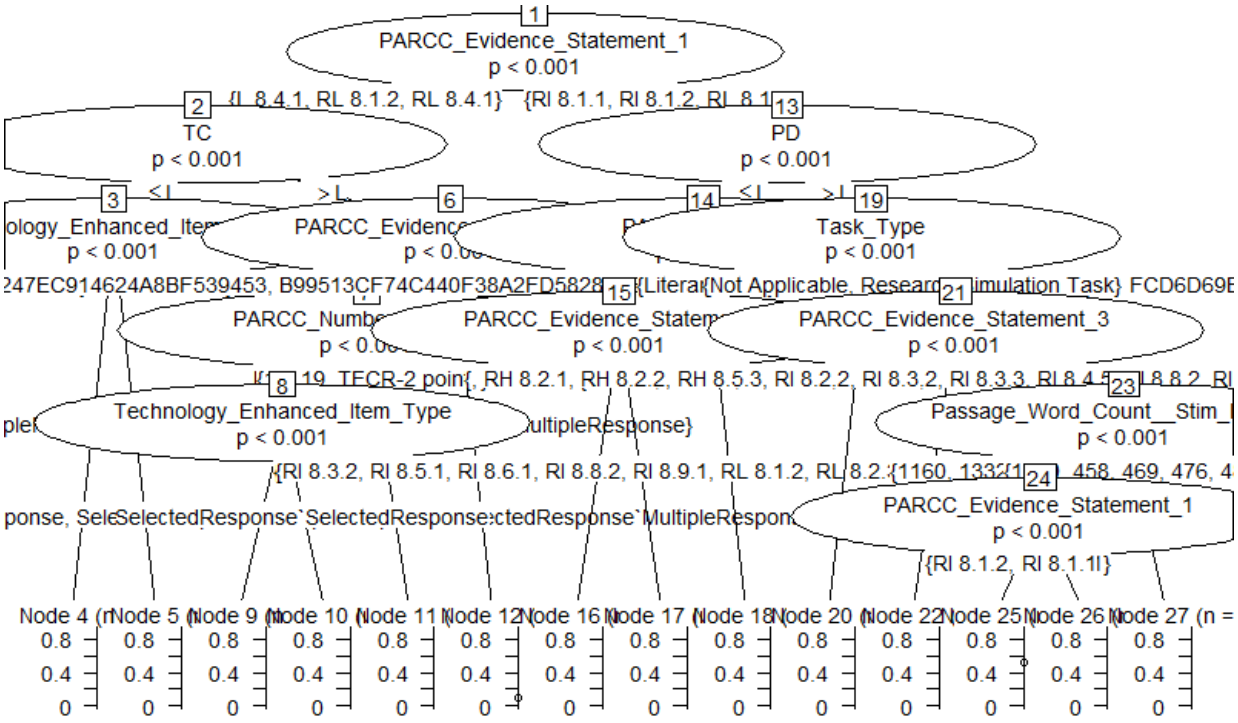


Figure G.6. Conditional Tree for Predicting Grade 8 ELA/L P-Values from Cognitive Complexity Source Codes and Metadata.

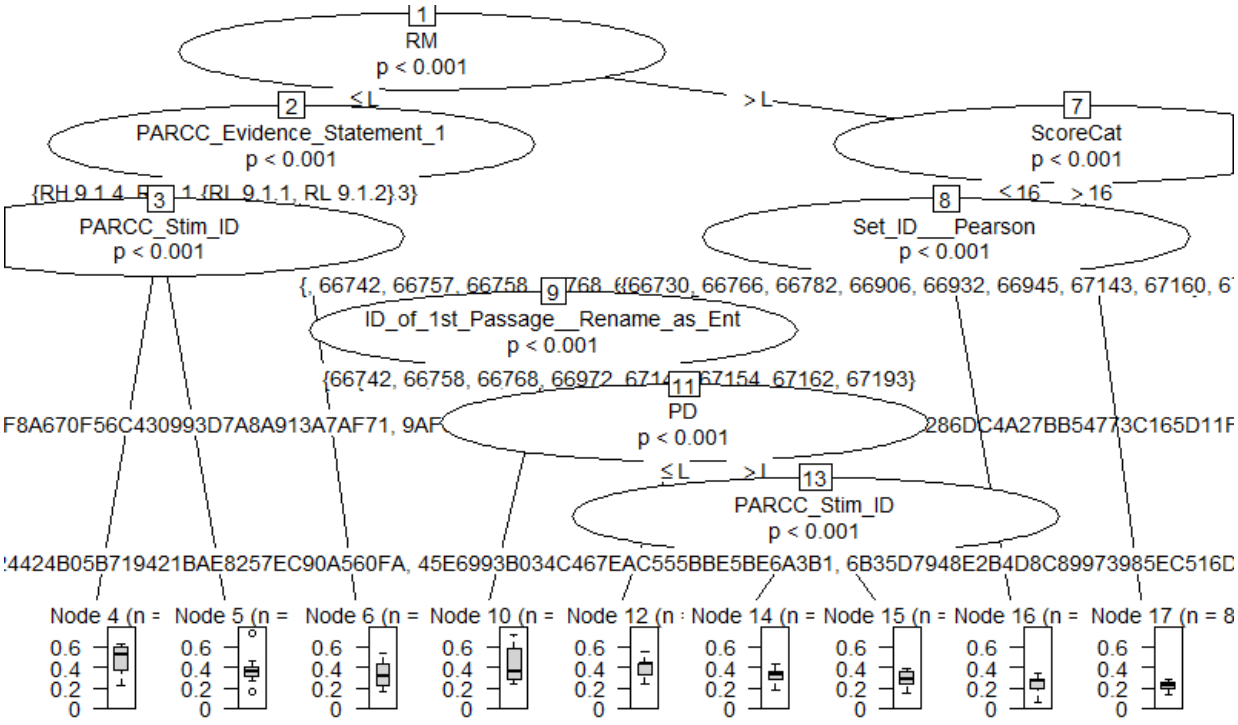


Figure G.7. Conditional Tree for Predicting Grade 9 ELA/LI P-Values from Cognitive Complexity Source Codes and Metadata.

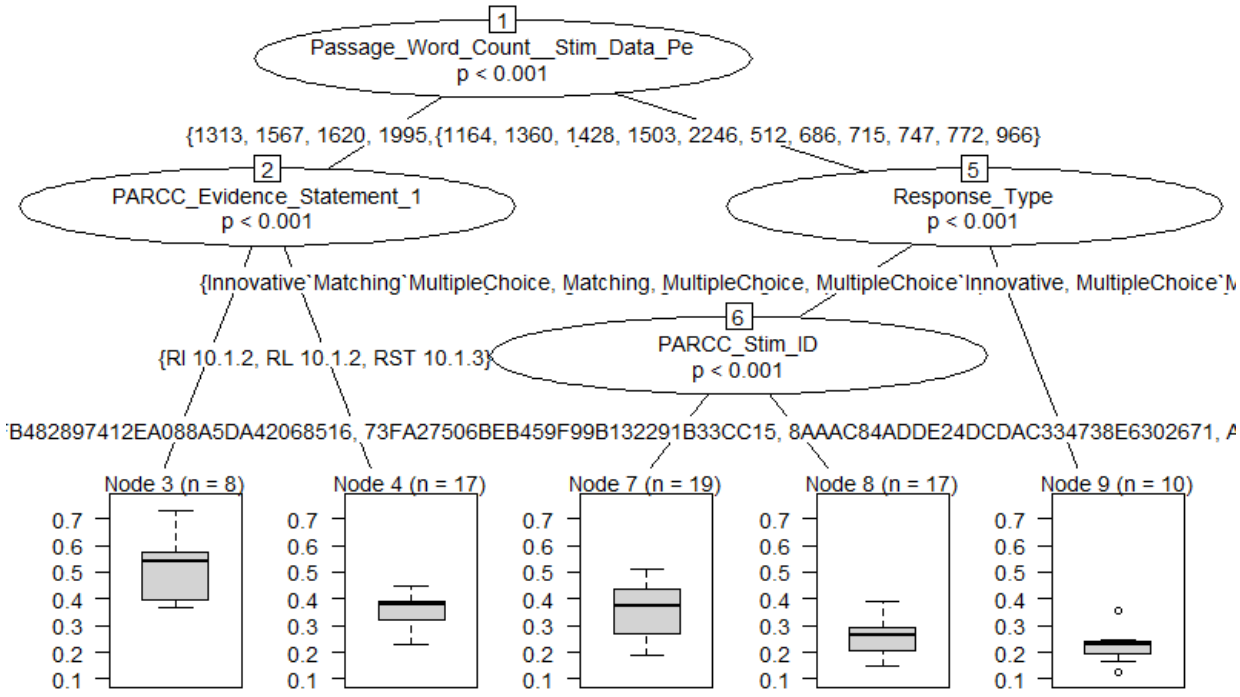


Figure G.8. Conditional Tree for Predicting Grade 10 ELA/L P-Values from Cognitive Complexity Source Codes and Metadata.

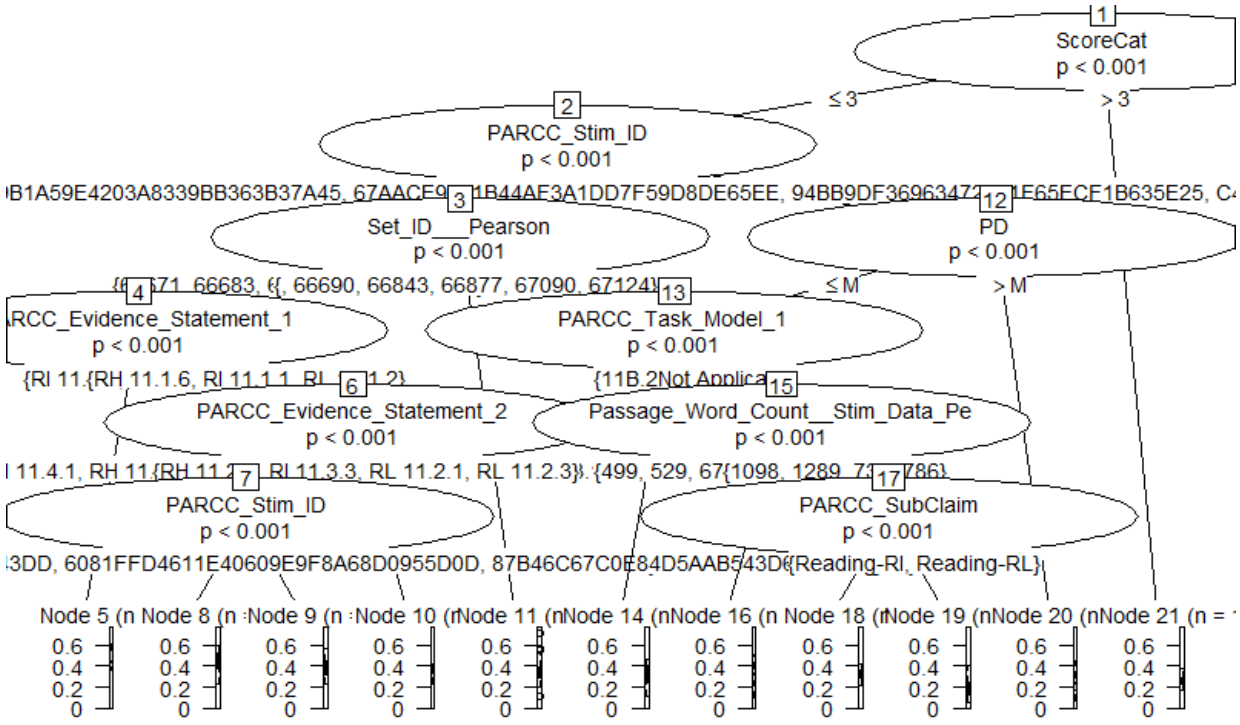


Figure G.9. Conditional Tree for Predicting Grade 11 ELA/L P-Values from Cognitive Complexity Source Codes and Metadata.

