A CRITICAL EXPLORATION OF THE POTENTIAL UTILITY OF RULE INDUCTION DATA MINING METHODS TO "ORTHODOX" EDUCATION RESEARCH

by

Emi Iwatani

Sc.B. Biology, Brown University, 2002

M.A.Ed. Science Education, Wake Forest University, 2003

M.A. Philosophy, Boston University, 2008

M.A. History and Philosophy of Science, University of Pittsburgh, 2011

Submitted to the Graduate Faculty of

the School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Emi Iwatani

It was defended on

November 30, 2017

and approved by

Suzanne Lane, Ph.D., Professor, Psychology in Education

Jerrold H. May, Ph.D., Professor, Business Administration

Jennifer E. Iriti, Ph.D., Research Scientist, Learning Research and Development Center;

Adjunct Assistant Professor, Learning Sciences and Policy

Dissertation Advisor: Clement A. Stone, Ph.D., Professor, Psychology in Education

2017

A CRITICAL EXPLORATION OF THE POTENTIAL UTILITY OF RULE INDUCTION DATA MINING METHODS TO "ORTHODOX" EDUCATION RESEARCH

Emi Iwatani, PhD

University of Pittsburgh, 2017

Despite some theoretical promise, it is unclear whether rule induction data mining approaches (e.g., classification trees and association rules) add methodological value to "orthodox" education research, i.e., research unrelated to computer-based education. To better understand whether and how rule induction methods could be useful to education researchers, I explored whether they, relative to regression approaches, (1) improve classification accuracy, and/or (2) offer new avenues of explanation. Additionally, I aimed to illustrate a practical and principled way to use the various rule induction approaches so researchers can more easily choose to use it. To these ends, I conducted an extended literature review on rule induction methods, and re-analyzed two regression studies (Byrnes & Miller, 2007; Thomas, 2006) on the National Educational Longitudinal Study of 1988 using ten rule induction approaches. Data mining happened in two rounds for each study: first, by using only the predictors used in the original study, and second by using all reasonable and available predictors. I compared results across methods and rounds to better understand whether, how, and why the rule induction may provide additional insights.

I found that while rule induction approaches can be labor intensive and not necessarily more predictive than regression, they can provide unique descriptions of the sample that shows at-a-glance, how key predictors relate to each other and to the outcome. They can also help identify relationships between variables that held for some subgroups but not others. For example: (i)

rulesets induced from Byrnes and Miller's dataset suggested that Algebra 2 and math self-concept were positively related to 12th grade math scores, but only for those who were higher achieving in 8th grade math; (ii) association rules mined from Thomas' dataset suggested that factors such as school safety and honors program participation were more strongly associated with 12^{th} grade achievement for lower income and students with lower parental education. Thus, when relationships between the predictors and outcome may not be uniform across the population, rule induction can provide more information than regression in exploring those relationships. Lessons learned and recommendations on how to apply rule induction approaches are also discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I am grateful to many mentors, teachers, friends and family, for being an inspiration for this project and the wind beneath my wings. It has been an honor and delight to work with my advisor and committee members, who have been so gracious with their support. Dr. Clem Stone and Dr. Suzanne Lane guided and supported me so much with this project, and throughout my years in the Research Methodology program. I was fortunate to have met you, and so proud to be your student. Dr. Jen Iriti's mentorship and uncompromising commitment to utilization-focused evaluation has been inspirational for this project, and more generally to me as a program evaluator. Dr. Jerry May has been a generous teacher and guide in data mining, introducing me to the Pittsburgh Supercomputing Center and providing very helpful feedback, advice and comedic relief throughout this process. The late Dr. Kevin H. Kim, who might as well have been on my committee, was my first statistics teacher and the person who pointed me to rule induction as a possible dissertation topic that would align with my interests.

Many have helped cultivate my professional lens in culturally responsive education and research/evaluation methods, which was vital to sustaining my motivation and interest in this project. Special thanks to Dr. Rich Milner, Dr. Stafford Hood, and former supervisors, colleagues and students from Boston Arts Academy and Voices Against Violence, including Dr. Linda Nathan, Carmen Torres, Richard Carrington Sr., Peter McCaffery, Devon Madden, and Kevin Alton.

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the

Pittsburgh Supercomputing Center. I thank Rick Costa for responding to my numerous questions about the supercomputer, which was made possible through the XSEDE Extended Collaborative Support Service program. I also am grateful to various travel awards that helped disseminate some of the early findings, including the Kevin H. Kim Memorial Travel Grant, School of Education Dean's Office Travel Grant, Council for Graduate Student in Education Travel Grant, Graduate and Professional Student Government Travel Grant, and the American Education Research Association Minority Dissertation Fellowship Travel Grant.

Dr. Barbara Means and Linda Shear at SRI Education have also supported me in many ways, including limiting my project involvement at SRI so that I am able to devote enough attention to this project. Dr. Feifei Ye and my cohort-mates at the University of Pittsburgh Research Methodology program have been a terrific moral support. I also benefitted tremendously from administrative support from Mary Mollo, Angie Chyne Bottles, Cole Cridlin, and Josh O'Malley at the University of Pittsburgh Psychology in Education Department. Finally, I thank my husband Eric, my parents Yuko and Shigeo, grandparents Rei and Sanshiro, brother Hiroshi, sister-in-law Chiho, niece Miwa, and my dear friends Hasna, Lisa, Jill, Kristina, Molly and Nonye for their unwavering love and support throughout.

1.0 INTRODUCTION

Can rule induction data mining methods be useful in orthodox educational research? Data mining is a set of approaches that automatically or semi-automatically detect quantitative patterns in datasets. It is intended for knowledge discovery in very large datasets, and therefore potentially useful for education research. While becoming more popular in education as a methodological approach, most applications of data mining have been limited to research on computer-based learning (e.g., intelligent tutoring systems, Gobert, Sao Pedro, Raziuddin, & Baker, 2013; learning management systems, Valsamidis, Kontogiannis, Kazanidis, Theodosiou, & Karakos, 2012). Many of the applied works on more traditional educational topics have been case studies by educational practitioners rather than researchers, or in other social or applied science fields such as business, computer science, and public health. To date, it is unclear whether data mining approaches add any methodological value to "more orthodox" research in education, i.e., research on non-computer-based learning. This could be, in part, because an inquiry of whether a new methodology is 'helpful' to applied orthodox education research requires several very different kinds of skills. It requires a very good understanding of what applied education researchers do and care about so that one can determine specific niches and questions that could benefit from the new method. It also requires solid knowledge of both the extant and new methods, and the ability to design studies that could demonstrate whether, when, and how (if at all), it would be helpful to apply the new method.

Rule induction methods—including decision trees, association rule mining and sequential covering—are a commonly used subset of data mining approaches. Described in more detail in the

next chapter, not only do these methods generate predictions about individual cases in a dataset, they are distinctive in that they generate potentially illuminating *if-then* rules about the predictors in the dataset (Tung, 2009).

Three main approaches to rule induction include sequential covering, decision trees and association rule mining. Conceptually, sequential covering works by identifying an *if-then* rule that applies robustly to many cases, separating out cases that are correctly identified, and repeating that process with cases that are not yet "covered" by that rule. The result is a list of rules that can be applied to new data to predict the outcome (Figure 1).

Decision tree algorithms take a divide-and-conquer approach: They begin by identifying the variable that is most predictive of the outcome, using that variable to split the cases into two or more subsets (e.g., more likely vs less likely to have outcome X), and repeating the process with each of these subsets until a stopping criteria is met. The result is a set of rules that are often expressed in the form of a tree (Figure 2).

Association rule mining algorithms identify any if-then patterns in the dataset that are sufficiently robust. Users typically need to specify the minimum threshold of accuracy and generality. They can also typically specify what must be included in the left-hand-side (*if* portion) and right-hand-side (*then* portion) of the rule. This generally results in a very large set of rules. Automatically or semi-automatically screening for interesting rules among them has been a large area of research in machine learning.

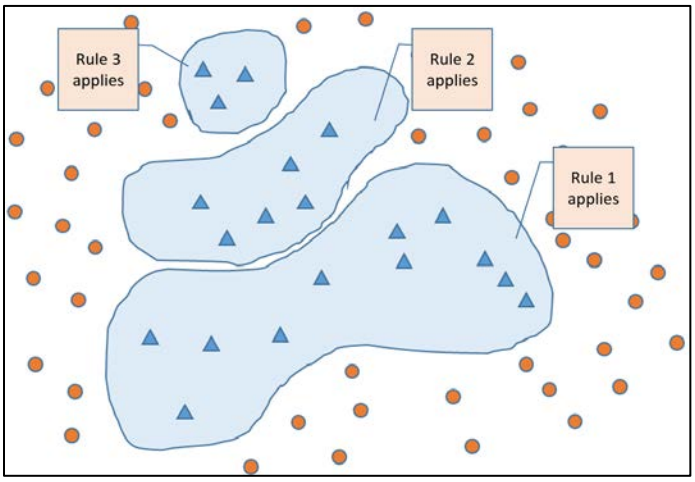
Prediction of academic achievement (fictional example)

Rule 1: *If* student attends school ABC *and* is a member of the debate team, *then* the student is likely to be high achieving.

Rule 2: *If* Rule 1 does not apply, but the student has taken Algebra 1 in 9th grade, *then* the student is likely to be high achiving.

Rule 3: *If* Rule 1 and 2 do not apply, but the student has received an academic award in 10th grade, *then* the student is likely to be high achieving

Rule 4: If Rules 1-3 do not apply, *then* the student is likely to be lower achiving.



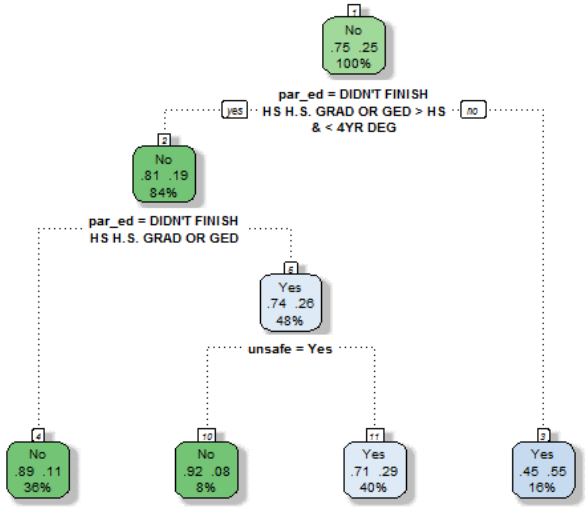
triangle = high achieving on standardized test

circle = lower achieving on standardized test

Figure 1. Example of an ordered decision list from sequential covering

Prediction of academic achievement

- *If* the student's parent has received at least a 4-year college degree, *then* the student is likely to be high-achieving (rule applied to 16% of the sample, and correctly predicted 55% of them).
- *If* the student's parent has received some college education but not a 4-year degree, *and* the student feels safe in school, *then* the student is likely to be high-achieving (rule applied to 40% of the sample, and correctly predicted 29% of them).
- *If* the student's parent has received some college education but not a 4-year degree, *and* the student feels unsafe in school, *then* the student is likely to be lower achieving (rule applied to 8% of the sample, and correctly predicted 92% of them).
- *If* the student's parent has a high school degree or less, *then* the student is likely to be lower-achieving (rule applied to 36% of the sample, and correctly predicted 89% of them).



Yes = high achieving
No = lower achieving

Figure created with R (R Core Team, 2016) *rpart* (Therneau, Atkinson, & Ripley, 2015) and *rattle* (Williams, 2011)

Figure 2. Example of a decision tree in ruleset form (left) and tree form (right)

Rule induction approaches have been popular because of their flexibility, robustness to outliers, and ease of use and understanding. Most distinctively, rule induction approaches can identify interesting relationships among predictors and outcomes that apply only to a subset of the data (Hand, 1997; Witten, Frank, & Hall, 2011) — relationships that may even contradict the general trend. This dissertation inquires how and to what extent rule induction data mining approaches can be useful to orthodox education research.

1.1 PROBLEM STATEMENT

1.1.1 Rule induction data mining—A promising addition to the toolbox of orthodox education research?

In theory, there are many good reasons to use rule induction data mining approaches in orthodox education research. As educational datasets become larger and more complex, data mining's capacity to identify complex patterns in such datasets could be especially helpful to accelerating and deepening our understanding of learning. Since learning is a complex process mediated by numerous internal and environmental factors, data mining could help screen for interesting and unexpected patterns that education researchers may want to further pursue. Rule induction methods are particularly useful since the link between the predictors to the outcome—i.e., the rules—have easily interpretable meaning. In contrast, many other data mining methods (such as neural networks) are considered "black box approaches"—the connection between the predictors and outcomes generally involve complex weighing and transformations, and are generally uninterpretable.

When studying education, it is often desirable to learn what works well for specific subgroups, even if they may not work as well for the general population. Education researchers studying the effects of a curriculum would be interested to know, for example, if its effect differed across geographic regions, if the geographic influence differed across students' primary learning style, and so on. Because education researchers tend to be interested in the nature of good learning, and because it is well-accepted that there are multiple and complex pathways to good learning, methods that help identify or clarify pathways are naturally of high relevance to their research (Martin & Sherin, 2013). Rule induction data mining methods, which help characterize some of the complex relationships among variables, therefore have the potential to be highly relevant to education researchers.

1.1.2 Current state of use

While most data mining in "orthodox" education research has involved some type of rule induction method, and while education researchers are certainly becoming aware of its potential value (e.g., Delen, 2012; Faulkner, Davidson, & McPherson, 2010; Flores, Inan, & Lin, 2013), its use has been infrequent, particularly in contexts outside of computer-based education. In reviewing the use of educational data mining (EDM) in the context of K12 dropout prevention. Márquez-Vera, Cano, Romero, and Ventura (2013) found that hardly any research had been done in that context, with most EDM research applied in the context of online or distance education in higher education. Further, while recent ERIC[1] searches for "regression" "logistic regression," and "HLM" resulted in 20169, 2675 and 948 hits, respectively, searches for "decision tree," "regression tree,"

[1] Search was conducted in August 2015.

"classification tree," "recursive partitioning" and "association rule mining," resulted in just 121, 17, 18, 8 and 15 hits, respectively. Among them, only 21 peer-reviewed studies applied rule induction methods on orthodox education topics.

In addition, there is room for improvement in how rule induction methods are used in education research. Over three-quarters of the aforementioned 21 peer-reviewed studies, used just one rule induction algorithm for their analysis. This increases the risk of algorithmic bias, since rule induction algorithms can highlight very different relationships among variables (Iwatani, Stone, & Shealy, forthcoming). Almost half of the studies had sample sizes and/or number of predictors that were likely to be too small to reliably detect predictor-outcome relationships that could be unique to subgroups. Eight of the studies did not cross-validate their results, which increases the risk that the results from these studies are not generalizable beyond that particular sample (Breiman, Friedman, Olshen, & Stone, 1984; Witten et al., 2011). Most studies did not attend to rule induction's unique ability to identify relationships between predictor and outcome among subgroups—relationships that may contradict the general trend. Only four (Delen, 2006; Flores et al., 2013; X. Liu & Ruiz, 2008; X. Liu & Whitford, 2011) used US national or international educational datasets, and just one (X. Liu & Whitford, 2011) was free of serious methodological concerns. These suggest that the (orthodox) education research community would greatly benefit from a methodologically informed, comprehensive and critical examination of the potential utility of data mining to their field.

1.1.3 Barriers to adoption and motivation for this project

The diffusion of innovations theory by Everett Rogers (2003) is helpful for assessing the barriers to widespread adoption of rule induction data mining approaches in orthodox education research.

According to the theory, between 49-87 percent of the variance in the rate of adoption of an innovation can be explained by five perceived attributes of innovation, including relative advantage, compatibility, complexity, trialability, and observability.

Relative advantage, which relates positively to the rate of adoption, is "the degree to which an innovation is perceived as being better than the idea it supersedes" (Rogers, 2003, p.229). Since most orthodox education researchers do not know what rule induction approaches are, they are not in the position to consider their relative advantages to other analytic approaches. That notwithstanding, even those who are familiar with rule induction methods would have difficulty assessing their relative advantage because there is little information on factors that would matter to potential adopters: how these methods compare to what they know and be usefully applied to the field and to *their* research.

To be sure, there are studies that compare data mining algorithms, including rule induction approaches, on real or artificial datasets (e.g., Curram & Mingers, 1994; Finch & Schneider, 2006; Holden, Finch, & Kelley, 2011; Lim, Loh, & Shih, 2000; Michie, Spiegelhalter, & Taylor, 1994; Vaughn & Wang, 2008). However, these studies generally reveal very little about the relative advantage of rule induction methods as they tend to focus solely on predictive accuracy and processing speed, and confirm what is now a truism in machine learning: that performance greatly depends on the nature of the dataset, and there is no universally superior learning algorithm (Wolpert, 2012; Wolpert & Macready, 1997). For example, in a comparison of over 30 data mining approaches on 32 datasets, Lim et al. (2000) found no statistical differences in the mean error rates, and concluded that "[the] differences are also probably insignificant in practical terms" (p.225). Moreover, very few such studies have used datasets of interest to orthodox education researchers (see Section 2.5, below for exceptions).

Compatibility is the extent to which an innovation is perceived as being consistent with the adopters' values, needs and past experiences. Orthodox education researchers are likely to need new ways to efficiently explore data as datasets are becoming larger, and consumers of research increasingly expect that such datasets be examined. Yet because most researchers are currently unfamiliar with the theory and potential applications of data mining, they are less likely to consider it as a solution to their need. In addition, researchers and statisticians in orthodox education have tended to favor theory-driven, confirmatory analyses over exploratory analyses. That data mining is more suited for exploratory rather than confirmatory purposes may serve as a barrier to adoption as it contradicts these consciously or subconsciously held values and approaches (see Section 2.2.2).

Complexity is "the degree to which an innovation is perceived as relatively difficult to understand and use" (Rogers, 2003, p.257) and is negatively related to its rate of adoption. Rule induction approaches are simple in the sense that they are non-parametric, and that they result in easily comprehended rules and rulesets. However, they may also appear complex to potential adopters in education research because, 1) there is lack of clarity on what they are, their relative advantages and disadvantages over other more commonly used analytical approaches, and their acceptability as a method for research, and 2) they require learning of new software and analytical frameworks.

Adding to these sources of complexity, there are multiple sub-categories and variations of rule induction, and multiple statistical software instantiating different subsets of these methods. Furthermore, results of rule induction approaches may appear more complicated than results of commonly used regression-based statistical methods. Instead of arriving at results that reject or confirm a null hypothesis, rule induction approaches result in a set of rules that describe the data,

and so far, there is little guidance on how such rules can and cannot support education research agendas. The typical use of such rules in business and marketing are practical—to make predictions about future customers or clients, and to target advertising or service offerings based on the general trends. Learning analytics and education data mining researchers tend to use data mining similarly, to identify ways to target educational content that is appropriate for individual students. Those interested in deepening the understanding of education outside the realm of computer-based learning may appreciate a simpler and more theory-oriented vision of how rule induction results can help them.

Trialability is "the degree to which an innovation may be experimented with on a limited basis" (Rogers, 2003, p.258). Innovations that are readily available for "test drives" are theorized to have a higher likelihood of adoption. Many rule induction algorithms are open-sourced and software are available for free or low-cost trials (e.g., *R*, *Weka*, *Orange*, *SPSS Modeler* student subscription). Even the latest standard version of SPSS includes four types of decision trees. However, those who are less familiar with programming and quantitative methods are likely to experience a large learning curve in both using and understanding the meaning of the results.

Observability refers to "the degree to which the results of an innovation are visible to others" (Rogers, 2003, p.258). The easier it is to see the innovation results—literally, and in the mind's eye—the greater its rate of adoption. Rule induction data mining has low observability because it is not physically instantiated—it is a set of methodologies, or ways of thinking-and-doing that can only be gained by the investment of time and attention. Even among methodologies, rule induction data mining is still low in observability since it has different assumptions and aims than inferential statistical methods that have been commonly used in education research.

Lack of clarity about relative advantage, questionable compatibility with values and aims of orthodox education research, seemingly high complexity, and low observability are current barriers to the adoption of rule induction methods. This study addresses several of these to help orthodox education researchers become able to make a better determination about whether to try the new methodology, and have a less difficult time adopting it if they so choose.

1.2 PROJECT OVERVIEW

I wanted to better understand how and to what extent rule induction methods could be useful to education researchers, from both theoretical and experiential standpoints. Thus, I conducted an extended literature review with a methodological focus (Chapter 2), conducted several rounds of rule induction data mining (Chapters 3 and 4), and reflected on the utility of the approach based on theory and experience (Chapter 5).

For the literature review, I examined what data mining is (Section 2.1), potential benefits and concerns of data mining vis-à-vis education research (Section 2.2), types of rule induction approaches (Section 2.3) methods to evaluate rules and rulesets (Section 2.4), and applications of rule induction in education research (Section 2.5). The literature was scattered across disciplines (machine leaning, education, statistics, sociology of science) with many claims still underdeveloped or underexplored, so threading them together turned out to be a rather formidable task. Interestingly, while I found several "orthodox" education studies applying rule induction, I did not see among them clear instances where rule induction provided insights that could not have been obtained by more traditional statistical approaches. I also saw that many of the studies had methodological shortcomings.

Thus, for the experiential component, I aimed to use sound methodology to identify some illustrative cases where rule induction data mining methods add value, beyond traditional statistical approaches. The purpose was *not* to generalize from such cases that rule induction is always (or even often) useful for education research, but rather to see whether it can so that I can use the experience to think more clearly think about the benefits and drawbacks of the approach. **My specific primary goals were to find illustrative example(s) in which rule induction methods, relative to regression approaches, (1) improve classification accuracy, and/or (2) offer new avenues of explanation through their unique ability to generate if-then rules about subgroups** (i.e., detect predictor-outcome relationships that could be unique to subgroups). My null hypothesis was that the rule induction results are not better in either of the two ways. **My secondary aim was methodological—to provide a practical and principled way to use rule induction in education research.**

To these ends, I reanalyzed two important regression studies that seemed, in theory, very likely to have gained from using rule induction. One study, by Byrnes and Miller (2007), explored the National Educational Longitudinal Study of the Eight Grade Class of 1988 (NELS:88) datasets (United States Department of Education National Center for Education Statistics, 1995, 1999, 2006a, 2006b), to better understand how opportunity, propensity and distal factors (represented by 37 variables), relate to STEM achievement in 10^{th} and 12^{th} grades. The other study, by Thomas (2006), explored NELS:88 to understand how student, family, peer, community and school factors (represented by 33 variables) relate to academic achievement of African American high school students. I re-analyzed their research questions using several rule induction methods in a principled way, to see whether the new methods, relative to the original regression approaches, improved

classification accuracy, and/or detected predictor-outcome relationships that were unique to some subgroups.

These questions and methodology were shaped by several background assumptions. First, I assumed there is inherent value in conducting exploratory quantitative analysis of data in contrast to some educational researchers and statisticians who have eschewed exploratory quantitative research as "data-dredging" or "data-fishing" (Section 2.2.2). I aligned myself with many contemporary researchers in believing that, 1) there is value to using such approaches, particularly when the dataset is very large in numbers of cases and predictors, and 2) that data mining and orthodox quantitative methods have complementary strengths (Grover & Mehra, 2008; Zhao & Luan, 2006). Attention had to be paid to reducing threats to statistical and ecological validity of inferences from data mining. Again, as discussed further in the next chapter, it was important to use multiple algorithms, cross-validate results and use alternative means to check the validity of any new inferences (e.g., check against literature). I also followed the recommendation to use a data mining framework (Azevedo, 2008) and articulate background assumptions when preparing the data, so to not blindly be searching for "anything that sticks."

My second background assumption was that a new methodology is useful to a field of research if it has the potential to provide information that cannot be gained, or cannot easily be gained, by existing approaches. Some researchers have used rule induction methods to investigate questions that could have been better answered by more widely used and known statistical approaches (especially regression). I assumed that a new methodology is useful only insofar as it adds methodological value to existing approaches.

Finally, I assumed that better understanding of subgroups is of high interest to many educational researchers, and therefore methodologies that help identify hitherto undiscovered

relationships that are potentially unique to some subgroups would be useful to educational researchers.

1.3 SIGNIFICANCE

I began this project believing that at the very least, it will be an example of why we should not be overly optimistic about rule induction, despite its theoretical promise, and demonstrate how to conduct rule induction using education datasets in a principled and cautious way. At best, I believed this project could be a solid example that illustrates how rule induction methods could be used and be useful in education; a project that could entice education researchers to use rule induction effectively, and more often, to improve education. As I dove deeper into the literature and began data analysis, I realized that my project also contributes to education research methodology by engaging more directly, comprehensively and deeply with the barriers to adoption of these methods than most current theoretical and applied research on rule induction.

My background section (Chapter 2) provides a comprehensive and critical exposition of benefits and concerns about the use of data mining in education, which should help readers form a balanced assessment of the compatibility of rule induction approaches with their values and favored methods for research (Section 2.2 and 2.5.1). It also discusses the evaluation of rules and rulesets in more detail, and with more relevance to applied education research than most expositions I have encountered (Section 2.4). Efforts were dedicated to placing standard machine learning validation approaches (namely, cross-validation and interestingness rules) into a greater context of theory validation, bringing in insights from a multi-faceted notion of validation used in the psychometrics. In addition, by being comprehensive and writing specifically for the education

research audience, the background and methodology sections should help reduce some of the perceived complexity associated with rule induction data mining.

My approach of re-analyzing important existing studies using rule induction and examining more than just predictive accuracy (Chapter 3), helps clarify the nature of any advantages this new approach may have over regression. I also incorporated hypothesis-driven, confirmatory elements to my data mining analysis, as it seemed to make it more compatible with approaches currently valued and needed in education research. This was done by (1) articulating in advance what I hoped to infer from rule induction, and evidence that would be needed to be reasonably confident in such inferences (Section 3.2.1), and (2) using a targeted, hypothesis-generation and confirmation approach to association rule mining (Section 3.2.4). My methodology also increases trialability and reduces perceived complexity of rule induction approaches, as it relies on a single, freely available statistical software (R), and streamlines the model-building and evaluation methods across algorithms. I share the codes used for the data mining (Appendices G and H), which mostly relies on programs created by others in the R-community, but also includes a subroutine I created to help evaluate results of the CBA algorithm.

The results section (Chapter 4) attempts to be comprehensive and accessible, with examples and visuals intended to reduce the perceived complexity and increase the observability of rule induction approaches and their potential benefits. Finally, the discussion section (Chapter 5) directly addresses major barriers to adoption of rule induction by expounding on its relative advantage and disadvantages, and providing concrete suggestions on how and why we may (and may not) want to incorporate these approaches into education research.

2.0 BACKGROUND

This chapter begins with an introduction to the main approaches and frameworks of data mining, and a review of the potential benefits and concerns for its use in education research (Sections 2.1 and 2.2). It then describes the main approaches to rule induction, and summarizes empirical studies comparing the approaches (Section 2.3). The section on rule and ruleset evaluation (2.4) describes methods to reduce the chance of mistaking artifacts for real patterns, and how to evaluate the extent to which rules and rulesets are valid. The chapter concludes with a review of the current applications of rule induction methods to orthodox education research, and a chapter summary (2.5 and 2.6).

2.1 INTRODUCTION TO DATA MINING

2.1.1 Definition and main approaches

Data mining is a process of systematically, and automatically or semi-automatically, uncovering patterns in data (Witten et al., 2011). It is typically conducted on very large datasets that would be difficult to exhaustively examine by relying solely on traditional hypothetico-deductive[2] methods. Some data mining approaches involve searching for patterns that are related to a particular outcome variable. These are called "supervised" search, since the outcome of interest can be construed as

[2] Method that involves the formulation and testing of a specific hypothesis that can be falsified by observations.

guiding, or supervising, the search process. Other data mining methods search for any regularity in the data, i.e., an "unsupervised" search. Data mining may uncover relationships that are unexpected and useful to the user, or generate new hypotheses about the variable relatedness.

That the discovered information is unexpected, and of value to the user, tend to be important criteria of success for data mining, such that Hand has defined data mining as "the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners" (1998, p. 112), or as "the science of finding unexpected, valuable, or interesting structures in large datasets" (2000, p. 442). What it means for a finding to be interesting, varies across users and the user domain, and could refer to notions such as:

> *Evidence*—the significance of a finding measured by a statistical criterion.
>
> *Redundancy*—similarity of a finding with respect to other findings and measures to what degree a finding follows from another one.
>
> *Usefulness*—relatedness of a finding to the user's goals.
>
> *Novelty*—deviation from prior knowledge of the user or system.
>
> *Simplicity*—syntactical complexity of the presentation of a finding.
>
> *Generality*—fraction of the population to which the finding refers.
>
> (Klösgen, 1996, p.252, referenced in Hand, 1998, p.115)

Major approaches for supervised data mining include stepwise or all-possible-subsets regression, discriminant analyses, decision trees, neural networks, and support vector machines. Approaches for unsupervised data mining include cluster analysis (e.g., k-nearest neighbor, hierarchical), and association rule-mining. These approaches are described in e.g., Witten et al. (2011), Provost and Fawcett (2013), and Michie et al. (1994).

2.1.2 Frameworks: KDD, CRISP-DM, SEMMA

While data mining is largely an exploratory approach, it can be considered to be part of a larger, systematic and goal-oriented process, with multiple feedback loops. Azevedo (2008) describes three such popular conceptualizations: Knowledge Discovery from Databases (KDD) process, Cross-Industry Standard Process for Data Mining (CRISP-DM), and the Sample, Explore, Modify, Model and Assess (SEMMA) process.

The widely accepted KDD process, introduced by Usama M Fayyad, Piatetsky-Shapiro, and Smyth (1996), considers data mining to be one of five stages of a knowledge discovery process. The first is selection of the dataset (e.g., the most relevant aspects of data among what is available in a large data warehouse) to which data mining is to be performed. An understanding of the problem domain and the data users' (in this case, the education researchers') aims and perspectives are critical to selecting the most appropriate data. The second is data pre-processing, whereby data are cleaned so to increase consistency and improve the capacity of the algorithm to discover patterns, if any exist. In the transformation stage, data may be transformed or reduced in dimension, again to improve the chances of pattern discovery. The data mining process involves application(s) of algorithm(s) to search for any patterns of interest that might exist in the data. The final stage is interpretation/evaluation of the mined patterns.

The CRISP-DM process (Chapman et al., 2000), conceives of data mining as part of a business market research cycle. Developed by a consortium that includes SPSS, DaimlerChrysler, and NCR Corporation, the 6-phase process begins with acquiring business understanding and data understanding to operationalize the business problem and plan a tractable methodology for using the data to better understand the problem. These are followed by what are typically considered as part of data mining, i.e., data preparation, modeling and evaluation phases, where data are cleaned,

and models are explored, selected and thoroughly evaluated on whether they adequately attain the desired business objectives. The final phase is model deployment, which typically requires the knowledge gained from the process be made accessible to the relevant company employees and customers.

As with the cases above, the SAS Institute recognizes the place of data mining within the business intelligence cycle, but also articulates distinct stages *within* the data mining process itself (SAS Institute, 1998). The SEMMA process, developed by the SAS Institute, identifies five distinct aspects within data mining, corresponding to the words used to create its acronym. The sampling phase involves selecting a subset of the data that is relevant to the objective of mining the data so that data can be explored and processed more effectively. This step is recommended, particularly when using the entire dataset would be inefficient, or when that would increase the possibility of uncovering spurious or otherwise useless relationships. Sampling can also help the data be more reflective of the population to which inferences are to be made, and thereby strengthen the generalizability of the results. In the exploration stage, data are searched for trends and relationships—both anticipated and unanticipated—so that one understands the nature of the data better, and gains insight about how to clean and model it. Data visualization, clustering, and factor analyses are commonly used at this stage. The next phase is to modify the data, which involves selecting variables and making necessary transformations to the data. The modeling phase is when algorithms are used to automatically or semi-automatically search for variables that reliably predict the outcome. One then must assess the model classification results to discern the extent to which they are reliable, valid and potentially useful given the use context.

2.2 DATA MINING IN EDUCATION RESEARCH: POTENTIAL BENEFITS AND CONCERNS

There has been some, albeit limited, scholarly discussion about the place of data mining in education research. This section reviews the main benefits and drawbacks that have been recently discussed. More emphasis is placed on explicating the drawbacks, since these are less commonly acknowledged but important to understand if one aims to use data mining in education.