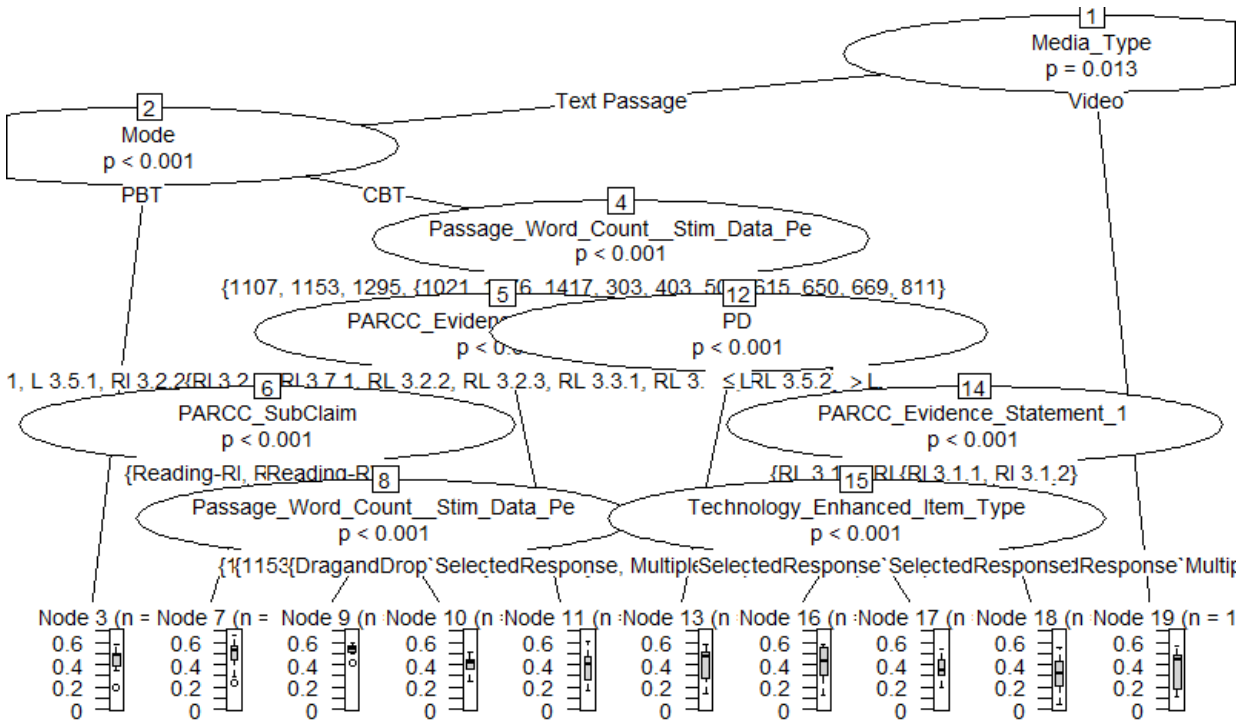


Figure G.10. Conditional Tree for Predicting Grade 3 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

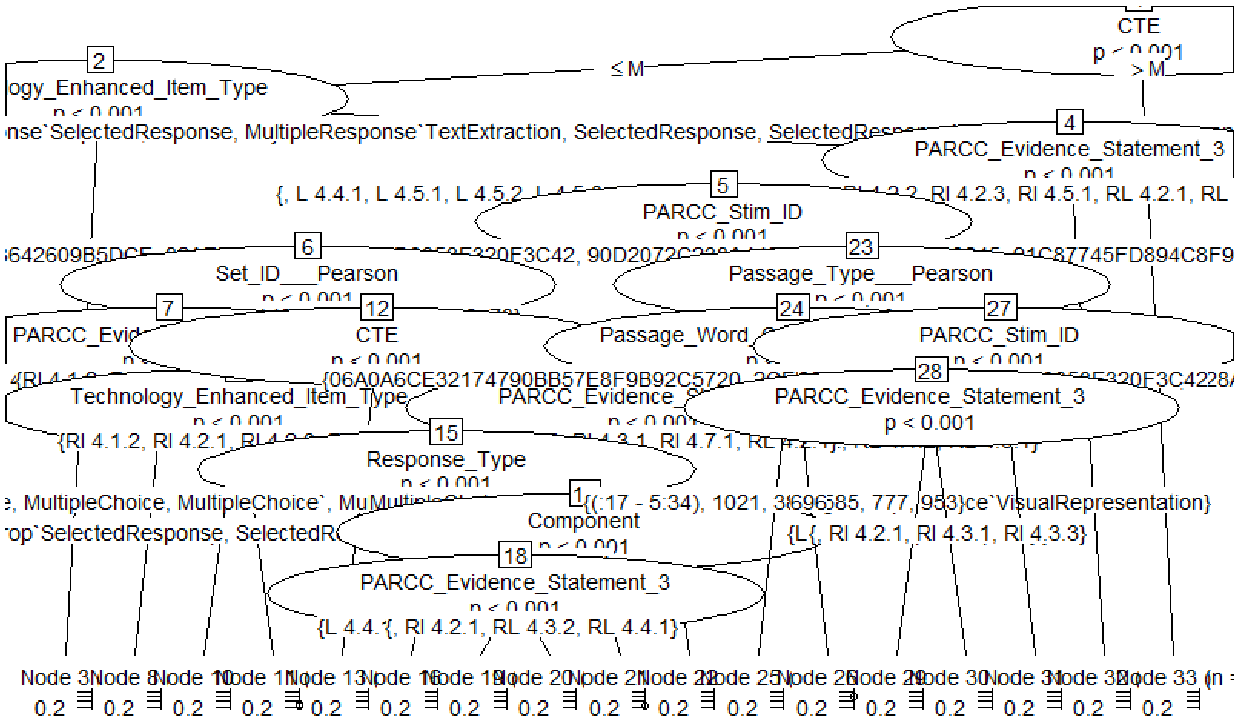


Figure G.11. Conditional Tree for Predicting Grade 4 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

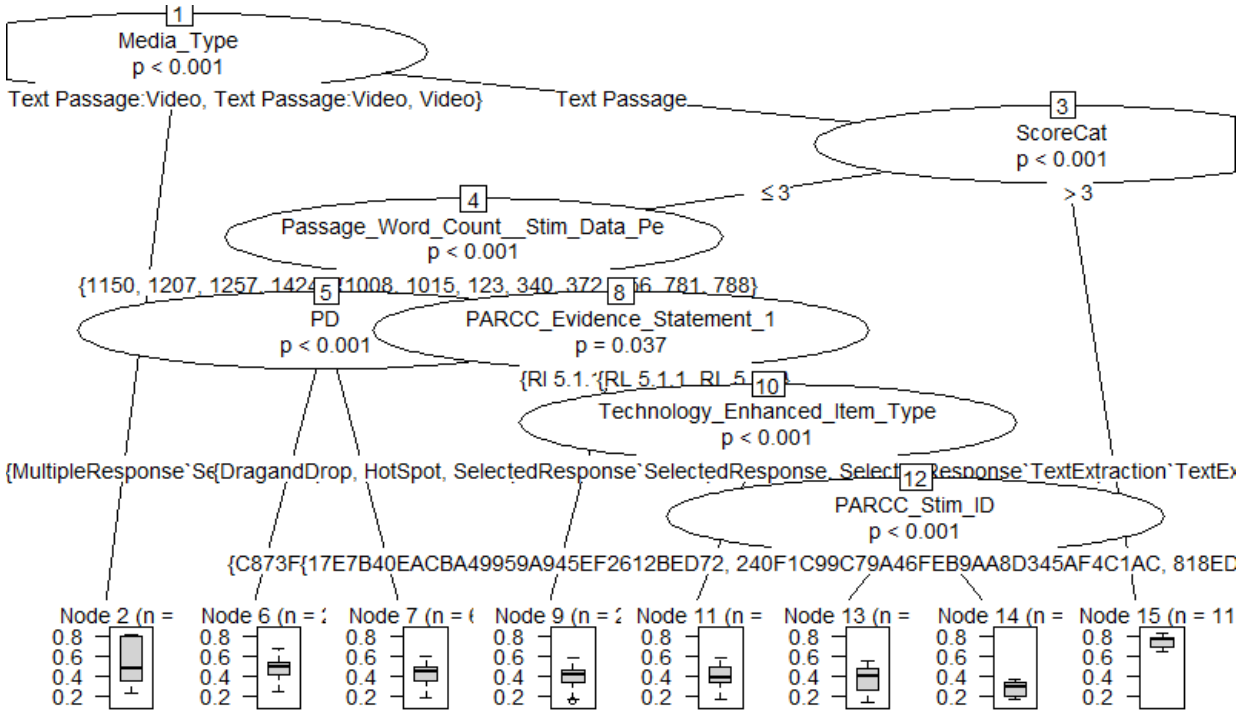


Figure G.12. Conditional Tree for Predicting Grade 5 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

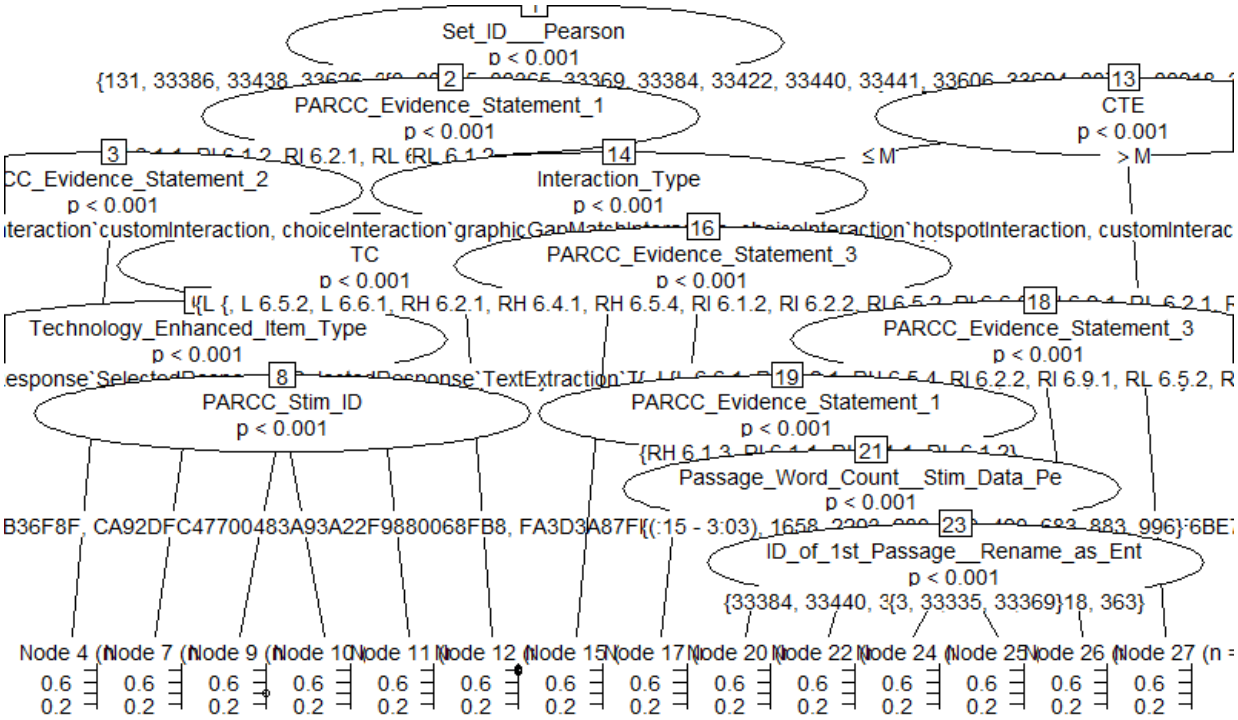


Figure G.13. Conditional Tree for Predicting Grade 6 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