I identified relevant articles by first searching for peer-reviewed works concerning "data mining" in ERIC database on August 6, 2015. ERIC was chosen because, sponsored by the US Department of Education, it is considered "the premier national bibliographic database of education literature" (University of Pittsburgh University Library System, 2015), and because it only includes references that relate to education. The search was restricted to works published between 2005 and 2015, identifying 137 academic journal articles and 1 ERIC document. Among them, 12 included substantive conceptual or theoretical discussions about the value of data mining as a methodology for education research. Key conceptual papers cited by these articles—some outside education—were also examined when appropriate. Rather than aiming to be exhaustive, this review aimed to identify a representative sample of voices in the field on the potential benefits and drawbacks of applications of data mining in education research.

The articles were generally optimistic about how data mining could contribute methodologically; a few seemed overly optimistic (AlShammari, Aldhafiri, & Al-Shammari, 2013; ElAtia, Ipperciel, & Hammad, 2012), and a few were critical (Gaševic, Dawson, & Siemens, 2015; Reimann, Markauskaite, & Bannert, 2014). There was general consensus on what it is and why it is used, and a shared sense of inevitability about the wide-spread use in education. Several compared and contrasted data mining to traditional statistics (Grover & Mehra, 2008; Zhao &

Luan, 2006), which turns out to be an important theoretical framework through which to understand the purported benefits and drawbacks of data mining. I begin first, however, by introducing the commonly acknowledged potential benefits of data mining to education.

2.2.1 Potential benefits of using data mining in education research

Most scholars were optimistic about the benefits data mining could confer to the field. An important reason for enthusiasm was that in theory, data mining could lead to deeper understandings of individual learners, which in turn can help improve their learning experiences (Papamitsiou & Economides, 2014). Data mining's capacity to identify patterns in very large datasets could be especially conducive to deepening our understanding of learning, which we know involves multiple and complex pathways (Martin & Sherin, 2013). This is all the more so, since educational datasets are becoming larger and more complex. Some have pointed out that given the increasing size of available educational datasets, we cannot afford *not* to mine data and use it for all its worth (Grover & Mehra, 2008).

The potential contribution of data mining can also be understood through its difference from traditional statistical methods. In contrast to traditional statistical approaches, which were designed to analyze small samples, data mining is designed to efficiently analyze very large datasets (Grover & Mehra, 2008). This allows data mining to provide "just-in-time" information (Luan & Zhao, 2006), and to detect *unexpectedly* useful information (ElAtia et al., 2012; Thuneberg & Hotulainen, 2006) . Data mining also requires fewer statistical assumptions, making it easier and more flexible to employ for analysis. Decision trees, for example, do not require the typical parametric assumptions of linearity, normality, and homogeneity of variance. In addition, being less hypothesis-driven, data mining allows one to examine data without heavy reliance on

theoretical frameworks. Explained in more detail below, this can benefit a field like education where theoretical frameworks are less strongly established (at least compared to the natural sciences) (Luan & Zhao, 2006).

Another unique benefit to data mining is that it can help analyze non-traditional forms of data in an efficient and effective way. Data mining can be applied to data on text, location, audio, images, interactions, and social relations (Grover & Mehra, 2008; Papamitsiou & Economides, 2014). This may help expand the analytic scope of traditionally qualitative sub-fields of education. Lang and Baehr (2012) used text-mining to better understand the relationship between writing composition instruction and student performance. By using data mining, they were able to analyze larger quantities of text data than what is typically analyzed in writing composition education research and have more confidence in their results.

The remaining sections review the cautions against blind use of data mining in education research. Concerns arise from considerations of traditional statistical principals, Sociology of Science, and from examinations of recent activities in learning analytics and educational data mining.

2.2.2 Concerns from the perspectives of traditional Statistics

> *Despite its obvious connection to statistics, data mining, which often employs "exotic" algorithms and seems to be operating mostly in a black box, has produced a fairly high level of discomfort in the statistical community. The major criticism of data mining centers on the lack of theory in the search for best predictions and, therefore, that too much power is given to the computer. This is directly contradictory to the traditional understanding of data analysis…* (Zhao & Luan, 2006, p. 8)

Data mining has been criticized in a number of ways, for having insufficient regard of traditional statistical theory. Hand (1998, 2000) and Zhao and Luan (2006) describe and address a series of

worries related to this point, by contrasting data mining with traditional statistical approaches. The essential exercise underlying traditional statistics is use of data to confirm a statement nested within a theoretical framework. It begins with a null hypothesis about a population based on some background theory and examines a random sample of that population to either reject or fail-to-reject the hypothesis. Variables to be included in a statistical model are also selected based on some background theory. Data mining, on the other hand, "shares a similar philosophical root" with exploratory data analysis being not as focused, or dependent, on theory confirmation (Zhao & Luan, 2006, p. 11). Its goal is typically to find immediately actionable information that accurately predicts behavior of the particular group of customers, students, or patients, with whom the company or institution must deal in the imminent future, *rather than* providing the best possible theoretical explanation of a complex social phenomena. As such, data mining does not necessitate that we have a well-defined background theory against which we select our model and interpret our results: although as Zhao and Luan (2006) take care to emphasize, data mining still requires a great deal of sound human input. As the leading data mining frameworks (Chapman et al., 2000; Usama M Fayyad et al., 1996; SAS Institute, 1998) make explicit, the researcher's understanding of the research context and dataset are critical to effective data mining. However, "compared with [traditional] statistics, data mining is less confined in presumptions about the relations among variables," and therefore it "[leaves] ample space for discoveries that might not occur otherwise" (Zhao & Luan, 2006, p. 11).

This difference in the role of theory underlies the concerns about data mining raised from the perspective of traditional Statistics (Grover & Mehra, 2008; Zhao & Luan, 2006). Data mining activities are typically not well grounded in prior research, and therefore have less to contribute in terms of theory confirmation, or explanation. They often do not assume a sampling theory so

cannot make convincing statistical generalizations about the population. Without reliance on background and sampling theories, there is no hypothesis testing, or significance values (often construed as "statistical rigor") to be attached to results. Finally, data mining may inflate the possibility of erroneously concluding that a finding is significant or important (inflation of Type I error). Such an error can be made either because the data miner has very little theoretical grounding so does not know what is or is not significant with respect to what is already known. It can also happen if the data miner repeatedly explores the same data, using different methods or conditions, which would increase the possibility of assessing a spurious relationship to be true or important.

However, as Zhao and Luan (2006) explain, the aforementioned limitations of data mining are not necessarily devastating. It is again the differences in the aims and approaches between data mining and traditional statistics, which help illuminate why. First, while theory can help guide what we observe and give us a level of comfort that we are examining important things that actually exist, it can also blind us in seeing what is important, or even guide us in the wrong direction. John Tukey made the analogy of a data analyst as a detective "open to a wide range of ideas, possibilities, and idiosyncrasies," and a (traditional) statistician as a judge "examining and testing clearly identified hypotheses" (Tukey, 1962, summarized by Zhao & Luan, 2006, p.11). To build on Tukey's analogy, detectives with strong preconceived notions about how criminals think and act, can miss important clues that don't align with their preconceptions, or weigh too heavily the evidence that strongly supports her/his particular viewpoint and fail to resolve a case. In social scientific research too, it is not always prudent to have too many assumptions about what exists and how things work. When it comes to understanding a phenomenon where good background theories are lacking, the atheoretical nature of data mining can prove to be a strength, rather than a weakness.

That data mining is not based on sampling theory is also not particularly concerning if data mining is used primarily to build specific models reflecting local conditions, rather than to build theories that apply more globally. When companies and institutions mine data, their purpose is typically to predict information about their own clients, and guide near-future decision making. Such organizations generally do not care whether that information is true more generally, across their entire industry, and therefore have no need to be taking random samples of companies within their industry. Zhao and Luan (2006) add that generating a useful global, rather than specific model is "an ambitious and even unrealistic task" (p. 12). They remark:

> A model is a simplification of reality, and a global model excludes low-level details, focusing only on a high level of abstraction that summarizes the data structure because it assumes homogeneity within the population. A globally generalizable model usually contains less detailed information than a specific model. But reality is extremely complicated, especially for social sciences, and fraught with difficulties and ambiguities stemming from deficiencies in measurement, design, and analysis. (Zhao & Luan, 2006, p. 12)

They continue that this general and quite crude nature of traditional statistical models explains the low threshold of acceptability of statistical models, and why it is not uncommon for social scientists to present results that explain less than 20 percent of the variance of the dependent variable. The contrast between data mining and traditional statistics then, is not simply that the latter attains more generalizable knowledge. Rather, it can be considered a tradeoff where, "typical statistical regression model uses a few variables to generalize to an entire population, [while] data mining provides the potential to take advantage of information at a more detailed and specific level" (Zhao & Luan, 2006, p. 12).

There is another way to think about the role of sampling in the data mining context: That as long as we have information and computing power necessary to analyze the entire population, there simply is no need for it. Traditional statistics was developed, in large part, as a pragmatic and economical means to understanding an entire phenomenon—it provided justification for making claims about a phenomenon, even if one looked only at a very small piece of it. Well into the 1970s much of statistical analyses were conducted by hand (Zhao & Luan, 2006), which meant there was a serious limitation to how much information one can reasonably consider in an analysis. Data collection and storage were expensive, especially before the use of electronic databases and online communication became routine, prohibiting analyses of rich population data. Over the past several decades, rich population information has become increasingly available, as has computing power to efficiently analyze such "big data." School districts, institutes of higher education, state and local education, health and social services departments, and criminal justice systems, now often have electronic records of every person who has been part of their system. Many research questions that may have required sampling, and associated statistical considerations to correctly account for it, no longer require sampling because data are available for the population. That data mining does not require adherence to a sampling theory, then, is not a serious concern *as long as* data is mined from all or most of the population that one hopes to understand.

The concern about the lack of statistical significance values attached to data mining results is a variant of the sampling concern. Statistical significance is a measure of uncertainty associated with sampling error. In some instances, there is no need to assess the possibility that the results are due to sampling error, e.g., when we: (i) have information on the entire population, (ii) have a large sample that quite adequately represents the population, (iii) have a large enough sample that almost any difference turns out to be statistically significant, and/or (iv) have no interest in generalizing

conclusions far beyond the particular sample at hand. However, if the above conditions are not met—i.e., if we want to generalize conclusions far beyond a small, potentially unrepresentative sample—trusting the data mining results wholesale, without regard to the possibility of sampling error, would be problematic.

The most serious concern about data mining from the perspective of traditional Statistics is the inflation of Type I error due to data dredging. As Hand (1998) describes:

> [Data mining] has a derogatory connotation because a sufficiently exhaustive search will certainly throw up patterns of some kind—by definition data that are not simply uniform have differences which can be interpreted as patterns. The trouble is that many of these "patterns" will simply be a product of random fluctuations, and will not represent any underlying structure. ... To statisticians, then, the term data mining conveys the sense of naïve hope vainly struggling against the cold realities of chance. (Hand, 1998)

The possibility of model over-fit and Type I error is increased when data mining is used to build precise models for local use (rather than less precise models for global understanding). Cross-validation of the results within and/or across datasets and across algorithms are essential to data mining, as are checking the feasibility of the model with domain experts (Luan & Zhao, 2006; Provost & Fawcett, 2013; Witten et al., 2011). Restricting model specificity during the model creation stage (e.g., using stopping rules or pruning when creating decision trees) are also ways in which model overfitting have and should be addressed.

To summarize, from the lens of traditional Statistics, data mining appears concerning because of its atheoretical nature of inquiry, non-reliance on sampling theory and increased possibility of Type I error. The concerns are not insurmountable, yet need to be understood and taken into consideration when designing studies involving data mining.

2.2.3 Concerns from Sociology of Science

> *All told, the generation, accumulation, processing and analysis of digital data is now being touted as a potential panacea for many current educational challenges and problems.* (Selwyn, 2015, p. 67)

A concern from Sociology of Science is that data mining contributes to a tunnel-versioned focus and regard of education data that has serious repercussions. In a discussion of the significance, merit and demerits of data mining and data-driven approaches in education, Selwyn (2015) raises concerns from a sociological (and the newly emerging "digital sociological"[3]) perspectives, regarding the "datafication" of education, or the increased data-reliance in our designs and understandings of education. Several of the concerns pertain directly to the topic of usefulness of data mining to education research. The first is that increased data-reliance of education may cause people to regard complex social and educational problems merely as complex but solvable *statistical* problems. Focusing too much on available data may prevent education researchers from considering important and relevant nuances, contextual factors, causal factors, and counter-narratives. Selwyn describes:

> The recording of social 'facts' into digital data, therefore, implies that some qualities and characteristics will be made better known than others. For example, as Ruppert (2012) notes, the core sociological constructs of race, social class, gender, sexuality and so on, do not translate easily into data categories, despite their constant use within data collection and analysis. Often digital data can be said to support little more than 'surface'

[3] This emerging subfield of sociology, and sociology of technology, tends to begin with the assumption that data is political, value-laden and power-conferring in nature, rather than objective, neutral and unproblematic. It also typically plays close attention to how data shapes and are shaped by social interests. Sewlyn (2014, pp.68-9) mentions Evelyn Ruppert (Open University), Kate Crawford (MIT), Susan Halford and colleagues (Southampton University), and others (Mike Savage, David Beer, Andrew Webster, Roger Burrows, Deborah Lupton, Theresa Sauter, Rob Kitchin, Mark Graham, Lev Manovich and Matthew Fuller) as key scholars shaping this field.

> understandings of sociological entities (Savage, 2009). … Much of the depth that is lacking from digital data could be argued to include issues of historical context and connections with past events, individualist and humanist accounts of the social, and an underpinning sense of moral knowledge (see Barnes, 2013; Ruppert, 2013). (Selwyn, 2015, p. 75)

Along the same lines, increased interest in data mining could consciously or subconsciously lure the minds of education researchers towards an unhealthy reductionism: researchers may begin to regard the realm of teaching and learning primarily in terms of easily operationalized attributes for practicality or other reasons. Worries about unhealthy reductionism and brute operationalization of complex constructs are not unique to data mining. However, the increased volume, variety and velocity of data processing (the classic descriptors of "big data," per Laney (2001)) increases attention and reliance on data-driven approaches, and therefore increases the magnitude of this concern. Important factors related to learning such as social interactions, agency, perception, attitudes, race, gender, historical context, cultural beliefs, are difficult to operationalize, and quality data will always be difficult and time-consuming to collect. As Selwyn (2015) and Manovich (2012) note, we do not want to neglect studies on "deep data" on just a few cases by focusing too much on "surface data" about many cases.

In addition, data mining raises concerns about differential power dynamics among those who analyze and are analyzed, and those who can and cannot analyze. Selwyn (2015), drawing from Lupton (2013), Manovich (2012), and Ruppert (2012), suspects that data, and the ability to use data, is a form of power that has the potential to be distributed inequitably and misused. It is conceivable that machine learning specialists involved in educational data mining come to obtain a disproportionate amount of power in deciding what happens in education (even if they are not familiar with many aspects of the field), just because of their technical knowledge of how to

manipulate large computer-based educational datasets. Governments, education policy makers, school districts, researchers and companies may provide machine learning specialists with more funding, attention and voice than is ultimately good for our teachers and students. Data open-access and privacy are related concerns for education researchers as they further explore the realm of big data in education (ElAtia et al., 2012). Open access would protect from too much data concentrating in the hands of the few, while privacy provides some protection towards those who are analyzed from those with the power to do the analysis against the interests of those who are analyzed.

2.2.4 Concerns from learning analytics and educational data mining

Related concerns about data mining have been raised from direct experience or familiarity with current practices of data mining in education. Educational data mining (EDM) and learning analytics are emerging and overlapping interdisciplinary fields, which involve harnessing knowledge from large educational datasets. Relatively speaking, EDM is more interested in finding new patterns, and/or developing new algorithms to discover new patterns, while learning analytics is more interested in finding applications of the patterns to improve teaching and learning (Bienkowski, Feng, & Means, 2012).

Upon reflecting on how EDM and related e-research methods have analyzed self-regulated learning, Reimann et al. (2014) noticed that many studies tend to assume a “flat ontology” that relies too heavily and assumes too much about simple user behaviors such as clicking, logging in, moving their eyes, typing and uttering. For example, while their previous study had found that “reading” for successful students was more strongly associated with “monitoring” and “elaboration” than with “repeating,” their models lacked explanatory power across contexts and

different student dispositions because their theoretical framework was ontologically impoverished. Reiman et al.'s general cautionary point was that "big data" and "more data" are not identical with conceptually rich data and deep data. They suggested enriching the EDM research ontology to include social structures and a range of cognitive and non-cognitive processes, which are beyond physical observable behaviors such as clicking and typing. This would also involve collecting richer data, which could involve data from multiple sources collected in a variety of ways, and analyzing them in a way that respects ontological complexity (they suggest system dynamics and agent-based modeling).

Martin and Sherin (2013) had raised similar concerns in their introduction to a special issue on learning analytics of the *Journal of Learning Sciences*. Their assessment of the EDM and learning analytics field was cautiously optimistic based on potential utility of these methods, rather than actual results to date:

> Although the educational data mining and [learning analytics] communities have produced a steady stream of interesting results, work in education has far to go in order to reap the benefits for student learning… (p.511-2).

Their discussion on the potential of learning analytics to learning science researchers, while on-the-whole positive, cautioned that there is increased temptation to conduct research on topics where big data are easy to collect: While learning analytics can be conducted on traditional data, "when we apply [learning analytics], we are more likely to restrict our study to learning activities that are conducted using computers" (p.515). Like Reimann et al. (2014), they urged learning analytics researchers to look beyond mouse clicks and key presses, to continue to research learning in a broad range of settings, and to make sure to allow the research questions to guide the methodology rather than the other way around.

Making progress in learning analytics has been difficult also because of its interdisciplinary nature (Gaševic et al., 2015). Consider, for example, an initiative to improve academic success by providing students with timely automated feedback about their coursework. For such an initiative to show impact there needs to be at minimum good analytics, a user-friendly implementation platform, and high-quality feedback aligned with the most current knowledge from Learning Sciences. The success of learning analytics depends upon substantive collaboration among machine learning scientists, education practitioners and educational researchers, making such initiatives riskier and more expensive.

A final concern about data mining raised among those in EDM and learning analytics pertains to unintended negative consequences to students. Corrin and de Barba (2014) found that high achieving students tended to underperform, relative to how they usually perform in a class, when their course data dashboard informed them of where they stood relative to the class mean. Along the same lines, learning analytics researchers have worried that constant reminders about poor performance may cause undue distress to students, and/or diminish the quality of teaching and learning to become narrowly focused on merely improving superficial metrics (Gaševic et al., 2015). Of course, conducting data mining in educational research *per se* is unlikely to be a direct cause of such consequences. However, just as educational and psychological assessment developers must carefully consider the unintended negative consequences of the instruments they develop (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Linn, 1997; Messick, 1975), quantitative education researchers more generally should take care to minimize negative implications of their research.

2.2.5 Implications

There is great optimism and momentum for applying data mining to investigate the nature of learning and education. The ability to analyze a large amount of data quickly and at once provides the possibility to find hitherto undiscovered relationships among teaching and learning variables that are useful or important. Data mining allows researchers to analyze visual, audio and text data without extensive recoding.

The concerns raised so far about data mining are not devastating, but provide guidance to those who hope to use it for research. Researchers should be principled in their use of data mining. It is possible to mine data with hardly any knowledge of the domain from which the data come—however, such reckless application is likely to be a hindrance to the field. While it is neither necessary nor always desirable for data miners to take a rigorous hypothesis-driven approach, the methodology and interpretation of results should be well-informed by what is known (or thought to be known) in the field. Data mining can be used for prediction, theory development, or hypothesis generation: The specific objective should determine the method, rather than conversely. Special attention should be paid to sampling, over-fit avoidance, and predictor set completeness.

Like any tool, the utility of data mining depends largely on the skill and imagination of the user. And like any tool, it may be used for a variety of goals and purposes. The verdict is still out on how useful data mining could be to educational research: even in learning analytics and educational data mining, convincing applications of data mining are still rare. As educational researchers explore the utility of data mining in their research, they should keep a balanced perspective, inform others even of null-results and unintended downstream consequences, and be vigilant in pursuing questions with answers worth knowing.

2.3 RULE INDUCTION APPROACHES IN DATA MINING

Rule induction methods are a particular class of data mining methods that involve induction of conditional (*if-then*) statements for purposes of classifying or describing observations (Tung, 2009). The antecedent (the '*if*' part), and consequent (the '*then*' part), are each a conjunctive statement about values of individual variables, or "attribute-value pairs." An education example of a rule is: "*if* middle school GPA > 3.5 *and* school = parochial, *then* student is likely to graduate (rather than drop-out)." As explained in more detail below, such rules can be detected from a dataset in virtue of its ubiquity and accuracy, among other measures.

The major advantage of rule induction among other data mining methods is that the results are descriptive and easily understood even by non-specialists. The non-parametric nature of rule-based approaches not only contributes to its ease of comprehension, but also increases their flexibility as an analytic tool. Rule-based approaches share fundamental commonalities with how some philosophers and cognitive scientists have conceptualized cognition (Sloman, 1996), which may explain part of its intuitive appeal as a modeling and knowledge discovery approach.

Three standard ways to generate (induce) rules from a dataset are sequential covering, decision trees and association rule mining. I describe each of these three types of rule induction approaches and review empirical research comparing these approaches. I also discuss ensemble approaches to rule induction, which are methods aimed to improve predictive accuracy by generating multiple trees through resampling the dataset and combining their results.

2.3.1 Sequential covering

First instantiated in the AQ algorithm by Michalski (1969), sequential covering is one of the most studied and used supervised rule induction approach in machine learning. Covering algorithms generate a predictive ruleset by discovering a highly predictive rule, separating the cases that were covered by that rule, and repeating the process until all or most of the cases are covered by some rule. Some refer to this approach as "separate-and-conquer learning."

A sequential covering algorithm typically comprises of four subroutines (Fürnkranz, Gamberger, & Lavrač, 2012). A **feature construction subroutine** generates viable "features," or atomic statements that can be selected as a part of the rule, based on the training data. For categorical variables, features have the form $\mathbf{A}_i = v_{i,j}$, where $\mathbf{A}_i$ is the i^{th} attribute (variable) of the dataset, and $v_{i,j}$ is the j^{th} value (level) of that attribute. For scale variables, features are expressed as inequalities (e.g., $\mathbf{A}_i > v$).

A **rule-learning subroutine** uses the available set of cases to search for a single "best rule" that correctly classifies the positive examples as positive, without misclassifying the negative examples. Most algorithms search for such a "best rule" using a general-to-specific, or top-down approach, where they specialize a general rule by adding features to the antecedent until there is no more increase in the rule quality (e.g., the rule no longer incorrectly misclassifies negative examples). Beam search is often used to improve the chance that the rule is globally optimal—it is a process whereby the top w best enhancements to the rule (where $w > 1$) are pursued in each iteration, instead of just one.

A **ruleset-learning subroutine** calls on the rule-learning subroutine to create a rule that best characterizes the uncovered cases, accepts rules that meet some adequacy criteria, and reduces the set of uncovered cases. The three subroutines described so far together identify a set of rules

that predicts one particular class (or level) of outcome. Finally, the **base subroutine** for rule-learning iteratively repeats this search for rulesets for each level of outcome resulting in a set of rules that predict the dataset.

Rulesets induced with covering algorithms can be ordered or unordered. Ordered rulesets are also called "decision lists" and have the structure of a branch with short twigs shooting off to the side (the twigs do not branch any further). After the first rule, the rules in a decision list are conditional upon all the previous rule(s) not being applied. Rules in unordered rulesets are not necessarily related to one another in such a way, and can result in some test cases receiving contradictory predictions by two or more rules. Usually these cases are resolved by some type of voting algorithm. Elaborations on the covering approach can be found in Witten et al. (2011), Fürnkranz (1999) and Fürnkranz et al. (2012).

Examples of covering algorithms include AQ (Hong, Mozetic, & Michalski, 1986; Michalski, 1969; Michalski & Kaufman, 2001), CN2 (Clark & Niblett, 1989), and RIPPER (Cohen, 1995). These algorithms differ primarily in the language they can use, the way in which they search for the "best" rules for the data, and the way in which they attempt to prevent rules from overfitting the data (Fürnkranz, 1999).

The earliest version of the AQ algorithm (Michalski, 1969) searched for ways to refine rules by considering differences between randomly drawn positive and negative examples in the dataset. Later algorithms, starting with PRISM (Cendrowska, 1987), have involved an exhaustive consideration of cases and features at each iteration. CN2, by Clark, Niblett and colleagues (Clark & Boswell, 1991; Clark & Niblett, 1989), incorporated a number of key innovations to reduce the chances of the rules overfitting the training dataset, including pre-pruning of insignificant rules based on likelihood ratio. In addition, CN2 developers introduced methods to handle outcomes

with more than two levels by voting using a decision list (Clark & Niblett, 1989) and through the use of an unordered approach where one level of outcome is pitted against all other levels for each round (Clark & Boswell, 1991).

The key innovation in RIPPER (Repeated Incremental Pruning to Produce Error Reduction; Cohen, 1995) was its ability to significantly reduce overfitting through post-processing, where earlier rules are checked on whether they could still be learned based on rules discovered later. OPUS (Optimized Pruning for Unordered Search) by Webb (1995) was the first rule-learning algorithm that exhaustively searched for rule sets. OPUS explores the search space efficiently by using an ordered search, and skipping explorations of a search space where more general versions of the rule are no good. FOIL (First-Order Inductive Learner) by Quinlan (1990), was the first learner that allowed rules to be discovered in first order logic, rather than propositional logic.

2.3.2 Decision trees

Decision tree rule induction involves taking a sample whose outcomes are known, and dividing them according to some shared predictor(s) so that those with who differ in the outcome are maximally separated into the newly created subsamples. For example, if we hope to build a decision tree that uses 9th grade information to predict whether a student graduates from high school, the first step is to get 9th grade information of a representative sample of graduates and non-graduates, and to identify which characteristic from their 9th grade year would best separate them into graduates vs non-graduates. The 9th graders are split into subgroups based on the "best attribute-value pair for the split," and the process is repeated for each of the newly created subgroups until there is no more predictors that differentiate the classes, and/or some other stopping criteria is met.

Tree approaches are also called “recursive partitioning” or “divide-and-conquer,” since it involves splitting the dataset into smaller and smaller subsets, with each split causing the subset to be more homogeneous in the outcome attribute. The rules generated from decision trees are mutually exclusive and collectively exhaustive in accounting for the predictor space—i.e., any case characterized by the same set of predictors will have an outcome designation as specified by exactly one rule. Detailed overviews of decision tree methodologies are provided by Hand (1997), Murthy (1998), Rokach and Maimon (2014), Michie et al. (1994) and Strobl, Malley, and Tutz (2009).

The way in which the splitting variable is chosen among all the available predictors, i.e., how exactly the tree is grown, is one of the largest factors that distinguishes one tree algorithm from another. Chi-squared automatic interaction detection, or CHAID, which is an older tree algorithm (Kass, 1980), compares all possible two-way chi-square values that can be created from the dependent and independent variables using a Bonferroni adjustment. If the independent variable has more than three levels, then the 2-by-*k* sub-tables are examined, where *k* is the number of levels of the dependent variable. For ordinal and continuous variables, only sub-tables consisting of neighboring levels of the independent variable are examined, while for nominal variables all possible sub-tables are examined. The sub-table with the lowest chi-square value is examined first, and if this is not significant the two levels of the independent variables are combined to form a single composite category. This process of combining categories is repeated until all the pairs of categories are significantly different on the dependent variable, or until there are two or fewer categories. The variable for which the probability is lowest (and chi-square value is highest), is chosen for the split.

CART (Breiman et al., 1984) has a number of options for creating splits, the most popular uses the Gini diversity index, which is a measure of how homogeneous the resulting subgroups ("child nodes") are in terms of the outcome. The index is a measure of the sum of squares of the proportions of the outcome levels within a node, calculated as Gini(n) = $1 - \sum(p_{j|n}{}^2)$, where $p_{j|n}$ is the probability of class j in the node n, and the summation is over classes. When the outcome consists of two classes, the node would be most impure (heterogeneous) when it consists of a 50-50 mixture of each of the classes, i.e., when Gini(n) = $1 - (.5^2 + .5^2) = .5$, and the node would be most pure (homogeneous) when it consists of 100-0 or 0-100 mixture of each of the classes, i.e., when Gini(n) = $1 - (0^2 + 1^2) = 0$. For each possible split, CART calculates a weighted sum of the Gini index of the child nodes, and selects the split that leads to the greatest reduction in impurity between the original ("parent") and child nodes. The Gini index can be calculated for both categorical and continuous outcomes (Michie et al., 1994, Chapter 5). C4.5 (Quinlan, 1993) and C5.0 (Quinlan, 2013) also base their splits on impurity reduction, but uses the Shannon entropy measure instead of the Gini index.

QUEST (Loh & Shih, 1997) tries to eliminate search algorithms' biases towards favoring variables with more levels (e.g., continuous variables) by incorporating a Bonferroni correction, and by decoupling the variable selection from the split selection. First, QUEST determines where the binary split should occur for each variable using quadratic discriminant analysis for ordered variables, or by applying a CRIMCOORD transformation (Gnanadesikan, 1977) and conducting χ^2 tests if the variable is non-ordered. To select the splitting variable, QUEST compares the p-values of a χ^2 test and F-test (with Bonferroni adjustment). If p-values are greater than a pre-desired threshold, Levine's F-test for unequal variance is further conducted on ordered variables to see whether this yields better results than a Bonferroni adjustment.

In addition to a subroutine for "growing", some algorithms (e.g., CART, C4.5, C5.0) also have a "pruning" subroutine, which eliminates small nodes that are likely to be overfitting the dataset. At least for some datasets, pruning an overgrown tree has been shown to produce better results than relying solely on a stopping criteria to prevent model overfit (Breiman et al., 1984). Many heuristics have been proposed for pruning, although empirical comparisons have generally concluded that no one pruning approach is superior over others (Rokach & Maimon, 2005). Similarly, empirical comparisons of different tree growing schemes have tended to show that performance across major algorithms are comparable, and when there is difference, what works relatively better/worse tends to depend on the nature of the dataset (Michie et al., 1994; Rokach & Maimon, 2005). In machine learning, the absence of a universally superior algorithm is considered a truism, and is referred to as the "no free lunch" theorem (Wolpert, 2012; Wolpert & Macready, 1997).

2.3.3 Association rule mining

Association rule mining is another popular rule induction approach, designed to identify good *rules* rather than good rulesets. Association rule algorithms scans the dataset to detect *any and all* conditional statements that apply fairly accurately to a substantive portion of the data, and/or meet some user-specified criteria. Unlike decision trees, association rule algorithms can perform an unsupervised search: the user need not specify a particular outcome variable, and the resulting rules are generally neither mutually exclusive nor collectively exhaustive in describing the predictor space. Since association rule mining generates all possible viable rules given the dataset, it tends to produce a greater number of rules relative to covering and tree approaches, and increase the possibility of detecting rules that are of interest to the user. However, it tends to also produce

a large number of obvious or otherwise uninteresting rules, which may be difficult for users to comb through.

Association rule mining approach can be used to generate predictive rulesets, for example by constraining the search for rules that have a particular outcome, and conducting a top-down, sequential covering search to arrive at a set of ordered or unordered rules that best describe the data. CBA (B. Liu, Hsu, & Ma, 1998; B. Liu, Ma, & Wong, 2000) is the best-known algorithm that takes such an approach, although there are quite a few others as surveyed by Bringmann, Nijssen, and Zimmermann (2009).

Because it is generally resource prohibitive to consider the viability of every logically possible rule that could characterize a dataset, and because not all associations among variables are interesting (in fact, most tend to be uninteresting), various attempts have been made to *efficiently* generate *only the interesting* rules. The Apriori algorithm (Agrawal, Imieliński, & Swami, 1993) is the oldest and most commonly used association rule algorithm, originating in market basket analysis. It first identifies attribute-value pair sets that occur frequently within the dataset, excluding any set that is a subset of another. Apriori efficiently identifies such sets by capitalizing on the notion that supersets of infrequent attribute-value sets cannot be frequent. It then considers all possible non-tautological rules that can be created using members of the set, screening for accuracy. The user specifies the minimum threshold for rule frequency (support) and accuracy (confidence).