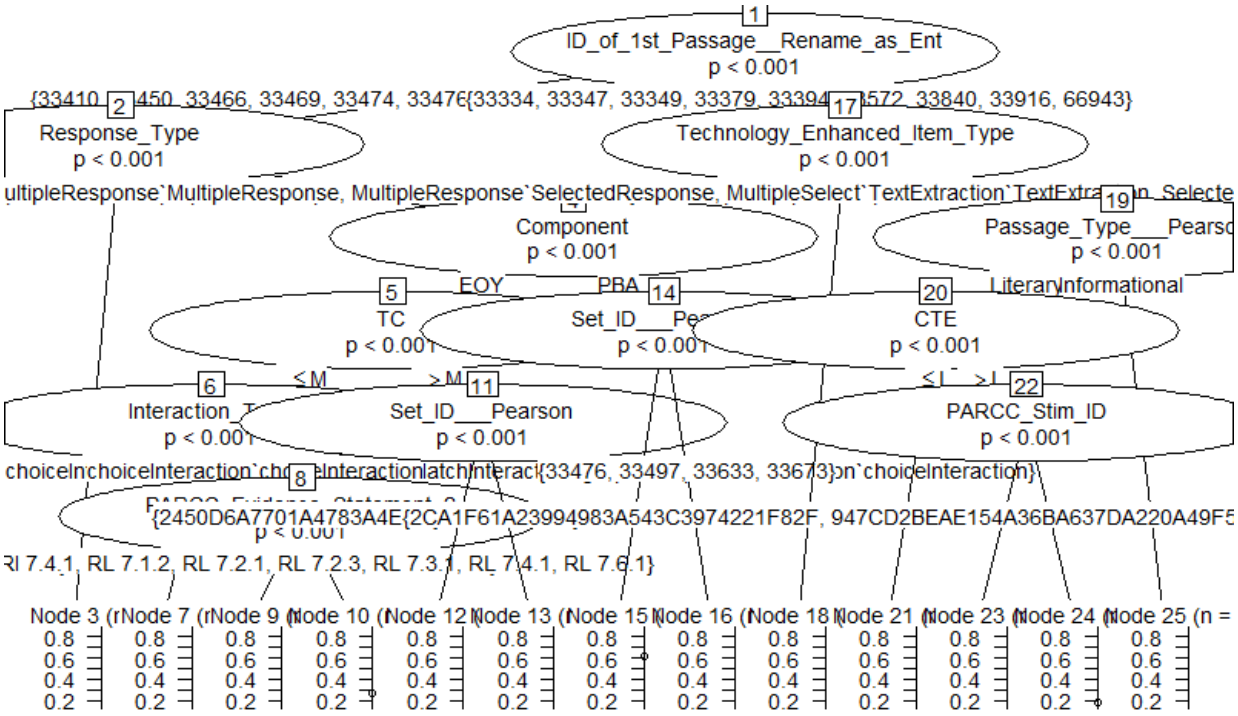


Figure G.14. Conditional Tree for Predicting Grade 7 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

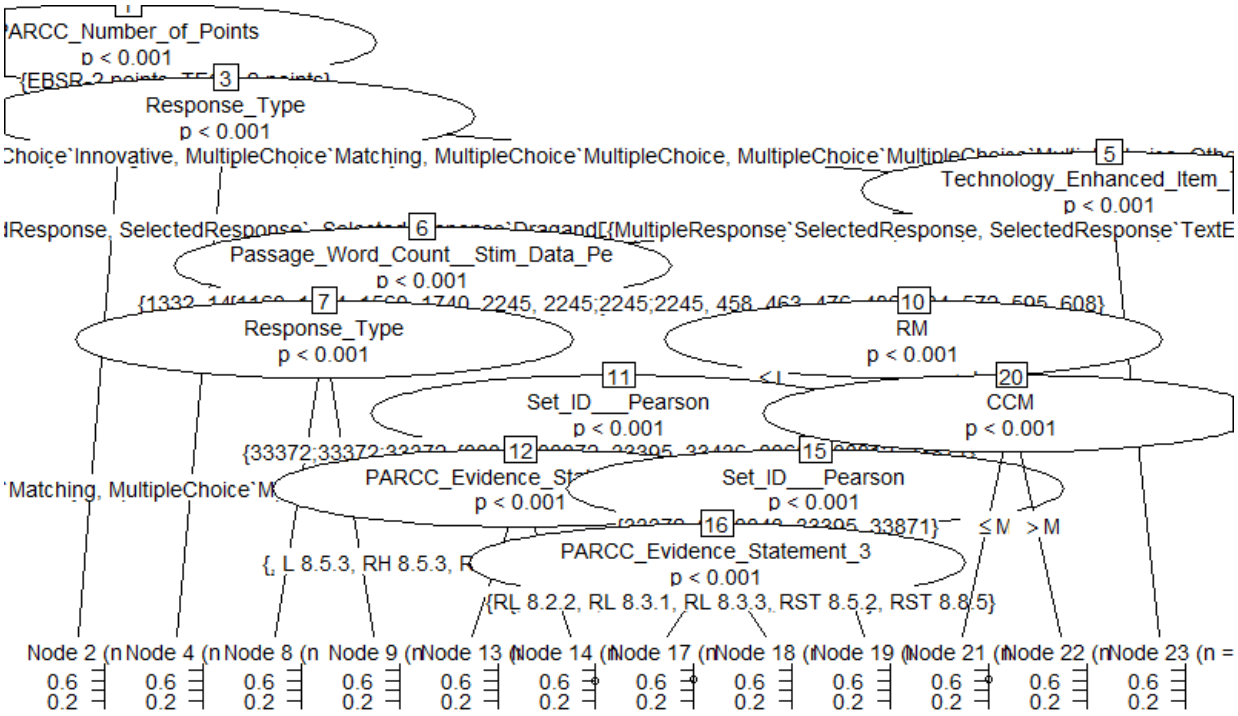


Figure G.15. Conditional Tree for Predicting Grade 8 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

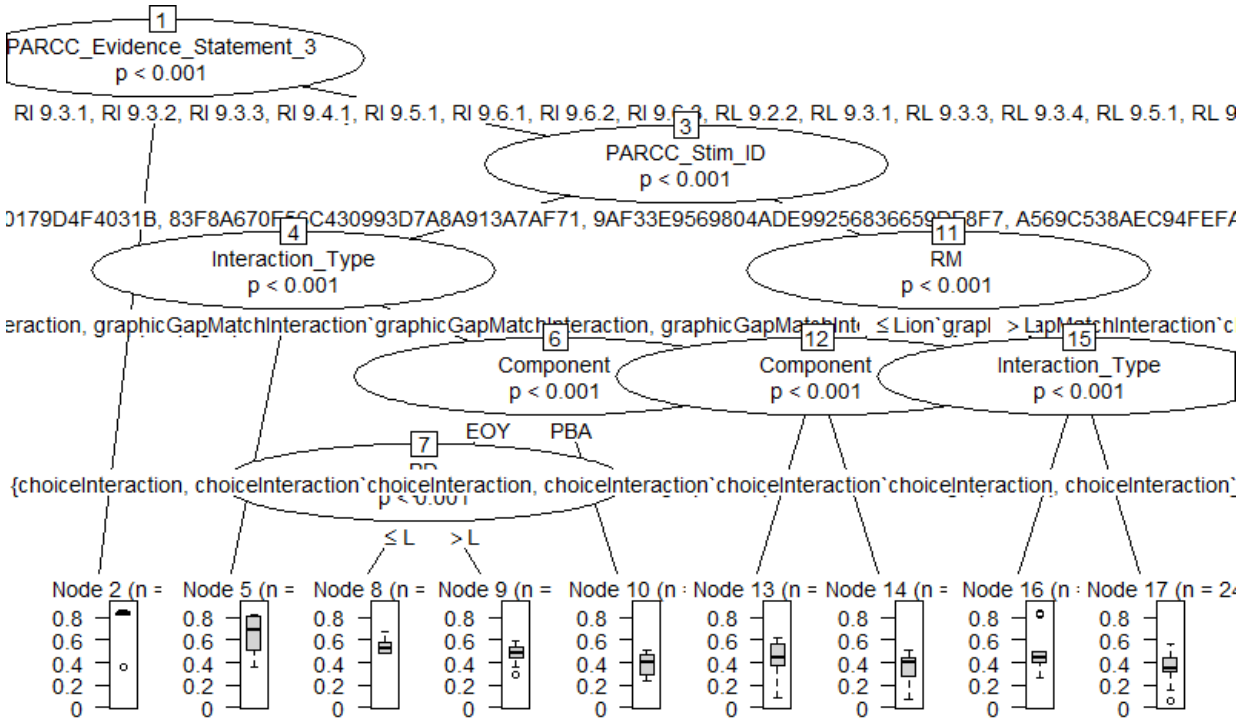


Figure G.16. Conditional Tree for Predicting Grade 9 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

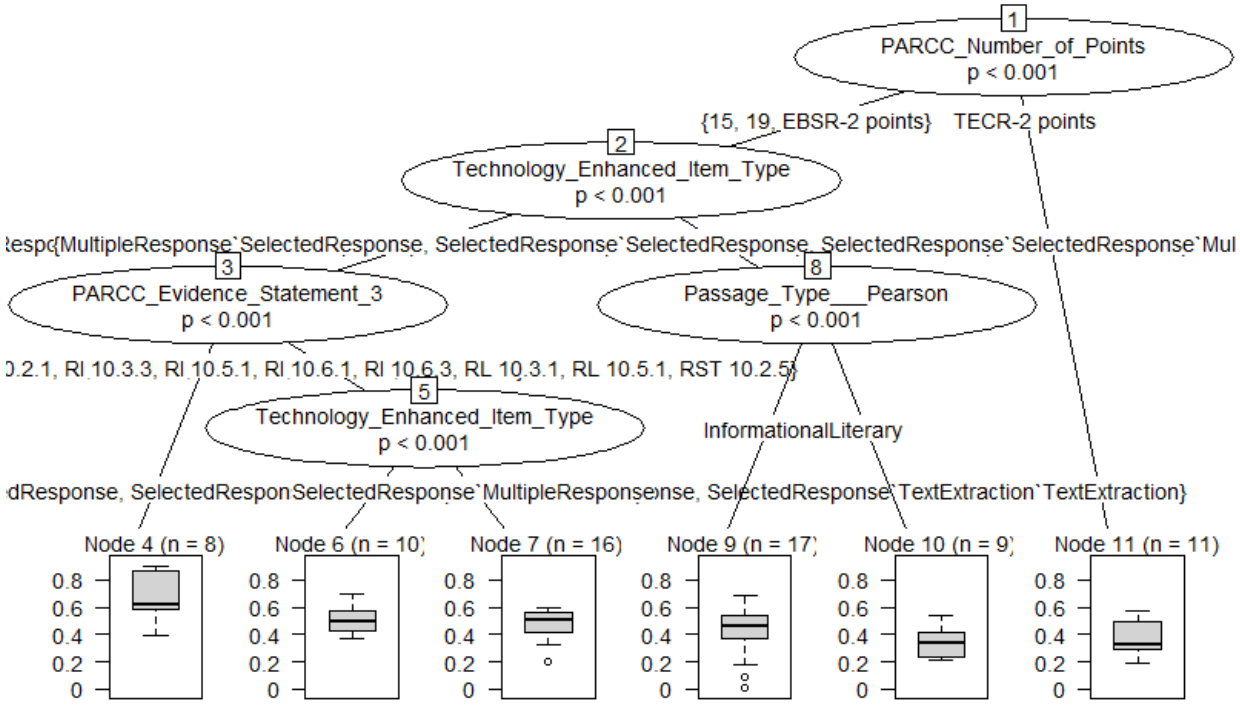


Figure G.17. Conditional Tree for Predicting Grade 10 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

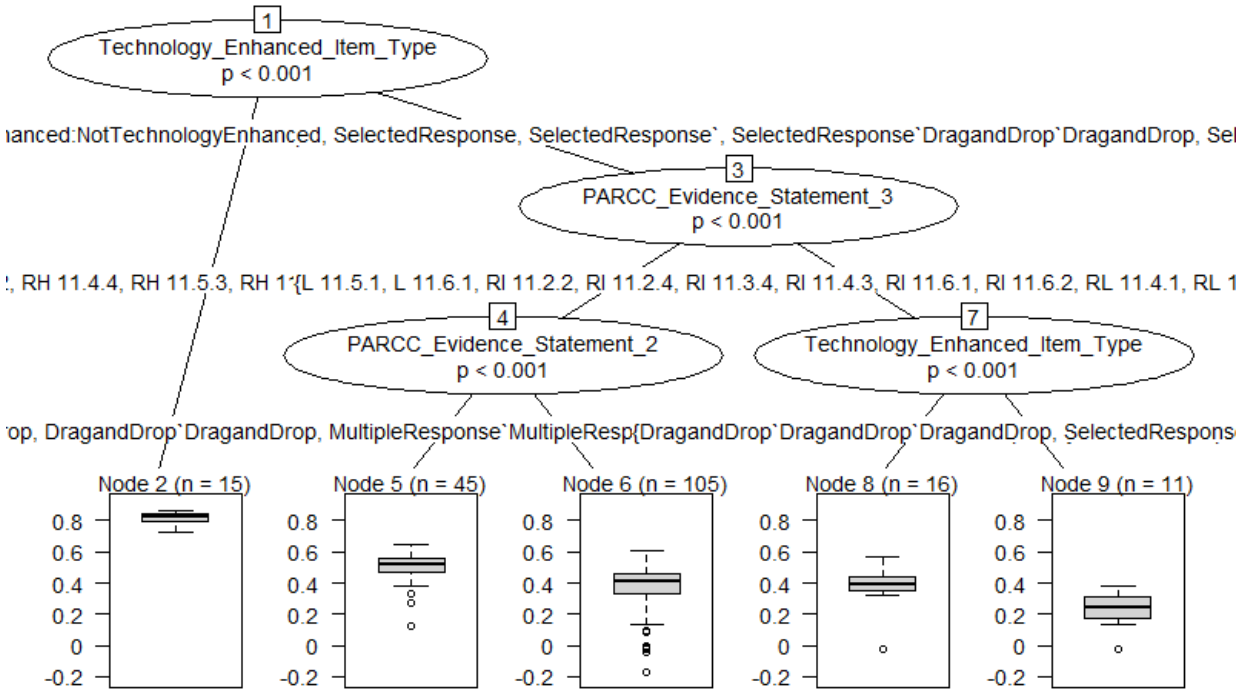


Figure G.18. Conditional Tree for Predicting Grade 11 ELA/L Point-Biserial Correlations from Cognitive Complexity Source Codes and Metadata.