As summarized by Goethals (2002), alternative approaches to improving the search efficiency include segmenting or sampling from the dataset to conduct searches (e.g., DIC by Brin, Motwani, Ullman, and Tsur (1997), CARMA by Hidber (1999), and a set of sampling algorithms by Toivonen (1996)), and generating associations between item sets and transaction IDs from

which to calculate frequent item sets (e.g., Eclat by Zaki (2000)). A large amount of research has also been done on ways to automatically or semi-automatically screen the rules for how interesting or relevant they might be to the user, as reviewed by Hilderman and Hamilton (1999), McGarry (2005), and Geng and Hamilton (2006).

2.3.4 Empirical research comparing rule induction approaches

There are some broad differences between covering, trees and association rule mining (Fürnkranz et al., 2012; Witten et al., 2011). If the goal is to search for good *rulesets* for the purposes of classification and prediction, sequential covering tends to be faster, and results in a simpler set of rules than trees. Ordered sequential covering approaches are more efficient and easier to apply to new cases than unordered sequential covering or tree approaches. However, when there is order dependence among rules, the validity of rules also become interdependent which could be problematic. Finally, ruleset builders based on efficient exhaustive searches (e.g., CBA) have the advantage over most commonly used tree and sequential covering methods by ensuring that the final model is globally optimal. However, as we shall see, globally optimal rulesets are not always better than locally optimal ones.

If the goal is to find good individual *rules*, association rule mining is advantageous in its ability to efficiently find all possible rules. However, this approach may produce too many rules for a user to effectively examine. In addition, sometimes the objective is not just to find rules that are predictive, but also to find rules that relate to other rules in some meaningful way. In such cases, ruleset induction approaches may be more beneficial than approaches that generate individual rules.

In the following subsections, I introduce two types of empirical studies examining the difference in efficacy of different ruleset induction approaches: (1) studies that compare accuracy and processing speed of trees and sequential covering, and (2) studies that examine whether exhaustive searches for rulesets lead to improved results.

2.3.4.1 Tree vs covering on accuracy and speed

Boström (1995): Trees may be more accurate and efficient than covering

Boström (1995) compared the accuracy and efficiency of tree and sequential covering approaches, showing that trees can more efficiently detect more accurate rulesets. Both algorithms relied on the information gain heuristic, and were implemented using first-order logic. Approaches were compared on three existing datasets—chess, tic-tac-toe, and on natural language parsing—each with dichotomous outcomes with an approximately even split, with 950 to 3200 cases. Holdout validation was used; the size of the training set was varied among 50, 25, 10, 5, and 1 percent, with each experiment repeated 50 times. Mean accuracy on the test set, CPU time, and "amount of algorithmic work" were compared across approaches, where the latter was operationalized as the number of times the algorithm checked whether a statement covered a case.

For the chess dataset, the tree approach was one to four percentage points more accurate than the covering, across all sizes of the training set. When half the data were used for training, the covering approach did over three times algorithmic work, and took over three times longer. For the untransformed natural language dataset, both covering and tree approaches had the same accuracy (relatively low, about 65%). Interestingly, however, when the variable set were enriched by including transformations of variables based on background knowledge of the domain, there was a large increase in accuracy (to about 70-97%, across training set size), with the tree model

performing three to ten percentage points better. Here again, the covering approach performed more algorithmic work than tree. In the tic-tac-toe dataset, the tree approach was consistently more accurate and efficient than covering, and generally by over ten percentage points. As with the natural language dataset, it suggested that inclusion of additional variables makes a large positive difference in the predictions, particularly when the size of the training sets range between 5 to 25%.

Cohen (1995): Covering can be more accurate and efficient than trees

Using a similar design, Cohen (1995) compared processing speed and accuracy of two covering approaches (IREP and RIPPER) and tree-based ruleset learning approach (C4.5rules). C4.5rules (Quinlan, 1993) generates rulesets iteratively greedily adding or subtracting single rules to a ruleset derived initially by an unpruned C4.5 decision tree. IREP (Fürnkranz & Widmer, 1994) was a then cutting-edge covering approach that had integrated a pruning mechanism to improve accuracy. Cohen had developed RIPPER as an improvement to IREP—by adding a rule optimization process that constructs and considers reasonable alternatives to rules generated by IREP (which can be repeated multiple times), applying a more liberal stopping criteria that reduced premature halting, and adding the ability to handle datasets with missing, numeric and categorical outcomes with three or more classes.

Accuracy and speed were compared on 37 diverse benchmark datasets (i.e., publicly accessible datasets used repeatedly across machine learning studies to evaluate new algorithms), many of which were from the UCI Machine Learning Repository. While IREP was much faster than C4.5rules, it had a higher error rate on 23 out of 37 datasets, with error rates approximately 13% higher than C4.5rules across datasets. RIPPER performed better than C4.5rules in 20 out of 37 data sets (tying in 2, and performing worse in 15), and had very comparable error rates to it

(about 1% higher across all datasets). When the rule optimization algorithm was repeated a second time (RIPPER2), the results against C4.5 rules seemed to improve slightly more. RIPPER was also much more efficient in handling large and noisy datasets than C4.5rules.

Section summary: No free lunch

Boström's (1995) study was limited in the variety of datasets and algorithms used. It suggested that the tree approach may be faster and more efficient than the covering approach, but seemed to more strongly support the no-free-lunch theorem. In contrast to Boström's general conclusion, Cohen's (1995) study showed that a carefully designed covering algorithm can perform comparably or better in both accuracy and efficiency to a standard tree-based algorithm. However, for some datasets trees did better, again supporting the no-free-lunch theorem. Rather typical to most such comparative studies, predictive accuracy and efficiency were the only measures used to assess the quality of rulesets.

2.3.4.2 Whether extensive searches find better rulesets

Quinlan and Cameron-Jones (1995): Exhaustive search may decrease accuracy

Quinlan and Cameron-Jones (1995) investigated whether more exhaustive searches lead to better rules. In their first sub-study, the authors generated a single conjunctive rule using a beam search with beam width that increased exponentially from 1 to 512. These approaches were tested on 12 small UCI datasets (about 100-700 cases each), with each type of search repeated 500 times. Rules were created using 50% of the dataset, and tested on the remaining cases. More exhaustive searches identified more accurate rule in just one dataset. In half of the datasets, greedy search was the most

accurate. In the remaining datasets, a beam width between 4 and 32 identified the most accurate rule.

In their second sub-study, the authors explored how the extensiveness of search related to the accuracy of *rulesets* of a covering algorithm. Using a similar design to the first sub-study, they tested three search conditions: greedy search; search with a beam width of 512 (approximating an exhaustive search); and "layered search" that widened the beam width to a roughly approximated optimal width (determined empirically based on the error, number of trials and number of rules examined in the training set across varying beam widths). The three methods were compared in accuracy, CPU time, and the total number of clauses in the ruleset.

Relative to the greedy search, the layered search was more accurate in their predictions in five datasets, and worse in three. Relative to the exhaustive search, the layered search was more accurate in six datasets and worse in just one. However, the error rates across the methods differed only by at most three percentage points, suggesting that the practical difference between the approaches are small. Layered search was approximately ten times slower than the greedy search, and 17 times faster than the exhaustive search. The theory size decreased by about 17% between the greedy and the layered search, but was essentially the same between the layered and extensive search.

The authors concluded that their results support their hypothesis that more searching can lead to worse rules and rulesets, by increasing the probability of detecting rules that are specific to the dataset. Furthermore, the decrease (rather than increase) in theory size with increased extensiveness of the search suggested that the problem of increased error due to over-searching was tangential to the problem of the notion of overfitting, the latter of which they construed as corresponding to rule/ruleset length or complexity.

Mutter, Hall, and Frank (2004): Exhaustive search does not significantly improve or worsen results, but takes much longer

Mutter, Hall, and Frank (2004) compared accuracy, classification speed and rule "compactness" of rulesets created by two different implementations of CBA (association rule based decision list algorithm) with C4.5 (decision tree), RIPPER (sequential covering) and PART (constructs decision trees from partial trees). Their main objective was to understand whether an exhaustive search using CBA would improve classification performance over other, more standard classification approaches. Performance of these approaches was compared across 12 datasets from the UCI repository. Compactness was operationalized as the number of rules that were used for classification. The CBA approaches were comparable in accuracy to the standard rule learning approaches across all datasets, but they required a much greater number of rules and took considerably longer than standard approaches (generally, over 1000-fold).

Janssen and Fürnkranz (2009): Efficacy of exhaustive search differs across heuristics and datasets

Janssen and Fürnkranz (2009) compared predictive accuracy and theory size across varying degrees of search extensiveness of sequential covering in 22 UCI datasets, using a 10-fold stratified cross validation approach. In addition to comparing performance across a wide range of beam widths (1, 2, 4, … 2^i, …, 2048, and exhaustive search), they also compared results across different rule-selection heuristics. The latter condition was a notable difference between the Quinlan and Cameron-Jones' (1995) study, which had used only the Laplace error as a criterion for rule improvement. Janssen and Fürnkranz expanded the set of search heuristics because searches with a narrow beam width would be particularly susceptible to any bias in the search heuristic.

The relationship between accuracy and beam width, averaged across datasets, varied considerably across search heuristics. The finding by Quinlan and Cameron-Jones (1995) that greedy or near-greedy searches led to higher accuracy than exhaustive searches, was replicated in five of the nine heuristics: for the other heuristics, the exhaustive search performed comparably (3 heuristics) or slightly better (1 heuristic) than less extensive searches. Comparison of results across datasets suggested that exhaustive searches may lead to increased accuracy when the dataset contains more variables, and variables with more levels. Janssen and Fürnkranz also found that exhaustive searches tended to find fewer but slightly longer rules. This difference in the discovered rule types may partly explain why greedy search may yield better results in some cases but not others: relationships that actually exist in the population may be more reflective of one rule type than the other.

Ordonez (2006): Exhaustive list of rules can be useful to practitioners

In an applied research study with a slightly different aim and design than those just described, Ordonez (2006) compared decision tree and association rule on prediction of medical diagnosis. The main objective was to explore whether association rules would uncover more useful rules to clinicians, being an approach less familiar to clinicians and medical researchers.

Ordonez compared accuracy, medical significance, and usefulness of rules generated by association rules and decision trees. The outcome of interest was heart disease, and the inquiry was conducted on a medical dataset of 655 patients with 25 predictors. The association rule approach produced over 1300 rules even after filtering for high support, confidence and lift, many of which were “medically significant” (defined by Ordonez as having at least 90% confidence and a lift greater than 2), including some that were surprising or unexpected to the domain. For the purposes

of identifying medically useful and/or interesting rules association rules was better, Ordonez concluded, since it produced greater number of such rules.

Section summary: No free lunch, but association rule mining may serve a unique purpose for practitioners

The evidence above suggests that exhaustive searches do not necessarily improve ruleset accuracy. However, here again, the no-free-lunch theorem seems to apply. For *some* datasets, *some* purposes, and with *some* heuristics, exhaustive searches perform better. In addition, Ordonez's (2006) study reminds us that particularly for applied researchers and practitioners, the interestingness of individual rules can be more important than producing a relatively more accurate theory. In such cases, exhaustive searches for rules (with cross-validation) seems more appropriate and helpful.

2.3.5 Ensemble approaches

Results of a single covering or tree model can be biased due to the sequential nature of model construction. This is because if a biased variable is added in the early stages of model construction, the accuracy of everything that follows becomes questionable since the choice of the latter variables depend on the former. Ensemble approaches were developed to compensate for such lack of robustness in decision tree and other data mining models. They generate multiple models from the same dataset by resampling and/or reweighing, and combining their results to make final predictions. Their general advantage is the improvement in classification accuracy relative to single trees and other data mining methods. Their general disadvantages are the loss of simplicity and interpretability of the results (Breiman, 1996; Tufféry, 2011) and increased computational burden (Strobl et al., 2009). As the loss of the interpretable rules makes ensemble approaches much

less attractive for use in orthodox education research, my study will include only a few examples of ensemble approaches.

Bagging, boosting and Random Forest are three major approaches to creating ensembles. The first two can be applied in conjunction with any supervised data mining approach, while the latter is instantiated in a tree-based approach. Several good reviews exist on these topics. For example, Dietterich (2000), provides an elegant conceptual introduction to general categories of ensemble methods, reviewing the process and rationale behind Bayesian averaging (the original ensemble method), error-correcting output coding, bagging and boosting, as well as empirical studies that compare these methodologies. Polikar (2006) reviews important topics on ensemble based methods, including the main ways in which diverse classifications can be created from the same data, ways in which predictions can be combined, and a summary of empirical literature that attempts to explore whether any of the methods are uniformly better than others. The clear organization, non-technical style of writing and visual treatment and style of writing makes this review particularly novice-friendly. Other helpful reviews and overviews are found in Strobl et al. (2009), Kotsiantis (2011) Kotsiantis (2014) and Galar, Fernandez, Barrenechea, Bustince, and Herrera (2012).

Bagging was a term coined by Breiman (1996), short for "bootstrap aggregating." It involves taking a bootstrap sample (i.e., random sample of the same size, drawn with replacement) or subsample (smaller samples without replacement), and building a model of each. Predictions of these models are combined—typically using majority voting for classification, and averaging for regression. Bagging works well for rule learners and other unstable classifiers where small changes in the data set can lead to large changes in the model (Bühlmann & Yu, 2002; Chandrahasan, Y, Sridhar, & L, 2011; Dietterich, 2000).

Boosting was developed by Freund and Schapire (1995, 1996; Schapire & Freund, 2012). Also known as "stage-wise additive modeling", it iteratively refines models such that the successive phases of the models are more likely to correctly classify cases that were mispredicted by earlier models. To do this, bossing reweighs the sample after each round so that cases that were mispredicted in the earlier round become more heavily weighted in the next round modified version of the sample. By focusing on improving predictions of the cases that are difficult to classify, boosting increases the chance of improving the overall accuracy. Boosting is fast and can be applied easily to any supervised classification approach (Breiman, 1998). It is also known to be stronger than bagging when the data is noise-free (Kotsiantis, 2011).

Random Forest is an ensemble algorithm specifically for decision trees and is considered "the most important" recursive partitioning method by Strobl et al. (2009, p. p.324). Introduced by Leo Breiman (2001) and later trademarked by Breiman and Adele Cutler, it combines bagging (Breiman, 1996) and random subspace methods (Ho, 1998), introducing additional diversity to the bagging approach by conducting it with randomly selected subsets of variables. Random forests generate more diverse trees than bagging, which result in a lower chance of error. It also allows for detection of relevant predictors that were masked by stronger predictors in a traditional algorithm which searches only for locally optimal predictors.

There are many empirical studies on the relative efficacy of ensemble approaches and their combinations. For example, Zaman and Hirose (2011) compared performance of bagging, boosting, and bundling (bagging + boosting) on 20 UCI Machine Learning repository datasets using small training sets and showed that bundling performs best on average. Kotsiantis (2011) showed that combining bagging, boosting, rotation forest and random subspace methods performed better than these methods alone statistically speaking, although the practically the

differences may not be so significant. Here again, the no-free-lunch theorem (Section 2.3.2) applies.

2.4 EVALUATION OF RULESETS AND RULES

When a decision tree, sequential covering algorithm, or association rule algorithm uncovers a set of if-then rules, how do we know the extent to which we may trust these rules? There are several things we might mean when we ask the extent to which rules or rule-sets are trustworthy. First, we might be wondering whether the patterns detected by the algorithm actually exist, or whether they are just an artifact. They could be an artifact of sampling, algorithmic bias, data representation, or data inadequacy. They could also be an artifact due to "leakage"—a data mining term that describes illegitimate correlations between the outcome and information used to predict the outcome (Elkan, 2012). Second, even if we believe that the patterns are not mere artifacts, we can still wonder about the extent to which they are "good." We can wonder about the extent to which individual rules are good, as well as the extent to which a group of rules, or rule sets, are good.

2.4.1 Assess whether the "discovered" patterns are likely to be mere artifacts

When data mining results strongly reflect the idiosyncratic features of the sample that do not exist in the population, they "overfit" the data and express patterns that exist only in the sample. Model overfit is considered the greatest threat to validity in data mining (Elkan, 2012). The most

commonly used approach[4] to detect and avoid overfit is to examine whether rulesets and/or rules induced from one sample apply correctly to other samples (Elkan, 2012). If one is working with a single large dataset, rules or rulesets can be created with a random subset (typically 70%) and tested for accuracy on the remainder. It can be reasonable to stratify the sampling across the levels of the dependent variable. Such an approach is referred to as the training-and-test-set approach, or the holdout method: the "training set" or "holdout sample" refers to the sample through which the rules or rulesets were created, while the "test set" refers to the sample used to evaluate the data. For a fair evaluation of overfit, it is essential that the test set is independent of the rule/ruleset generation process.