Appendix H: Mathematics Conditional Trees Predicting Overall Cognitive Complexity using Cognitive Complexity Source Codes and Metadata as Predictors

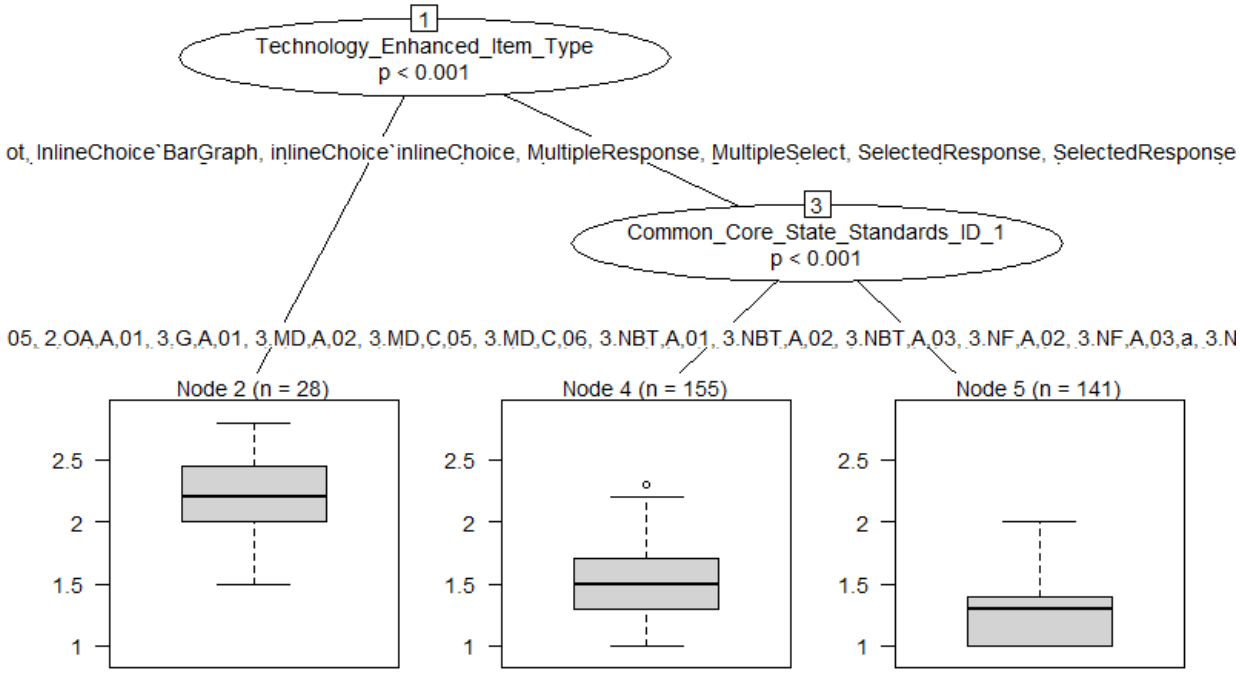


Figure H.1. Conditional Tree for Predicting Grade 3 Mathematics Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

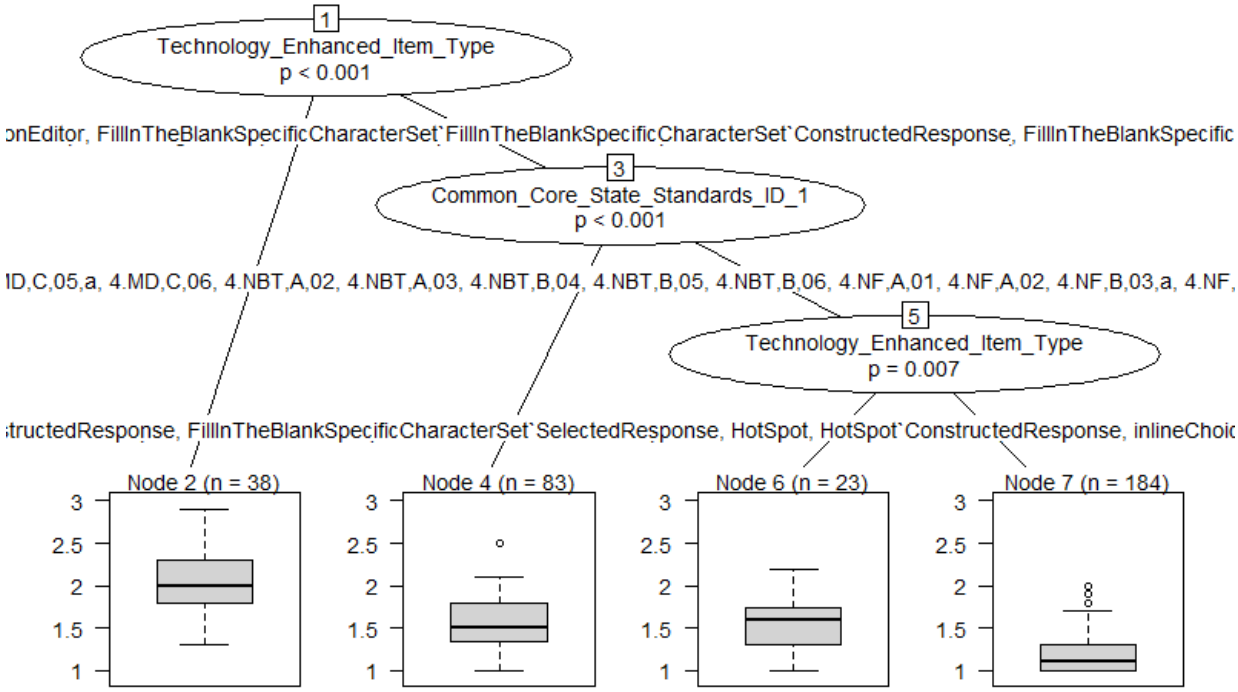


Figure H.2. Conditional Tree for Predicting Grade 4 Mathematics Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

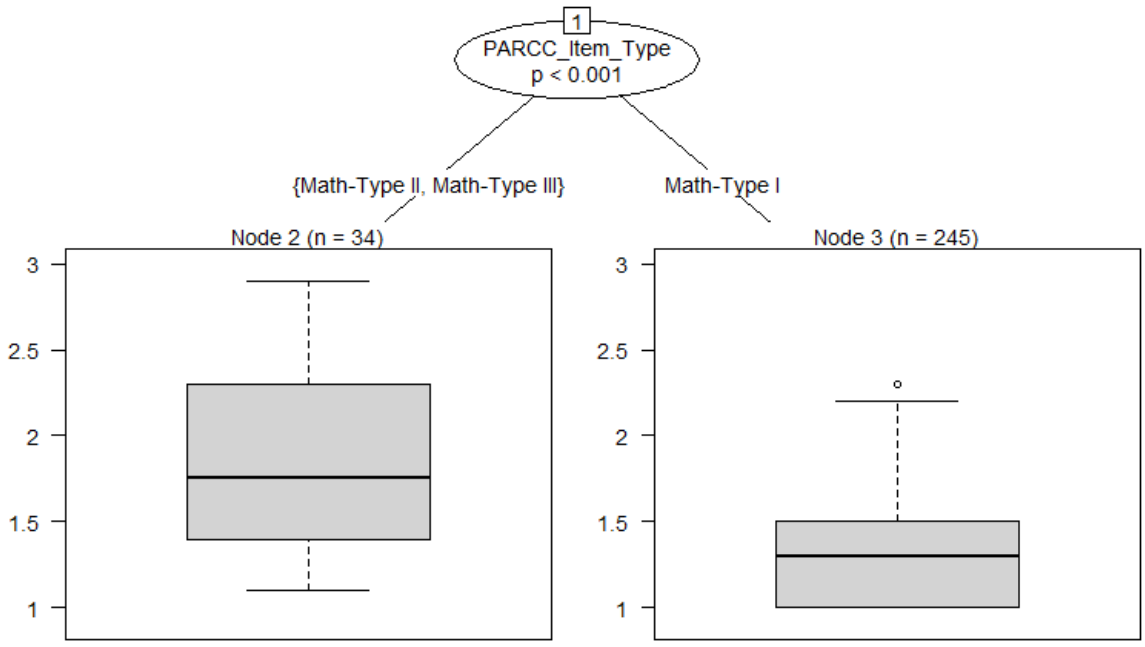


Figure H.3. Conditional Tree for Predicting Grade 5 Mathematics Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

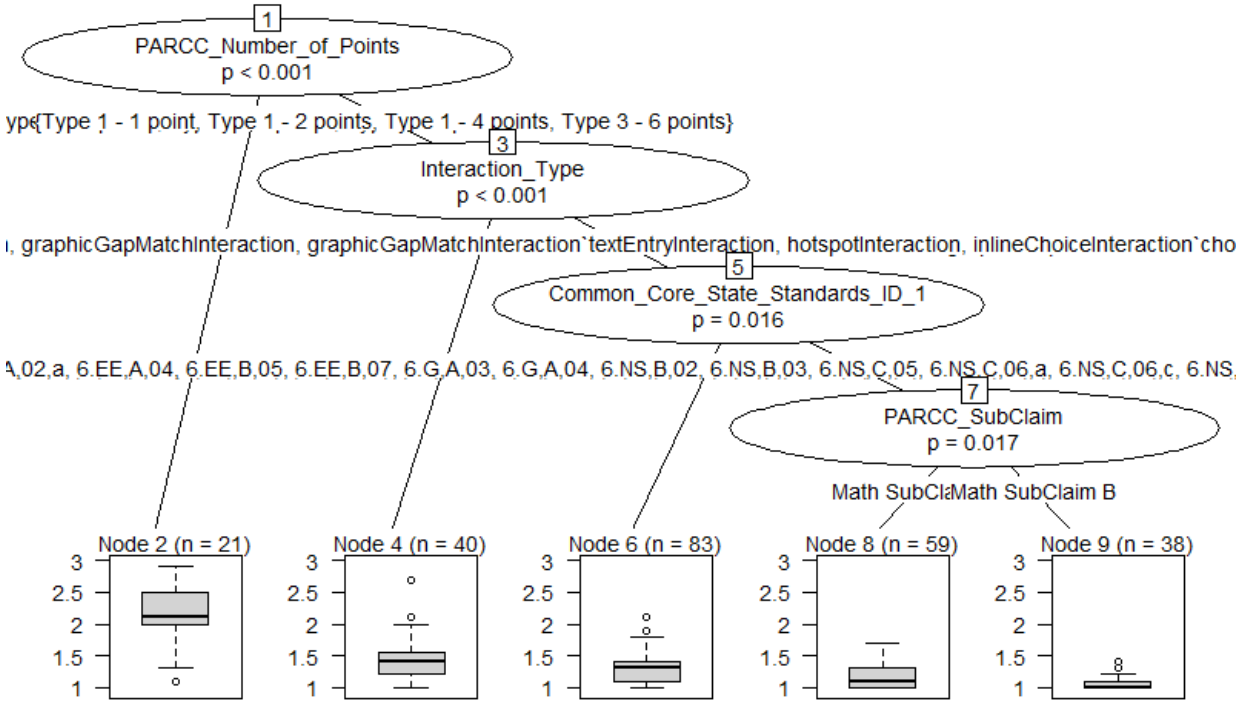


Figure H.4. Conditional Tree for Predicting Grade 6 Mathematics Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

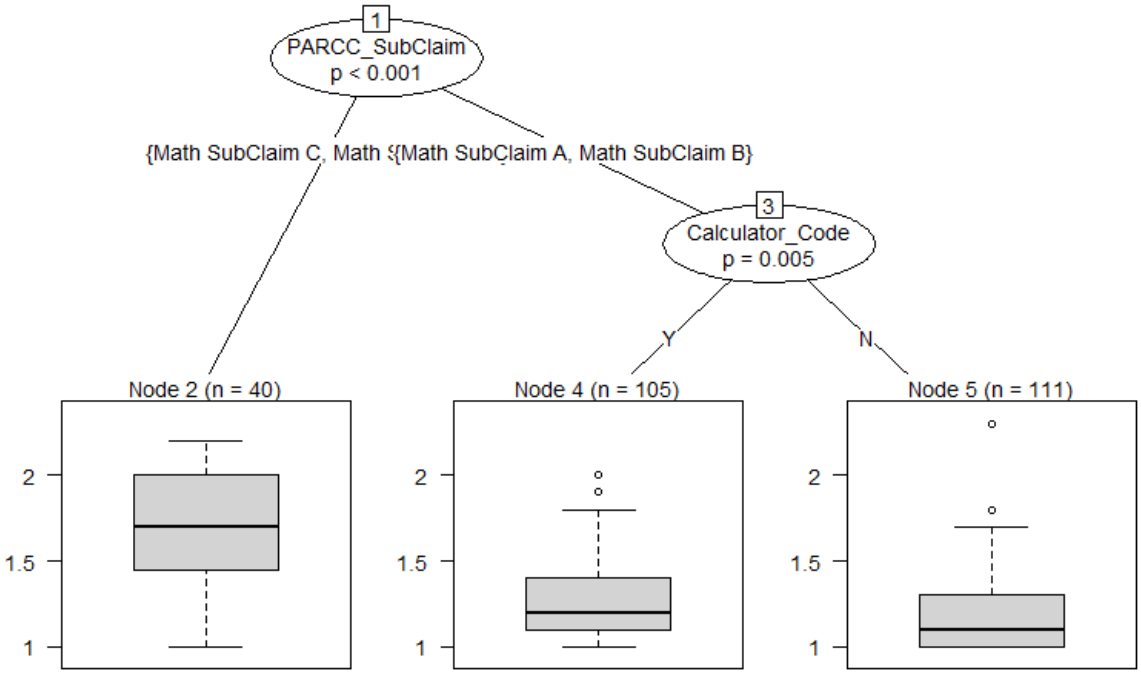


Figure H.5. Conditional Tree for Predicting Grade 7 Mathematics Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

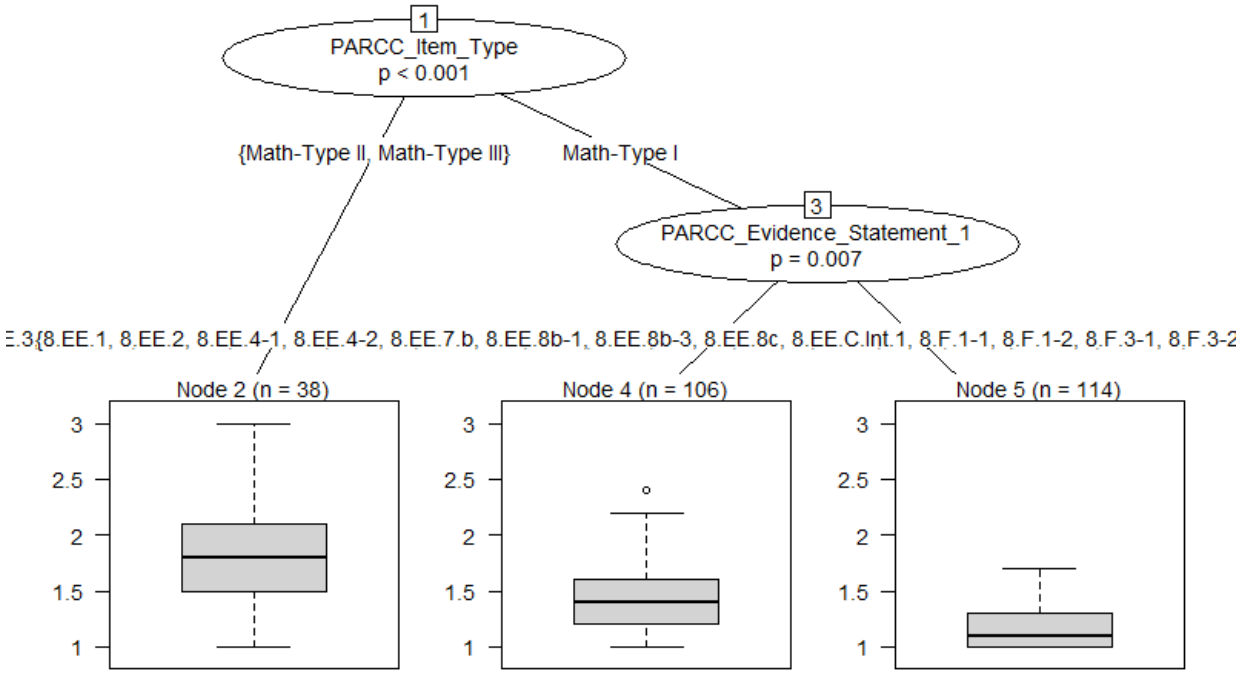


Figure H.6. Conditional Tree for Predicting Grade 8 Mathematics Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

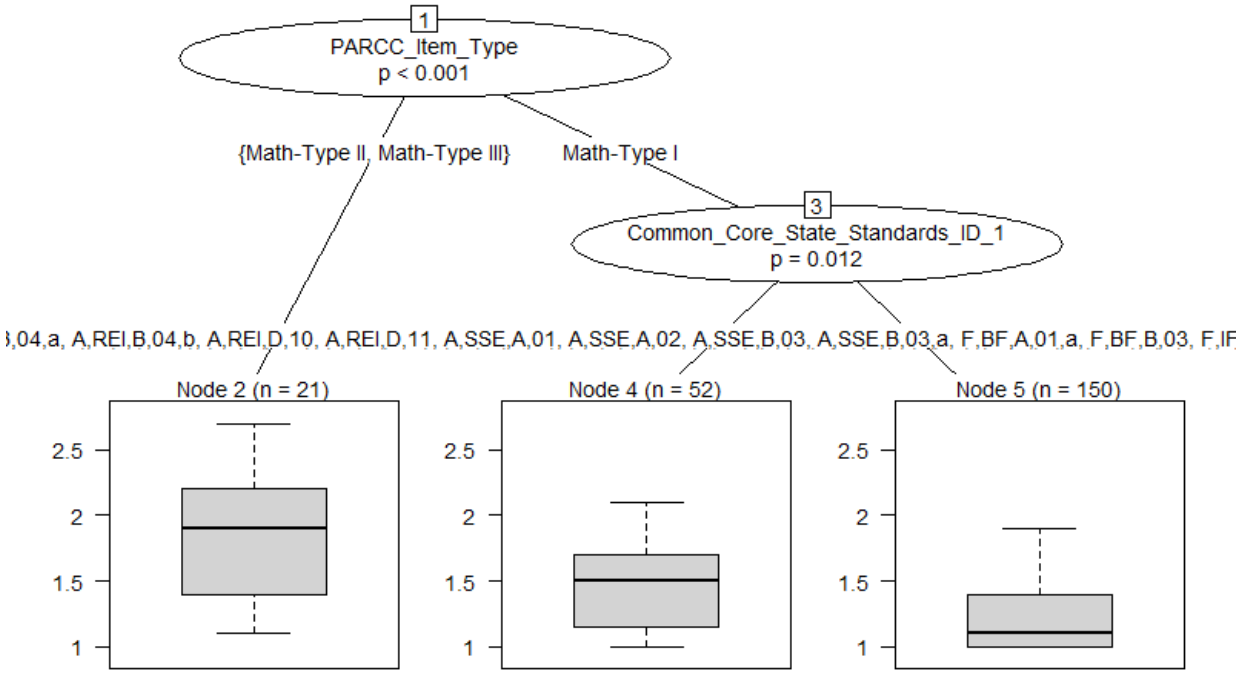


Figure H.7. Conditional Tree for Predicting Algebra I Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

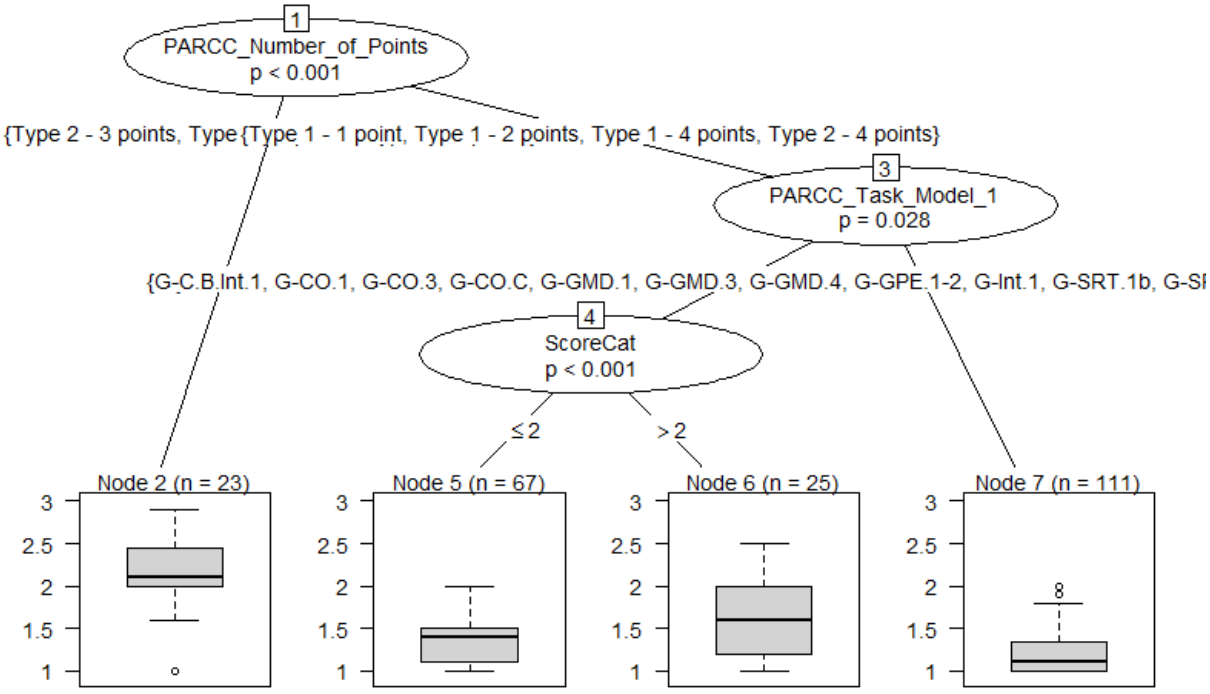


Figure H.8. Conditional Tree for Predicting Geometry Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