The rationale underlying the holdout method is that if the rules or rulesets are just an artifact of sampling, they would not be detected in another, independent sample. The test set serves as this independent sample. Often a particular algorithm has several parameter settings that need to be predetermined by the user. As the settings can affect the outcome, it is recommended that the researcher uses *only* the training set to explore and optimize the settings, and to leave the test set untouched for evaluating the final model (Elkan, 2012; Salzberg, 1997). Again, the bottom line is that the test set should be independent of the rule/ruleset generation process to ensure a fair evaluation of sampling bias and detection of model overfit.

A widely used variant to the holdout method is k-fold cross-validation. Used when the dataset is too small such that using only a subset might not yield reliable results, it involves subdividing the dataset S into k equal-sized subsets S_1 through S_k for $i = 1$ through k, and running the algorithm k times, each time with $S\text{-}S_i$ as the training set, and S_i as the test set. The average

[4] In addition to the empirical methods described here, there are also theoretical methods for estimating generalization error of models, although less commonly used. See Maimon and Rokach (2005), pp.153-155.

classification error rate among the k tests are typically considered to be an informal, conservative estimate of the error rate of the model created using the entire dataset S. The most commonly used number for k is 10, with some reporting that increasing the number of folds beyond 10 makes little difference in the final estimate (Breiman & Spector, 1992; Kohavi, 1995). Stratifying the sampling, and repeating the 10-fold cross validation process 10 times may improve the error estimate slightly (Witten et al., 2011).

Leave-one-out cross-validation (LOOCV) is a special case of k-fold cross-validation, when k is equal to the sample size minus 1. In theory, LOOCV could yield a more accurate estimate of the error rate as it uses the maximum amount of data for model creation in each fold, and because the process is deterministic (i.e., does not involve random sampling, which reduces sampling error). However, LOOCV has some drawbacks, including computational burden, and inability to create reliable models under conditions where stratification matters (Elkan, 2012; Witten et al., 2011).

The rules and rulesets may also be artifacts of algorithmic bias. Artifacts of algorithmic bias are patterns that do not actually exist but are "detected" by an algorithm because of an interaction between idiosyncratic features of the algorithm and dataset. For example, it has been demonstrated that all else being equal, the CART algorithm (Breiman et al., 1984) tends to favor continuous variables over discrete variables (Loh & Shih, 1997). In a previous study, my colleagues and I have found that three different decision tree algorithms produced three different trees (Iwatani et al., forthcoming). To prevent erroneous inferences due to algorithmic bias, researchers should have a good understanding of the strengths and weaknesses of the algorithms they are using, and pursue the same research question using a number of different algorithms that differ fundamentally in their approach.

Rules and rulesets may also be an artifact of data representation, meaning that patterns may seem to appear only because of the idiosyncratic way in which the data were transformed. Data miners often transform variables before trying to detect patterns among them. If patterns are detected using transformed variables, it is prudent to check whether alternative, equally justified methods of representing that variable would yield the same results.

Data inadequacy—including missing, erroneous and, irrelevant data—also affect the trustworthiness of the rules and rulesets generated from data mining. Data miners conducting secondary data analysis typically do not have much control over the data content and collection methodology. However, understanding the nature of the data and how it was collected is essential for data mining, as it helps determine which variables are reliable or relevant to the task at hand, and what might be missing. Data miners can try to ensure internal validity of the study from which the data was collected by avoiding the use of datasets, variables, or portions of variables that are highly erroneous. If potentially relevant variables or constructs appear to be missing, that information might be supplemented through other sources (e.g., mean district data, or neighborhood socioeconomic status might be relevant, but missing, from student-level data provided by a state Department of Education). It is probably good practice for data miners to explicitly note the limitations of the data, along the three dimensions (what is likely to be missing, erroneous and irrelevant), before exploring the data.

Leakage is a term used to describe circumstances where good predictions were made because information utilized in the model were illegitimately correlated with the outcome. Elkan (2012) describes several ways through which leakage can occur. An illegitimate predictor variable can be a cause of leakage. For example, predicting high school graduation with number of total credits earned would be leakage, since credits are a main determinant of whether a student

graduates. Essentially, the outcome is hidden within the predictors. Human background knowledge can also be a cause of leakage, as in some situations where a good model is attained after a researcher drops a subset of cases from the dataset for being "anomalous." If exclusion of this subset was based on the researcher's knowledge about their anomalous performance on the outcome—and if there were no other way in which the exclusion could be justified—this would be considered leakage.

Illegitimate cases contained in the training set can also be a cause of leakage. As an extreme example, the performance of the model on the test set would look *excellent* if the test set contained exact clones of the training set, but this would only be because knowledge about the test set had (metaphorically) "leaked" into the training set. Less extreme illegitimate linkage among cases commonly surfaces as a problem in orthodox education research as when members of the sample attend the same school, have the same teachers, or share some other characteristic that might affect the outcome. Inferences from such datasets would only be generalizable to those who also possess these shared characteristics, unless there is sufficient reason to believe that the characteristics are not related to the outcome. Data miners must be careful to know their data were created (e.g., how cases were sampled), especially when analyzing secondary data, since linkages between cases are not immediately visible from the dataset.

To summarize, there are many reasons for patterns discovered in data mining to be mere artifacts; data miners should be careful to avoid confusing shadows with monsters under the bed. Best practices include: cross-validation, testing the same question with a range of different algorithms, understanding the algorithms' basic operational principles, exploring different variable representations, examining and addressing dataset limitations before analysis, and scrutinizing the results to minimize the possibility of leakage.

2.4.2 Evaluate the extent to which discovered patterns are valid

Even if the mined patterns are not artifacts (i.e., they really are patterns), the question remains as to how "good" they are. This section reviews several criteria in which the validity of mined rules and rulesets are evaluated. Much of the discussion concerns how to quantify model performance under cross-validation, since this is the main way in which the data mining field validates rules and rule sets inferred from the data (e.g., Grzymala-Busse, 2005; Hand, 1997). This includes ways to quantify predictive accuracy of the model based on final class assignment, as well as probability assignments, of individual cases. However, just as validity is viewed as a multifaceted concept in the field of educational measurement (Messick, 1989), there are conceptions beyond predictive accuracy that comprise rule and ruleset validity, including computational complexity, comprehensibility and interestingness (Maimon & Rokach, 2005). I first discuss how to evaluate the validity of rulesets (i.e., models made from decision trees and sequential covering approaches) using predictive accuracy, and other approaches. I then summarize approaches for evaluating the goodness of individual rules.

2.4.2.1 Predictive accuracy of rulesets

Most data mining approaches make two types of predictions about an individual case: a final class assignment (e.g., whether a student will graduate or drop out), and probabilistic assignment to each case (e.g., the probability that the student will graduate, and the probability that the student will dropout). Either can be used for as the basis for quantifying how good a model is in predicting the outcome of cases. The most commonly used approach to validate a model consisting of a set of rules (e.g., decision tree models and sequential coving rule sets) is by examining the extent to which the rulesets, taken together, make correct final classifications of the test set. In general,

higher classification accuracy would imply higher ruleset validity. Examining only the overall accuracy rate (number of cases classified correctly / number of total cases) is too simplistic, however, since it does not take into account the proportion of cases that belong to each of the classes, and the relative importance of correctly classifying each of the classes (Provost & Fawcett, 2013). For example, in a school where 95% of the students graduate, a model predicting that "everyone graduates" will be 95% correct but utterly useless, as it does no better than chance, and provides no insight about what we really want to know (i.e., who might drop out). Thus, instead of making inferences about validity based on the overall accuracy, the entire pattern of classification accuracy, broken down by class, must be examined and reported.

Classification results for an outcome variable with *k* levels are expressed in the form of a *k* by *k* contingency table (called a "class confusion matrix"), with its actual classification and model-predicted classification as rows and columns (there is no convention about their positioning). For example, a confusion matrix for an outcome with two levels—a class of interest and a class of non-interest, would express the extent to which cases were classified as belonging to the class of interest correctly (true positive) or incorrectly (false positive), and the extent to which cases were classified as being of non-interest correctly (true negative) or incorrectly (false positives) (Figure 3). Sometimes more than one, and/or all variables are of equally high interest. In those cases, the labels of the cell value (e.g., "true positive") would not readily apply, and different metrics may be calculated, but the confusion matrix will be constructed just the same.

Model A (N = 224)

		Predicted: Drops out	Predicted: Graduates
Actual	Drops out	2 *(True positive)*	21 *(False negative)*
	Graduates	4 *(False positive)*	197 *(True negative)*

Model B (N = 224)

		Predicted: Drops out	Predicted: Graduates
Actual	Drops out	10 *(True positives)*	13 *(False negatives)*
	Graduates	51 *(False positives)*	150 *(True negatives)*

Figure 3. Two examples of a 2 by 2 confusion matrix

Many predictive accuracy metrics can be calculated using confusion matrices. Some of the common metrics that can be calculated from a 2 by 2 confusion matrix are summarized in Table 1. Overall accuracy (or simply, "accuracy"), precision, recall, and F-measure are most commonly reported in the field of Data Mining (Elkan, 2012; Forman & Scholz, 2010; Japkowicz, 2013; Provost & Fawcett, 2013). Accuracy is simply the percentage of correct predictions by the model, or the percentage of the sum of the true positives and true negatives. In Figure 3, the accuracy of model A is (2+197)/224, or 89%, and that of model B is (10+150)/224 or 71%. Since the true positives and true negatives are aggregated in the calculation, unless there are equal numbers of the positive and negative class in the sample, accuracy is not informative about how well the model predicts positive cases and negative cases separately. Thus, accuracy is generally insufficient for the purposes of model evaluation.

Precision is the proportion of cases that were classified as positive that were actually positive. The precision of model A would be 2/(2+4), or 33%, while the precision of model B would be 10/(10+51) or 16%. This suggests that when model A predicts someone to be a graduate, it about twice as likely to be correct than model B. However, this metric does not take into account the proportion of dropouts the model correctly flags as being a dropout, or recall. Recall, or true positive rate for model A is 2/(2+21) or 9%, while it is 10/(10+13) or 43% for model B. In this metric, model B is over 4 times successful than model A.

Since neither precision nor recall by itself provides a complete picture of how well the model predicts the positive instances, the average of the two metrics—referred to as the F-measure—is often used to assess the relative predictive validity of the model.[5] The F-measure of model A is 2/[(1/33%)+(1/9%)], or 21%, while that of model B is 2/[(1/16%)+(1/43%)], or 30%. This implies that if precision and accuracy are equally important, model B is more accurate in predicting the dropouts than model A. The F-measure can be weighted by the relative importance of precision and recall, if one is more important than the other (Table 1). It is a reasonable measure with which to evaluate models if we are primarily interested in how well models correctly identify one particular class, since it is not as biased as the two other most popular summary metrics—accuracy and AUC (described below)—when classes are imbalanced (Forman & Scholz, 2010). However users should be aware of its limitations including its bias towards the majority class and the fact that it does not take into account true negatives (Powers, n.d.). It is best to examine and report the entire confusion matrix, rather than only reporting summary metrics (Elkan, 2012).

[5] The harmonic mean must be used, since precision and recall are both rates. For a simple and thoughtful exposition of the F-measure, see Sasaki (2007).

Table 1. Commonly used evaluation metrics of model predictive accuracy for binary classification

Metric	Formula*	Notes
Accuracy	(TP+TN)/n	Should not be the only metric reported.
True positive rate, recall, or specificity	TP/(TP+FN)	Proportion positive instances that were correctly classified.
False negative rate	FP/(TP+FN)	Proportion positive instances that were incorrectly classified.
True negative rate, or sensitivity	TN/(TP+FN)	Proportion negative instances that were correctly classified.
False positive rate	FN/(TN+FP)	Proportion negative instances that were incorrectly classified.
Precision, or positive predictive value	TP/(TP+FP)	Proportion of cases classified as positive that were actually positive.
F-measure	$2/(\text{precision}^{-1}+\text{recall}^{-1})$	Harmonic mean of precision and recall.
F-measure with weighting	$[(1+\beta^2)*\text{recall}*\text{precision}]/[(\beta^2*\text{precision}) + \text{recall}]$, where β = (importance of recall) / (importance of precision), and $0 \leq \beta \leq +\infty$.	β is the relative importance of recall to precision, commonly set to 1, 2 or .5. See Sasaki (2007) for an explanation.
Expected benefit	(TP/n)*(benefit of TP) + (FN/n)*(benefit of FN) + (TN/n)*(benefit of TN) + (FP/n)*(benefit of FP)	Benefit model confers per case.
Youden's index	sensitivity – (1-specificity)	Arithmetic mean between sensitivity and specificity

*TP = number of true positive cases; FP = number of false positive cases; TN = number of true negative cases; FN = number of false negative cases; n = total number of cases.

So far, I have discussed ways to quantify predictive accuracy of models based on how correctly or incorrectly they make discrete, final class assignments. As mentioned, it is also possible to quantify predictive accuracy based on how good they are in assigning probabilistic predictions. This method is useful if accurate probabilistic assignments matter at least as much as accurate final class predictions. This is often the case when we wish to rank order cases. For

example, if the goal is to identify 5 students who would most benefit from a summer academic enrichment camp (assuming that more than 5 students would benefit), we would want to know the relative extent to which the students would benefit, over and above "whether or not" the students would benefit. In this case, we want a model that correctly makes probabilistic predictions about each student so that we can select the students who have the greatest probability of benefitting.

Witten et al. (2011) describe two of the most commonly calculated metrics that summarize the extent to which models make accurate probability estimates of a sample: the quadratic loss function, and the information loss function. Both functions reward models that give better predictions, i.e., higher probabilities to individuals' actual outcomes and lower probabilities to non-outcomes. The quadratic loss function is calculated by squaring the difference between the actual prediction and the model-predicted probability for each class, where the actual prediction and summing those within and across individuals:

$$\sum_n \sum_j (p_j - a_j)^2$$

The probability vector, p_1, p_2, … p_j, denotes the predicted probabilities of the j possible outcomes or a particular case, and sums to 1. The vector a_1 … a_j denotes the actual outcome for an individual, where a_i would be 1 for the actual outcome, and 0 otherwise. n denotes the number of cases. Thus, for example, if a model predicts a dropout to have a 99% chance of dropping out, its quadratic loss for that individual will be $(.99\text{-}1)^2 + (.01\text{-}0)^2$, or .0002. On the other hand, if a model predicts that a dropout has a 51% of dropping out, the quadratic loss would be much higher $(.51\text{-}1)^2 + (.49\text{-}0)^2$, or .4802. Models that make more accurate predictions for each individual will have a smaller loss function.

The information loss function, based on Shannon entropy, is:

$$-\log_2 p_i$$

where p_i denotes the probability for the correct prediction. This function, also referred to as the negative log-likelihood function, expresses the average number of bits (yes-no signals) necessary to discern what the actual outcome is, given the probability distribution. If the model predicted probability distribution provides highly accurate information, there would be little/no need to ask further yes-no questions to correctly identify a person's outcome. For example, the information loss function of a model that predicts a dropout to have a 99% chance of dropping out would be $-\log_2(.99)$, or .0145. This means that on the whole, for this individual, we can rely solely on the model predicted probability to assess that he drops out. The information loss function would be $-\log_2(.51)$, or .97143, for a model predicts that a dropout has a 51% of dropping out. This means that in addition to relying on the model prediction, we must ask approximately one additional yes-no question to correctly identify whether this individual will be a dropout.

Witten et al. (2011) contend that there is no universal agreement on which of these information functions one ought to use—that the choice is "a matter of taste" (p.162). However, one should keep in mind that the quadratic loss function, unlike the information loss function, takes into account the probability predictions of the non-outcomes; it will favor predictions that distribute probabilities evenly among the non-outcome levels. Additionally, the penalty for assigning a low probability to a correct outcome is much more severe for the information loss function (maximum penalty is infinity) than the quadratic loss function (maximum penalty is less than 2).