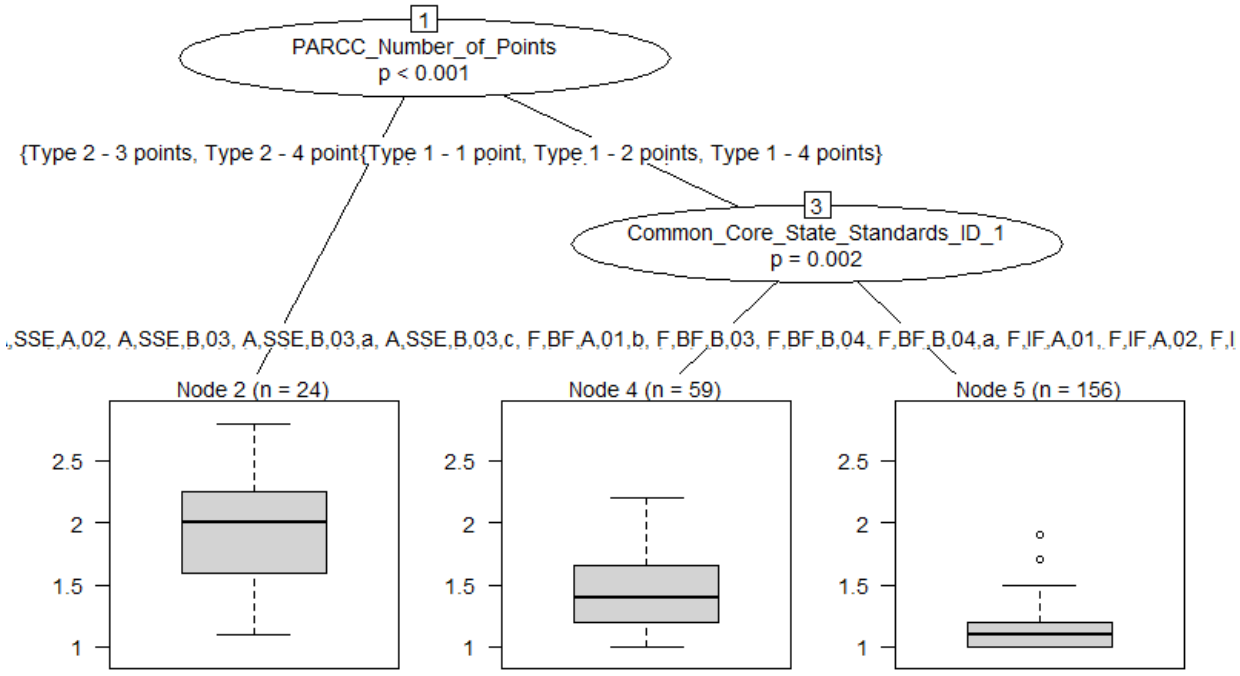


Figure H.9. Conditional Tree for Predicting Algebra II Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

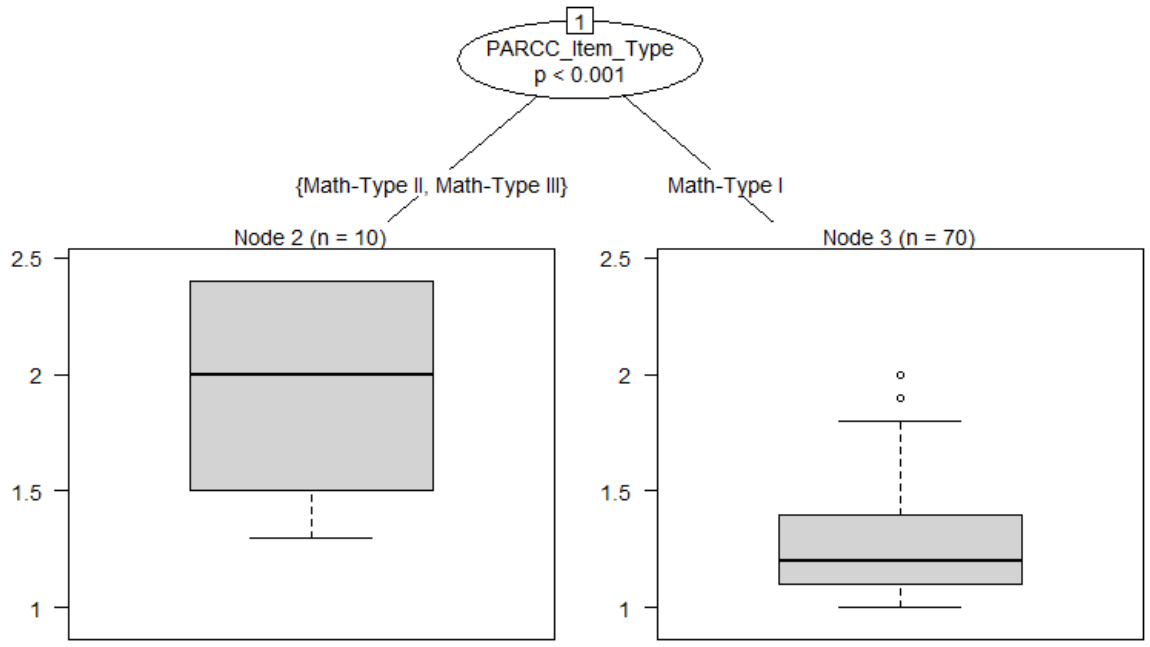


Figure H.10. Conditional Tree for Predicting Integrated Mathematics 1 Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

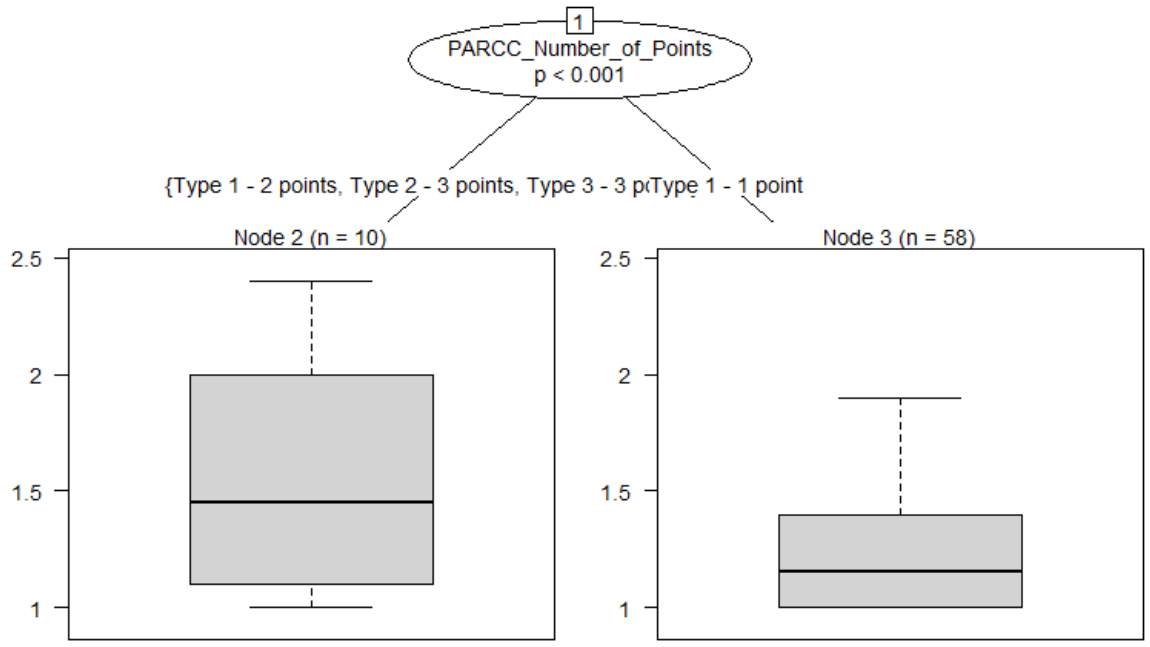


Figure H.11. Conditional Tree for Predicting Integrated Mathematics 2 Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

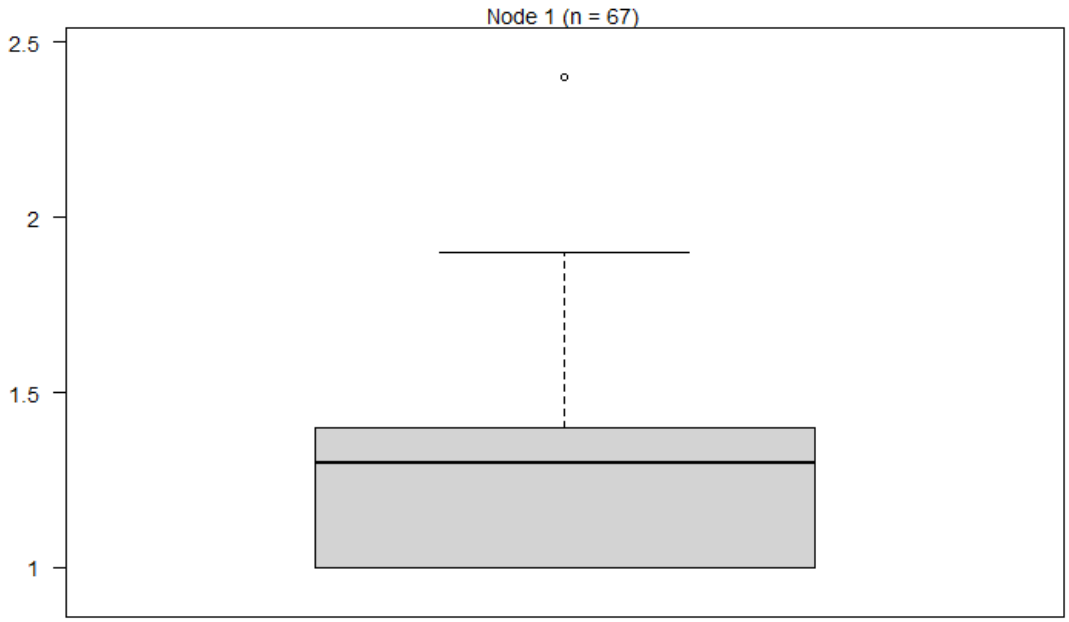


Figure H.12. Conditional Tree for Predicting Integrated Mathematics 3 Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

Appendix I: ELA/L Conditional Trees Predicting Overall Cognitive Complexity using Cognitive Complexity Source Codes and Metadata as Predictors

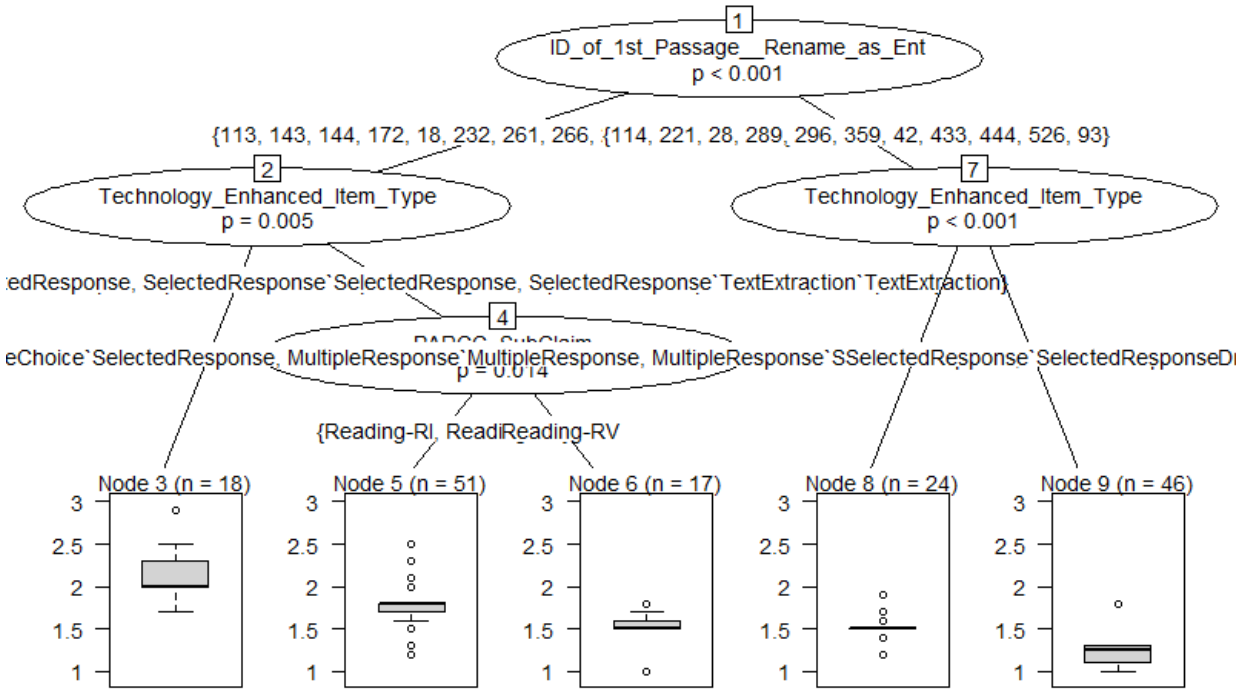


Figure I.1. Conditional Tree for Predicting Grade 3 ELA/L Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

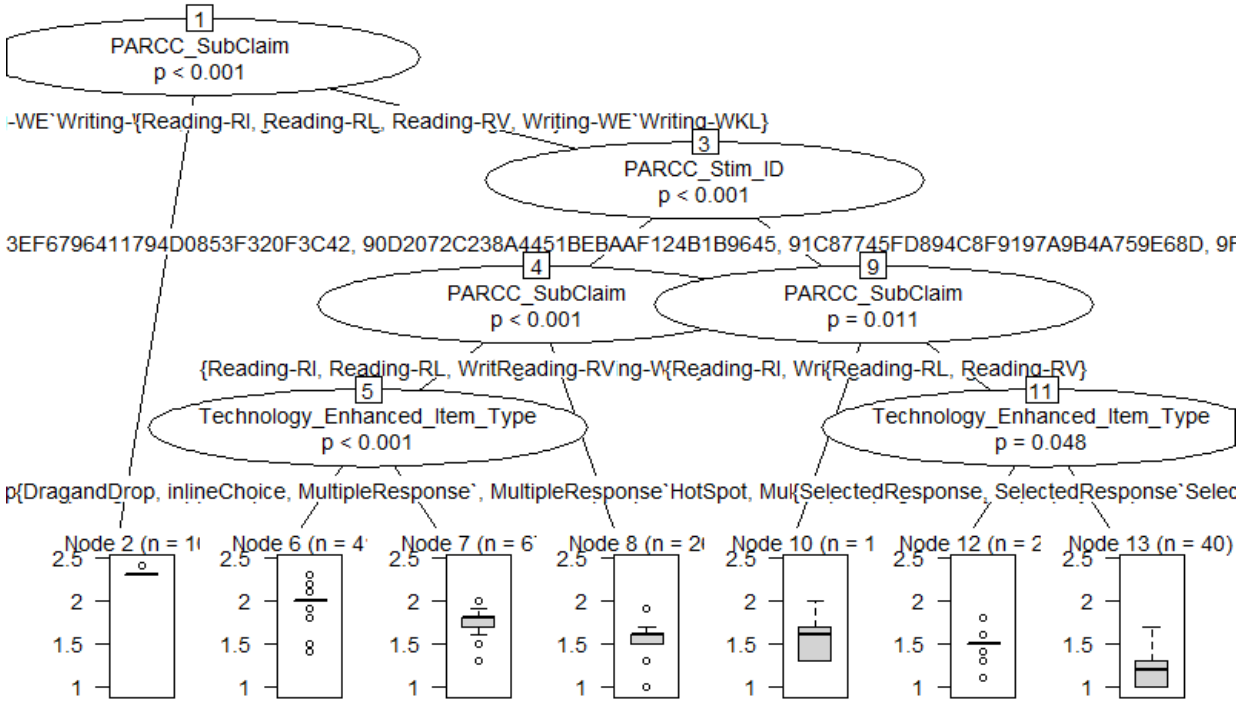


Figure I.2. Conditional Tree for Predicting Grade 4 ELA/L Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

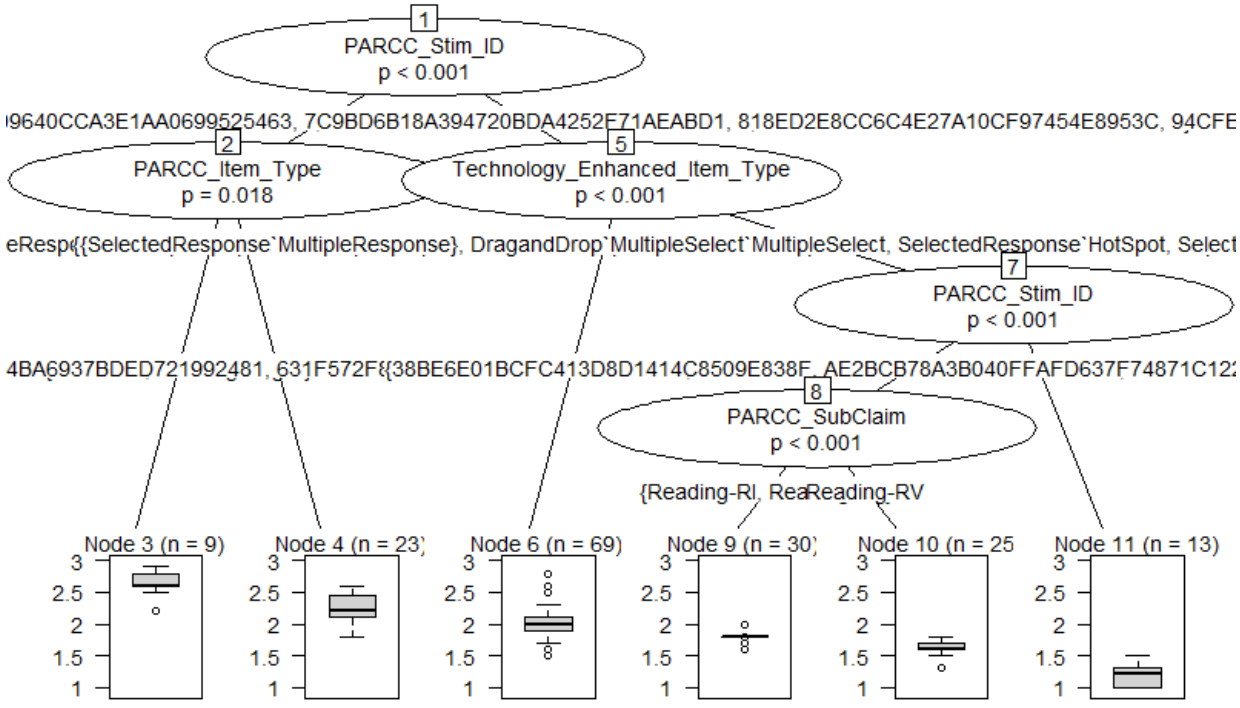


Figure I.3. Conditional Tree for Predicting Grade 5 ELA/L Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

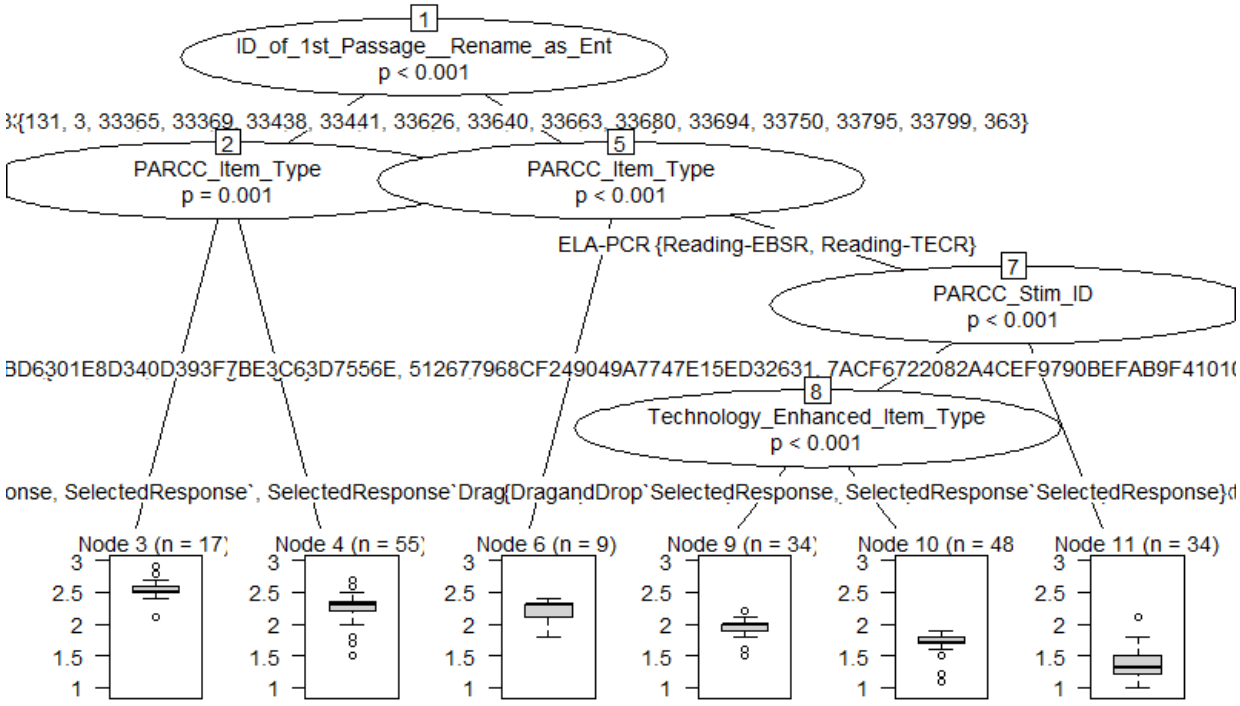


Figure I.4. Conditional Tree for Predicting Grade 6 ELA/L Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

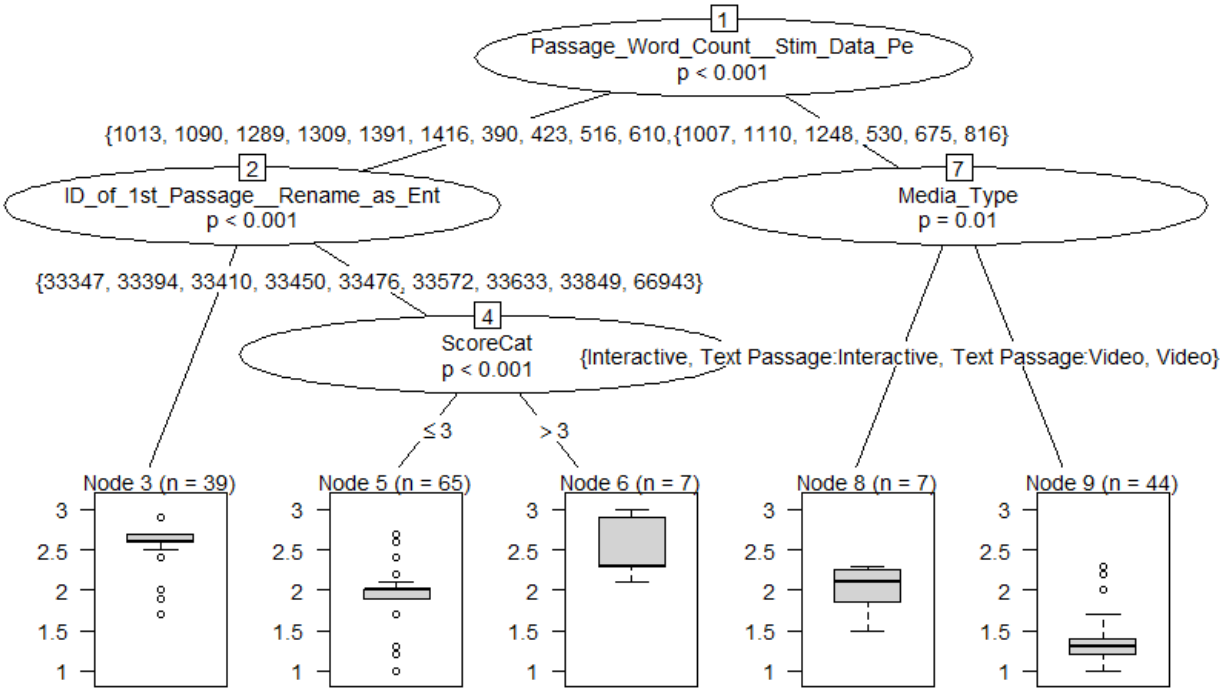


Figure I.5. Conditional Tree for Predicting Grade 7 ELA/L Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