Sometimes, the goal of modeling is not simply to maximize the prediction of a particular category (e.g., dropouts), but to find the optimal balance of correct and incorrect class predictions

such that it produces the maximum amount of benefit. The "expected benefit[6]" of a model can be calculated for a set of predictions, and be used to evaluate model performance, as long as the costs and benefits associated with each outcome can be defined (Provost & Fawcett, 2013). The average expected benefit is calculated by multiplying the benefit of each type of outcome with the probability of that outcome, and summing these across outcomes type and individuals, and dividing by the number of cases. For example, the expected benefit of results summarized in a 2 by 2 confusion matrix would be: (proportion of true positives)*(benefit of a true positive) + (proportion of false negatives)*(benefit of a false negative) + (proportion of true negatives)*(benefit of a true negative) + (proportion of false positives)*(benefit of a false positive). It is also possible to consider costs differentials during the model building process (so that it can be ignored during the evaluation process). Either the subsample ratios can be adjusted so that the sample size for each outcome is proportional to the cost of the outcome, or (at least in some algorithms) weights can be set in advance so that the algorithm takes into account the relative costs of inaccurate classifications (Witten et al., 2011).

Instead of comparing how well models would do in classifying the entire population (estimated by examining how they classify the entire sample), it is sometimes relevant to compare how accurate they are in *ranking* the population according to some metric of interest (e.g., probability of belonging to a certain category, or amount of expected benefit). A model can be good at ranking everyone, or ranking just those in certain score ranges. Depending on what the user is interested in doing with the model, it might be more valuable to have a model that is very good at ranking part of the population but not the rest, rather than a model that is moderately good

[6] The conventional term in Data Mining is "expected profit." I use "benefit" instead of "profit" to make it relevant to a wider range of applications encountered in education research. Benefit must be quantifiable, but does not necessarily need to be in terms of money.

at ranking everyone. Common approaches to assessing ranking accuracy in data mining include examination of profit curves, lift curves, Receiver Operating Characteristic (ROC) curves and precision-recall curves (Japkowicz, 2013; Provost & Fawcett, 2013; Witten et al., 2011). Ranking accuracies is a reasonable way to evaluate model performance when the outcome distribution are skewed and when that particular imbalance may not be encountered in data for model deployment (Chawla, 2005; Japkowicz, 2013).

Profit curves involve plotting the cumulative expected benefit across percentage of the population, after rank-ordering the population in descending order of some model-estimated score of interest. It is used in business to examine how much the expected profit will change as they target more people with a particular promotion. The most valuable model for the company would be one that predicts the maximum cumulative profit within the population range that it has the resources to target.

Similarly, one can plot any of the metrics in Table 1, across percentage of the population (rank ordered by model-estimated score of interest). A chart that examines the change in true positive rate (percentage of correctly classified positive cases) across the population is referred to as a cumulative response curve, and sometimes as a lift curve. Lift is the ratio between the true positive rate of the model, and true positive rate according to random chance. (As Provost and Fawcett (2013) point out, a chart that directly examines the lift across the population is also called a lift curve.) Lift curves helps visualize how true positive rates of different models compare for the top 1%, 5%, 10%, …etc., of model-predicted scores the sample.

The ROC (Swets, 1988) is similar to a lift curve, but plots true positive rate across increasing values of false positive rates. As with the other plots introduced above, it is created by sorting the cases according to the model-predicted score (model-predicted likelihood for instance

to be positive), and examining the confusion matrices made by: assuming that none of the cases are positive; the first observation as positive; first two observations as positive, first three observations as positive; and so on until all observations are positive. ROC can help examine the tradeoff between true and false positive rates within an algorithm, and compare the true positive rates between models, given different thresholds of false positive rates. Ideally, models should have perfect true positive rates (i.e., should correctly classify positive instances as positive), regardless of its true negative rate (i.e., regardless of how correct we are in classifying negative instances). The expectation according to random chance is, however, that we have the same probability of calling a positive instance a negative, as we do the probability of calling a negative instance a positive.

Intuitively, the ROC is more difficult to understand than cumulative response curves or lift curves. However, ROCs can be drawn prior to incorporating costs, and are thus advantageous over the latter types of curves when one is not completely sure of the costs and benefits of the possible decisions (Provost & Fawcett, 2001). The area under the ROC curve (AUC) is sometime used as a summary statistic to evaluate the model's average performance across all cases. It ranges between 0 and 1, and a greater value corresponding to a greater overall model performance, and a value of .5 corresponding to chance performance. AUC corresponds to the probability that the model ranks a random positive case above a random negative case. Like accuracy, AUC is misleading when the ratio between classes are skewed or if good model performance is desired for a particular subset of cases.

Relative accuracy of models that make numeric predictions (e.g., regression trees) can be compared based on summary statistics such as the mean squared error, root mean squared error, and correlation coefficient (correlation between predicted and observed outcomes). Different

measures of error can be useful depending on, e.g., how much weight the user wants to give to different kinds of outliers, and whether error should be weighed relative to the size of the predicted outcome. See Table 2 for measures that can be used to assess models that predict numeric outcomes.

Table 2. Evaluation metrics for models with numeric outcomes

Mean-squared error	$\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}$
Root mean-squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$
Mean-absolute error	$\frac{\lvert p_1 - a_1\rvert + \cdots + \lvert p_n - a_n\rvert}{n}$
Relative-squared error	$\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \cdots + (a_n - \bar{a})^2}$
Root relative-squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \cdots + (a_n - \bar{a})^2}}$
Relative-absolute error	$\frac{\lvert p_1 - a_1\rvert + \cdots + \lvert p_n - a_n\rvert}{\lvert a_1 - \bar{a}\rvert + \cdots + \lvert a_n - \bar{a}\rvert}$
Correlation coefficient	$\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA} = \frac{\sum_i (p_i - \bar{p})(p_n - \bar{a})}{n-1}$, $S_P = \frac{\sum_i (p_i - \bar{\boldsymbol{p}})^2}{n-1}$, and $S_A = \frac{\sum_i (a_i - \bar{\boldsymbol{a}})^2}{n-1}$.

Adopted from Witten et al. (2011, p. 180). p_1, p_2,…p_n are the predicted values and a_1, a_2,…a_n are actual values for each case. $\bar{a}$ is the mean over the training data, while $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{p}}$ are the means over the test data.

Whether and to what extent the predictive accuracy metrics are practically significant can depend on the baseline to which the metrics are compared. Different baselines are appropriate for different circumstances, so the researcher should articulate their baseline (e.g., random chance; everyone classified as the majority class; results of the best existing method of classification) and

argue for its appropriateness in the context of her research. For example, the kappa statistic, often calculated in conjunction with confusion matrices, is the proportion of cases the model classifies correctly, after eliminating the number of cases we expect to get correct by random chance (Witten et al., 2011). This measure could be appropriate if the goal is to do better than chance.

Caution needs to be exercised when using predictive accuracy metrics with cross-validation, as the order of calculation may yield biased results. Forman and Scholz (2010) was perhaps one of the first to clarify this point. They demonstrated that the F-measure should be calculated using the sum of the number of true positives and false positives over all the folds, rather than averaging the F-measures of each fold and taking the average, or using the average the precision and recall over the folds. They also showed that the AUC should be calculated by averaging across all the folds, rather than creating a big ROC curve that combines the information from all of the folds. The bias using the other calculation methods is particularly large when the data are skewed.

2.4.2.2 Other notions of ruleset validity

So far, we have only discussed how to evaluate rulesets based on how accurate they are in classifying the individuals in a test set. Given two or more rulesets that are adequate in this regard, however, depending on the research objective, there could still be other reasons for an orthodox education researcher to prefer one ruleset over others. If the researcher is trying to identify a ruleset that describes key patterns within a domain, she may wish to examine the extent to which the ruleset is *comprehensive* in including variables and/or patterns known to be important in the domain, and/or she might want to check that all rules within the ruleset are *relevant* to the domain. She would also probably want to check that the rules discovered are generally *consistent* with what is known in the field—i.e., that they don't outright contradict what is agreed upon by scholarship

to date. For example, if the objective is to arrive at a comprehensive model that explains student achievement among US public school students, a comprehensive model should include main explanatory factors that are known to be relevant in the field, such as family income, parental education, school funding and student attendance. A model that includes few or none of the known key variables, however accurate, may be considered "not good" for the purposes of that particular research. Similarly, a model that contains irrelevant factors (e.g., color of school mascot or student ID) or erroneous relationships (e.g., "low student attendance leads to high achievement"), are also likely invalid. Domain knowledge is required to assess the extent to which rules are comprehensive, relevant, and/or consistent with what is already known.

Usefulness, *novelty*, and *simplicity* are some other criteria in which rule induction models can be evaluated. The usefulness and novelty of a rule induction model greatly depend on the extent which individual rules are useful and/or novel. Generally, a ruleset is useful or novel if the underlying rules are useful or novel. The way in which rule usefulness and rule novelty have been defined and assessed will be reviewed in more detail below. However, even if the individual rules are not particularly novel or useful, there is a chance that a model can combine or highlight these rules in a novel or useful way. I do not know of literature that discusses this in any detail, but I presume that domain knowledge would be essential for judgment of this kind.

Finally, if the research goal is to find out rulesets or rules that reflect actual patterns in the world, the *generalizability* of the rulesets and/or rules to other populations could also serve as validity evidence. The underlying logic is that if rules or rulesets discovered for one population also apply in other populations, they are more likely to be true. For example, if a researcher made an unexpected finding about achievement among Californian 8th graders in 1988, the finding would

seem more valid if the same relationship were also discovered in other years, in other states, and/or among different subgroups of students.

2.4.2.3 Interesting measures for rule evaluation

There has been much scholarly discussion on how to assess the goodness of individual rules created by rule induction algorithms (reviews include: Bourassa, 2011; Carvalho, Freitas, & Ebecken, 2003; Fürnkranz & Flach, 2005; Geng & Hamilton, 2006; Hilderman & Hamilton, 1999; McGarry, 2005; Murthy, 1998). Much of the research on this topic is motivated to improving the results of association rule mining algorithms, which typically produce too many rules, including many that are imprecise, obvious, useless or otherwise uninteresting to the user. Metrics designed to quantify the extent to which individual rules are good have broadly been referred to as "interestingness measures." As mentioned previously, since "good" or "interesting" can mean several different things, interestingness metrics differ correspondingly in what they measure. The appropriate measures for any given data mining project depends greatly on the goals for which the rules are being mined.

A large number of interestingness measures work to identify the extent to which the rules correctly predict the sample. There is a lot of overlap between this class of interesting measures, and the predictive validity metrics for rulesets. On the other hand, scholars have also tried to quantify the extent to which rules are new, surprising, unexpected, and/or useful. Some such approaches require user input, while others do not. In this section, I introduce the three classes of approaches proposed so far on how to evaluate rule interestingness (which could also be regarded as rule "goodness" or "validity.") Some approaches have been labeled "objective," meaning they do not require additional user knowledge or input. In contrast, some of the "subjective" approaches involve user input on factors that would make rules more interesting. I will also comment on

advantages and disadvantages of the specific methods, from the perspective of an applied researcher.

The first approach to identifying better rules is by applying an objective filter to the rules to eliminate redundancies among them. Elimination of redundant rules allows the user to more easily scan the rules for further validation. Generally, the redundancy-elimination method works by detecting and eliminating logically redundant rules. Furthermore, in the case of association rule mining, algorithms can determine the minimum set of rules or patterns to adequately describe the data (e.g., Padmanabhan & Tuzhilin 2000; Bastide et al 2000; Li and Hamilton 2004). In addition, scholars have developed methods to balance the simplicity of the theory (set of rules) with the proportion of data which it can account (e.g., see accounts in Forsyth et al 1994; Vitanyi and Li, 2000). The underlying rationale for these approaches are that unique and simpler rules are to be preferred, all else being equal.

The advantage of the redundancy-elimination approaches is that they can be done for any set of rules, regardless of whether they were generated from decision trees, association rule mining algorithms or sequential covering. However, since this approach is conservative in its identification of non-interesting rules, it does not necessarily reduce the number of rules to a manageable size, or a set that is interesting in some other regard. Non-redundant and simple are often not enough to consider something interesting. Furthermore, this approach does not result in a meaningful rating or ranking of rule interestingness.

The second class of approaches to determining rule interestingness concerns predictive accuracy and coverage of the rules within samples. The underlying assumption of this class of approaches is that for a rule to be interesting, it must be accurate, precise, and/or applicable to many. These approaches are objective, easily generated for any rule, and so very commonly used

to evaluate rules produced by association rule mining (Geng & Hamilton, 2006). I conceptually describe the metrics in the paragraphs that follow, and list more formal expressions of each in Table 3.

Table 3. Examples of probability-based interestingness measures for rule A→ B

Measure		Tan et al. (2002) properties*
Support	$P(AB)$	2, 4
Confidence/Precision	$P(B\|A)$	2, 8
Coverage	$P(A)$	-
Prevalence	$P(B)$	-
Recall/Sensitivity/True positive rate	$P(A\|B)$	-
False positive rate/Fallout	$P(A\|\neg B)$	-
Specificity	$P(\neg B\|\neg A)$	-
False omission rate	$P(B\|\neg A)$	-
Accuracy	$P(AB) + P(\neg A\|\neg B)$	-
Lift/Interest	$P(B\|A)/P(B)$ or $P(AB)/P(A)P(B)$	1, 2, 3, 4
Odds ratio	$P(AB)P(\neg A\neg B) / \{P(A\neg B)\, P(\neg AB)\}$	1, 2, 3, 4, 5, 6, 7
Positive likelihood ratio	$P(A\|B)/ P(A\|\neg B)$	-
Relative risk/ Relative probability	$P(B\|A)/ P(B\|\neg A)$	-
Leverage	$P(B\|A) – P(A)P(B)$	-
Added Value	$P(B\|A) – P(B)$	1, 2, 3
Cohen's Kappa	$P(AB)+P(\neg A\neg B)-P(A)P(B)-P(\neg A)P(\neg B) / \{1-P(A)P(B)-P(\neg A)P(\neg B)\}$	1, 2, 3, 4, 6
Laplace Correction	$\{N(AB)+1\} / \{N(A)+2\}$	2
Gini Index	$P(A) * \{P(B\|A)^2 + P(\neg B\|A)^2\} + P(\neg A) * \{P(B\|\neg A)^2 + P(\neg B\|\neg A)^2\} – P(B)^{2+} - P(\neg B)^2$	1, 7
J-Measure	$P(AB)\, log\{P(B\|A) / P(B)\} + P(A\neg B)\, log\{P(\neg B\|A) / P(\neg B)\}$	1

Table 3 continued

Piatetsky-Shapiro	*P(AB) - P(A)P(B)*	1, 2, 3, 4, 6, 7
Information Gain	*log{P(AB) / (P(A)P(B))}*	-

Adopted from Geng & Hamilton (2006, p.9) and Tan, Kumar, and Srivastava (2002).
*Refer to note 7 and associated text on p.74. Indicated only if available in Tan et al. (2002).

The most commonly examined metrics are support, coverage and confidence. Both support and coverage indicate the generality of the rule, while confidence is an indication of its precision. ***Support*** of a rule is the proportion of cases in the population for which the rule is valid. This metric could be interesting for stores who want to identify item sets that are frequently brought together by customers, journal publishers who want to know which articles are frequently viewed and/or downloaded together, and social media websites who are interested in identifying people who tend to be friends with each other. The disadvantage of this metric is that patterns that very frequently happen are generally well-known, because of their ubiquity. Additionally, this metric does not factor the proportion of cases for which the rule applies but is not true. ***Coverage*** of a rule is the proportion of cases in the population that satisfy the rule's antecedent. The disadvantage of this metric is that not only do widely applicable rules tend to be already known (and so uninteresting), it does not include any information about the consequent. However, coverage is conceptually useful, and practically useful for rule evaluation when combined with other metrics. ***Confidence***, a measure of rule precision, is the probability that the rule consequent is true, given the antecedent. It is a metric that is useful to look at in conjunction with support and/or coverage, since rules that are extremely precise may not be so useful if it only applies to a very small proportion of the population.

One of the caveats of confidence as a metric of precision is that it does not consider the correlation between the antecedent and the consequent. To illustrate the importance of accounting for the correlation, suppose the rule "if a student attends a summer program, they graduate" is

applicable to 35% of the population, and applies correctly to 30 percent of the population. This means that the coverage is 35%, support is 30%, and confidence is (30%)/(35%), or 86%. While confidence seems reasonably high, intuitively, the rule would not seem as interesting if it were the case that 100% of the students graduated rather than 32%. Thus, a measure of the correlation between the antecedent (summer program) and the consequent (graduation) is highly relevant for assessing the goodness of the rule.