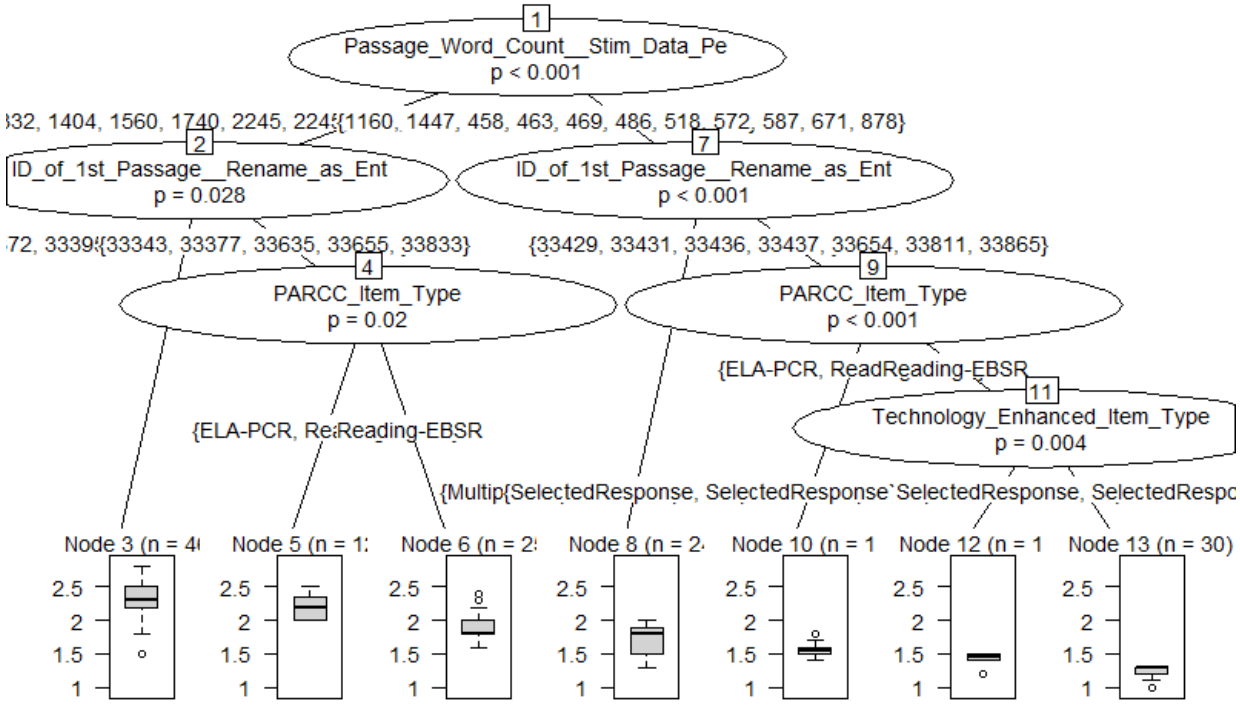


Figure I.6. Conditional Tree for Predicting Grade 8 ELA/L Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

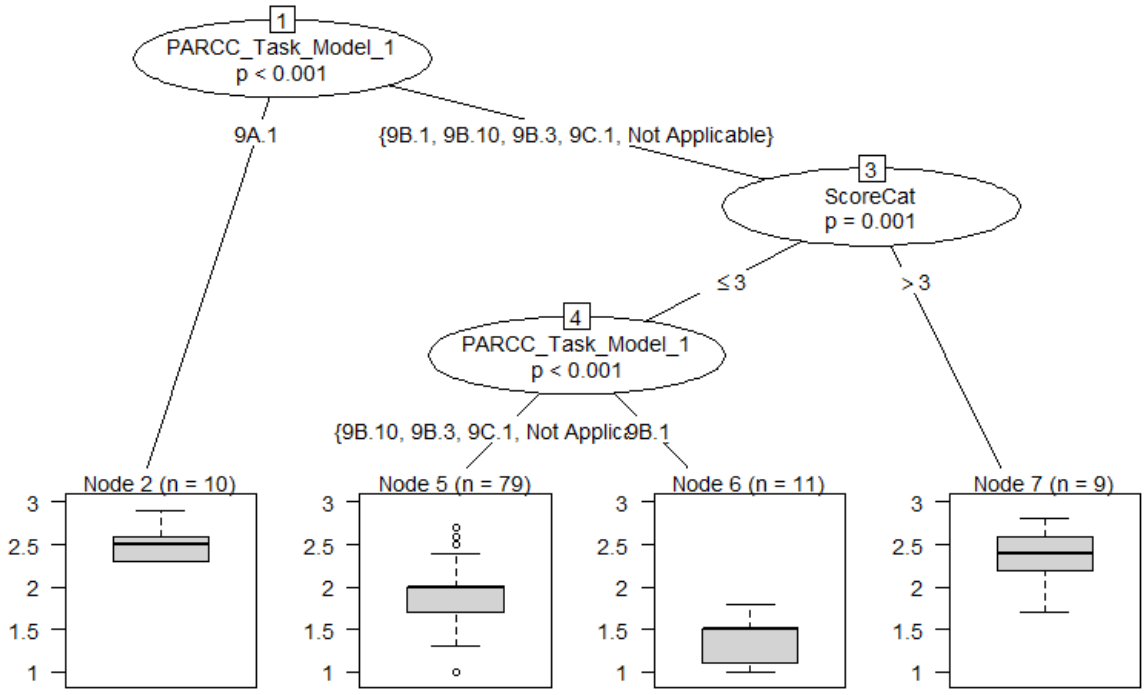


Figure I.7. Conditional Tree for Predicting Grade 9 ELA/LI Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

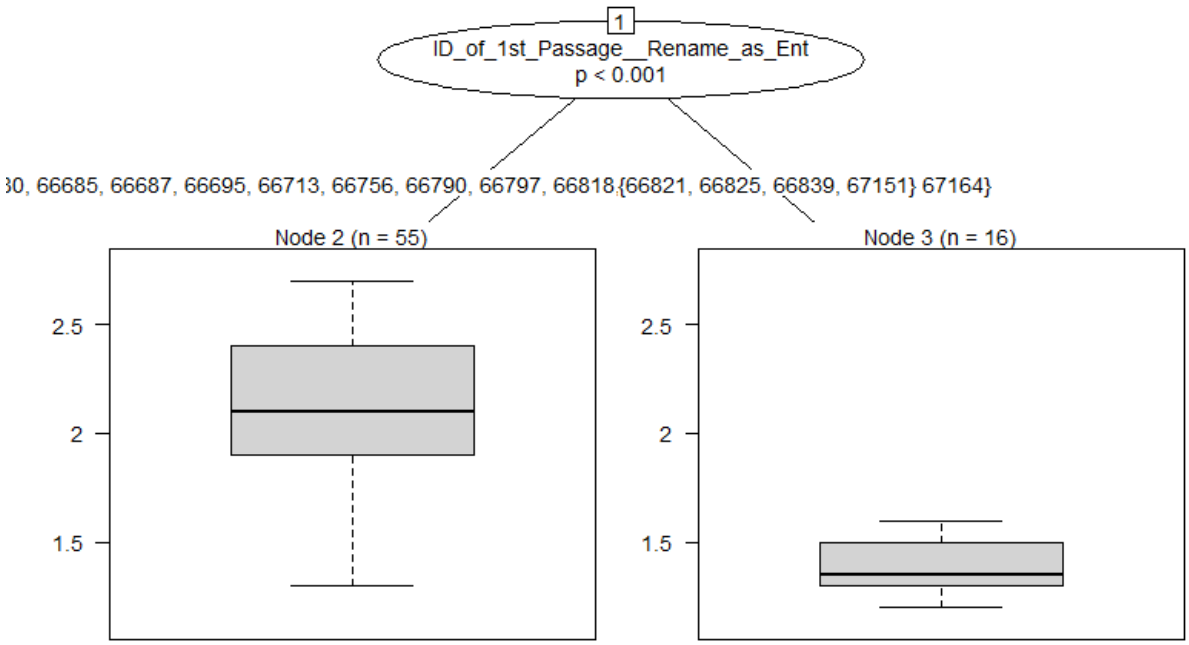


Figure I.8. Conditional Tree for Predicting Grade 10 ELA/L Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

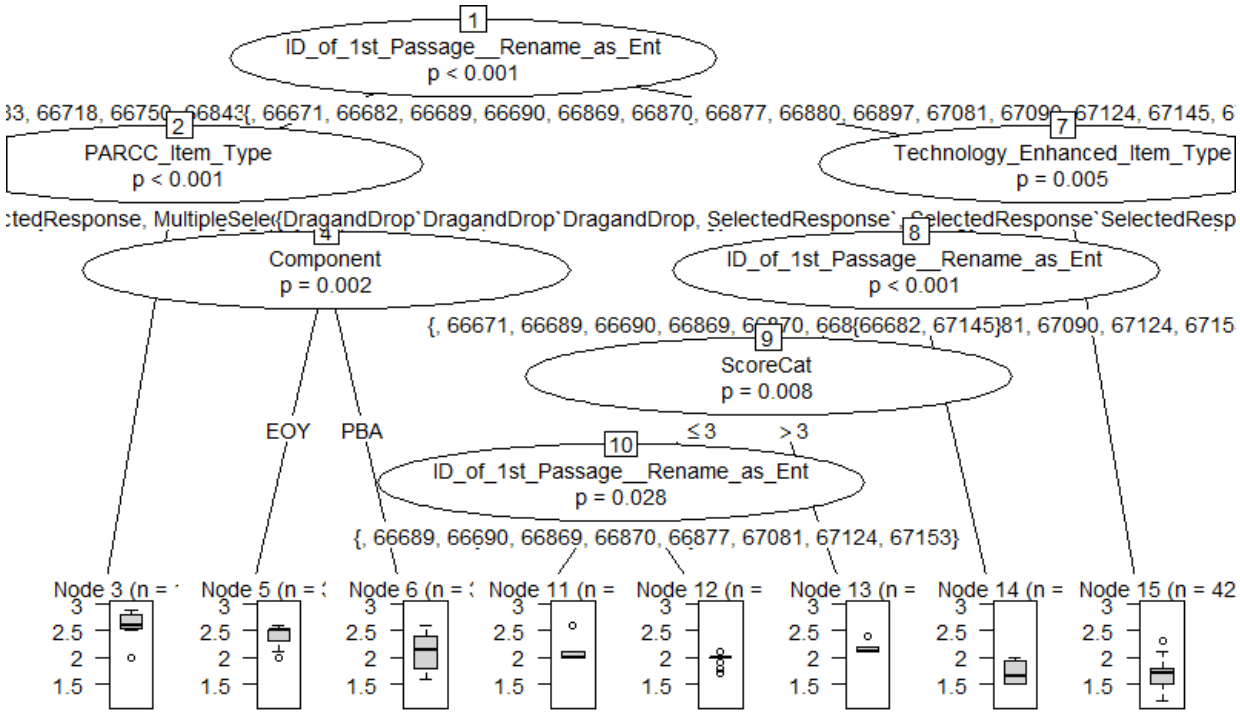


Figure I.9. Conditional Tree for Predicting Grade 11 ELA/L Overall Cognitive Complexity from Cognitive Complexity Source Codes and Metadata.

Appendix J: Survey Questions

See the separate file *z z 2. Appendix J SurveyMonkey_061715.pdf*

Appendix K: Survey Response Frequencies

See the separate Excel worksheet file *z z 3. Appendix K SurveySummary_06162015.xls*

Appendix L: Raw Responses to Open Ended Survey Items

See the separate Excel worksheet file *z z 4. Appendix L PARCC survey data Sheet_1 NO COMMAS.xls*

Appendix M: Item Metadata Variables Included in Analyses 1 and 2

See the following tabs in the separate Excel worksheet file *z z 5. Appendix M SurveySummary_06162015.xlsx*: Math_Vars and ELA_Vars

Appendix N: Presentation Slides with Summary of Coding Training and Validity Check Set Results

See the separate PowerPoint slides in file *z z 6. Appendix N PARCC CC for CA SIG session 2014 version B 04-04-14.pptx*

Appendix O: Presentation Slides for the Operational Working Group Briefings

See the separate PowerPoint slides in file *z z 7. Appendix O PARCC CC slides July briefings 07-23-15.pdf*