Two commonly used measures of association (correlation) between the antecedent and consequent are added value and lift. ***Added value*** is a measure of the extent to which the confidence is different from random chance, and is calculated by subtracting the probability of the consequent occurring in the population (i.e., random chance occurrence of the consequent), from the confidence. Positive values indicate that the rule does better than random chance in its precision of detecting the consequent and vice versa. ***Lift***, again, is a measure of how well the rule performs against chance, and is the ratio between the probability of an outcome for the particular subgroup identified by the rule, and the probability of the outcome of that overall sample. For example, if my rule uncovers that the graduation rate of school district A is 90 percent, while the average graduation rate including all school districts in the dataset is 60 percent, the school district A's lift is 90/60 = 1.33. Lift of 1 indicates that the antecedent and consequent are independent. Of course, larger added value or lift rules are not always a sufficient condition for the rule to be interesting. Rules with relatively high interest values may be uninteresting due to other criteria of interestingness (e.g., low support, low utility).

Many interestingness measures are combinations of a measure of generality (either support or coverage), and a measure of precision that considers the correlation between rule antecedent and consequent. For example, the measure by Piatetsky-Shapiro (1991) is the product of coverage

and added value, while the two-way support measure by Yao and Zhong (1999) is a product of support and the natural log of lift (see Geng and Hamilton (2006) for a review). In a review of 21 such interestingness measures, Tan et al. (2002) found eight mathematical properties that distinguish these measures including some that were already known (Piatetsky-Shapiro, 1991) and some that were not as well known.[7] There was no measure that could objectively be considered as "superior" across all eight properties, in part because the relevance of each property depends on the characteristic of the pattern that the user is looking to find. They also found, however, that the measures were highly correlated with one another when support was held constant, and particularly when the support was low.

What I consider to be a third class of rule interestingness measures attempt to identify rules that have a greater potential to be unknown and/or valuable to the user, and do this by quantifying what can be considered novelty, peculiarity, or surprising-ness of rules. These metrics are rather heterogeneous, both in what they attempt to measure, and how the measures are operationalized. In fact, they are probably more similar in what they are not, rather than what they are—all of these measures are *not* measures of redundancy, generality, precision or their combination. For the sake of simplicity, I will hereon refer to this motley group of interestingness measures as *noteworthiness* measures, where noteworthy refers to some conception of interestingness beyond non-redundancy,

[7] The eight properties, described conceptually, are: (1) The measure is 0 if the antecedent and consequent are statistically independent; (2) The measure monotonically increases with support, when coverage and relevance (the proportion of cases in the sample for which the consequent applies) are held constant; (3) The measure monotonically decreases with coverage when support and relevance are held constant, and similarly with relevance when support and coverage are held constant; (4) For all attribute-value statements A and B that characterizes the sample, the measure is consistent for the rules A → B and B→ A; (5) The measure is consistent even when particular rows and/or columns of the rule applicability confusion matrix are multiplied by a constant; (6) The measure does not between positive and negative correlations of the rule antecedent and consequent—the value remains the same whether rows *or* column of the confusion matrix are swapped; (7) The measure remains the same even when rows *and* columns of the confusion matrix are swapped; (8) The measure remains the same when even when irrelevant cases (i.e., cases where neither the antecedent nor consequent apply) are added to the sample.

generality and/or precision. It is certainly possible to use noteworthiness measures in conjunction with other aforementioned types of interestingness measures. For example, one could rank the rules in order of some novelty metric after excluding redundant rules and rules that do not meet certain thresholds of generality and precision. I will first describe several objective, or data-driven approaches to quantifying noteworthiness measures, then describe subjective, or user-driven noteworthiness measures, keeping my explanations at a conceptual level.

One approach to objectively identifying noteworthy rules is by examining how "far away" rules are from one another, and locating rules that are isolated from the rest. There are at least two ways to think about what it means for a rule to be "far away" from the rest. On a semantics-based notion of distance (Toivonen, Klemettinen, Ronkainen, Hätönen, & Mannila, 1995), rules that describe many of the same cases are considered to be closer to one another than rules that have very little overlap in the cases they describe. If data mining revealed many precise rules for students in a high school, but only several of them applied to a subgroup of 30 students, this subset of rules would be considered more distant from the other rules and so have a higher noteworthiness index. Since this measure is sample-dependent, it has a high chance of capturing sampling error. A syntactical notion of distance (Dong & Li, 1998), on the other hand, is the extent to which the same attribute-value pair occurs within and across different rules. The more syntactically similar the rules, the closer the distance. For example, if many of the mined rules about academic success in college included subsets of the same 10 variables, but there were also a handful of them include none of these frequently detected variables, those latter rules would be flagged as being noteworthy. The syntactical distance between two rules do not change regardless of how often and to which cases the rules apply.

Another popular objective approach to finding noteworthy rules is to identify reliable "exception rules," or rules that contradict rules that are known or likely to be known (Carvalho et al., 2003; Freitas, 1998; Hussain, Liu, Suzuki, & Lu, 2000; Suzuki & Kodratoff, 1998). An intuitive example of a commonsense and exception rule pair, commonly known as Simpson's paradox, is mentioned by Suzuki and Kodratoff (1998): the commonsense rule that seatbelts are safe is contradicted by the rule that seatbelts are risky for a small child. As this illustrates, exception rules can be noteworthy precisely because it contradicts commonsense, even though it may not score so high on generality or precision.

One approach to flag rules that are likely to be exception rules is to begin by identifying commonsense rules (e.g., rules exceed high generality and reasonable precision thresholds), and for each, "zoom in" to cases where see whether there are more specific version of these rules that contradict the outcome with reasonably high precision (Carvalho et al., 2003; Hussain et al., 2000). Conversely, one can identify very specific rules with high precision (but low generality), and see whether the relationship between the antecedent and consequent becomes contradicted by examining more general cases of the antecedent i.e., by "zooming out" (Carvalho et al., 2003; Freitas, 1998). Different measures can be used to quantify the extent to which the exception rule contradicts the commonsense rule including the number of cases where the exception applies calibrated by the specificity of the exception (Carvalho et al., 2003; Freitas, 1998), information gain (Carvalho et al., 2003; Hussain et al., 2000) and "intensity of implication" (Suzuki & Kodratoff, 1998).

The approaches to finding noteworthy rules mentioned above are applied post-hoc, after the rules are generated. In contrast, some approaches are integrated into a search algorithm. The STUCCO algorithm by Bay and Pazzani (2000) is an alternative to decision trees that induces

noteworthy rules by mining for contrast sets, or differences among groups. Rather than looking for which factors best separate the respondents into levels of the outcome (which is what typical decision trees do), their algorithm looks instead for the relative differences in the characteristics of the two groups, and figures out which set of predictor characteristics distinguish them the most. The hybrid algorithm by Carvalho and Freitas (2002, 2004) uses a decision tree (C5.0) first to classify large disjuncts, then uses a genetic algorithm to classify "small disjuncts." Small disjuncts are rules that reliably apply only to a small number of cases, which tend to be looked over by typical interestingness measures due to their lack of generalizability.

Finally, there are several subjective, or user-driven noteworthiness measures, where rules are assessed on the extent to which they are like user-specified rules. Some approaches work by weeding out rules that are the same as or similar to what the user already knows to be true (e.g., B. Liu, Hsu, & Chen, 1997; B. Liu, Hsu, Chen, & Ma, 2000; B. Liu, Hsu, & Ma, 1999). Another set of approaches attempt to find Simpson's paradoxes: the user specifies rules she is interested in, and the algorithm looks for either more general or more specific rules that contradict them (Padmanabhan & Tuzhilin, 1998, 1999, 2000). There have also been proposals that require users to rate the noteworthiness of some of the rules that have been generated so that the algorithm can "learn" the types of rules that are more or less noteworthy to the user (Silberschatz & Tuzhilin, 1995, 1996).

Subjective approaches might increase the likelihood that results will be surprising to the user, and seems particularly appropriate when there is strong and justified user conviction about which rules are valuable than others. However, specifying relevant rules from which to judge noteworthiness could be tedious and error-prone. The approach would also be hard to justify if there is no strong consensus within the field about what kind of rules would be noteworthy or not.

2.4.2.4 Other notions of rule validity

A rule that is general, precise and/or noteworthy in a population may lack validity in other regards. Depending on the objective of the study, the rule may have to generalize to different populations or accurately characterize sub-populations. The rule may have to be actionable, practical, cost-effective, lead to no unintended consequences, or be logically consistent with or derivable from other rules that are already known in the field. Just as it has become very important for assessment developers to articulate in advance how validity of inferences of test scores would be ascertained (American Educational Research Association et al., 2014; Kane, 2006), it would be beneficial for educational researchers to articulate in advance *what it means for a mined rule to be valid* in a way that makes sense given the context of their study.

2.5 APPLICATIONS OF RULE INDUCTION IN ORTHODOX EDUCATION RESEARCH

I have discussed how rule induction data mining methods work, including the potential benefits they can bring to orthodox education research. I now turn to a review of how rule-based data mining approaches have been applied to education research so far. Here I review educational research that applies rule-based data mining in contexts outside of online learning.

I conducted the review in August 2015, by searching the ERIC database for peer-reviewed works published between 2005 and 2015, using keywords: “data mining” “decision tree” “recursive partitioning” “classification tree” “regression tree” “association rule” “sequential covering” or “rule mining.” ERIC was chosen because sponsored by the US Department of Education, it is considered “the premier national bibliographic database of education

literature"(University of Pittsburgh University Library System, 2015). Of the nearly 180 articles that were suggested from this search, just 24 studies were education research that were *not* about online learning. Among them, the vast majority (21) used at least one kind of rule induction approach.

My review focused on the methodological aspects, examining research purpose, type of datasets used, model and rule induction approaches, evaluation and validation approaches, and types of inferences drawn from the results. I paid special attention to the extent to which rule induction seems to (or seems not to) add methodological value relative to more traditional statistical approaches.

2.5.1 Purposes and rationale for using rule induction

The orthodox education researchers utilized rule induction data mining for three main reasons. The vast majority used it to explore possible predictors of important educational outcomes. Nine of these studies explored factors associated with postsecondary student outcomes, including GPA, attrition, retention and graduation rate (Bailey, 2006; Delen, 2006, 2012; Guruler, Istanbullu, & Karahasan, 2010; Herzog, 2006; Kopiez, Weihs, Ligges, & Lee, 2006; Schumacher, Olinsky, Quinn, & Smith, 2010; Vandamme, Meskens, & Superby, 2007; Willett & Hom, 2007). Six studies explored outcomes of K12 students including junior high school academic achievement (Pai, Lyu, & Wang, 2010), reading achievement (Compton, Fuchs, Fuchs, & Bryant, 2006; Streifer & Schumann, 2005); science achievement (X. Liu & Ruiz, 2008; X. Liu & Whitford, 2011), and math achievement (Flores et al., 2013). Three studies explored predictors of students' non-cognitive outcomes, including elementary students' commitment to musical instrument learning (Faulkner et al., 2010), secondary students' internet dependence (Kayri & Gunuc, 2010) and

delinquency behavior (Y. C. Liu & Hsu, 2013). There were also studies on school disciplinary practices (Horner, Fireman, & Wang, 2010), alumni giving (Weerts & Ronca, 2009), and teacher education (Masunaga & Lewis, 2011).

Some researchers used data mining to examine very specific hypotheses about the relationships between specific independent variables and outcome variables. Most were related to student achievement, although one was about teacher education. Eykamp (2006) examined how students, if at all, are using AP course credits to their advantage in college. Compton et al. (2006) examined the relationship between seldom used reading diagnostic measures and first grade reading, and whether the former adds value to more commonly utilized measures. X. Liu and Whitford (2011) examined the relationship between opportunity to learn at home and science achievement. Flores et al. (2013) examined the relationship between computer use and math achievement, although they also explored other possible predictors at the student, classroom and school level. Kopiez et al. (2006) examined the hypothesis that musical sight reading expertise is necessary but not sufficient for sight-reading ability. Finally, Masunaga and Lewis (2011) examined the relationship between disposition towards teaching and success in student teaching.

In addition to applied work that were exploratory, and/or hypothesis-driven, there were two studies that were methodological. Herzog (2006) compared the predictive efficacy between traditional and various data mining approaches. Compton et al. (2006) analyzed whether there was added benefit to using classification tree analysis over logistic regression, and what approach might be most relevant for the field of higher education.

2.5.2 Algorithms used, and rationale

Ten studies used only a single tree algorithm to mine their data. The algorithms included CART (Delen, 2006; Masunaga & Lewis, 2011; Streifer & Schumann, 2005; Weerts & Ronca, 2009), C4.5 (Faulkner et al., 2010; X. Liu & Whitford, 2011), CHAID (Horner et al., 2010), Microsoft Decision Trees (Guruler et al., 2010), and an unspecified algorithm (Flores et al.).

There were many reasons that these authors, and others who used multiple algorithms, raised to justify mining data using decision trees. Most wanted to identify hidden, relevant patterns in a large dataset (Delen, 2006, 2012; Eykamp, 2006; Guruler et al., 2010; X. Liu & Ruiz, 2008; Streifer & Schumann, 2005; Weerts & Ronca, 2009; Willett & Hom, 2007), or wanted to know the most relevant predictors in a large dataset (Horner et al., 2010; Weerts & Ronca, 2009). Some mentioned that they used trees because they wanted to explore data using what has been regarded as a helpful alternative to traditional statistical methods. (Masunaga & Lewis, 2011; Streifer & Schumann, 2005), or because their attempt at using a traditional approach did not yield interesting/useful results (Eykamp, 2006). Some mentioned the practical benefits of decision trees as reasons for use, including the ease in data processing (Schumacher et al., 2010; Weerts & Ronca, 2009) and ease of interpretation and usefulness for practitioners and policy makers (Eykamp, 2006; Guruler et al., 2010; Weerts & Ronca, 2009). Some also appealed to the successful use of decision trees in other studies (Faulkner et al., 2010; X. Liu & Whitford, 2011), and curiosity of whether it would work for their data at hand (Streifer & Schumann, 2005). Some studies mentioned all of the aforementioned typical advantages of rule-based mining (Flores et al.; Vandamme et al., 2007), adding that decision trees and/or data mining methods are not often used in the field (Delen, 2012; Faulkner et al., 2010; Kopiez et al.).

Three studies used two different tree approaches for the same analysis. X. Liu and Ruiz (2008) and Kopiez et al. (2006) used a classification tree and a regression tree approach, presumably to account for errors that may be associated with data transformation. Kayri and Gunuc (2010) used CART and CHAID to compare different approaches, and presumably to increase statistical validity by controlling for algorithmic bias.

Five studies used one tree approach, and one or more non-tree classification approach for the same analysis. Delen (2012) used logistic regression and artificial neural networks in addition to C5, following the CRISP-DM recommendation to develop model using comparable analytical techniques. Compton et al. (2006) used CART and logistic regression to compare results, and to examined whether tree analyses provided different/additional insight. Eykamp (2006) supplemented regression with decision tree analyses when the former did not provide great insight. When the decision tree also turned out not to be so helpful, he used additional data mining approaches, including cluster analysis and neural network analysis. Schumacher et al. (2010) used neural networks, logistic regression and CART to replicate their previous study and contrast a new approach with those they had already taken. Vandamme et al. (2007) used ID3, neural network and linear discriminant analysis, presumably for some of the reasons mentioned by others, above.

Only two studies compared multiple tree and non-tree approaches for the same analysis. Herzog (2006) used three kinds of decision tree (CART, CHAID, C5.0), three kinds of neural network algorithms and regression, in order to be able to compare relative prediction accuracy of different approaches. Pai et al. (2010) used rough set theory, C4.5, ID3, CART and PART to examine efficacy of lesser known data mining approach (rough set theory) with better known approaches.

Finally, just one study used association rule mining. Association rule mining & time sequence analysis. Y. C. Liu and Hsu (2013) used association rule mining and time sequence analysis to predict adolescent behavior from counselor notes. Association rule mining was used because it allows for semi-automatic analysis of text, which are too large in volume to be coded manually.

Decision trees are clearly the most popular approach among orthodox education researchers who have ventured out into the field of Data Mining. As perhaps to be expected in early-adoption stage of data mining in the field, it appears that orthodox education researchers so far tend to use too few algorithms, and have insufficient justification for using the approaches that they do. Only one approach referenced to a data mining framework to justify their methodology, and only two used more than two decision tree approaches. Furthermore, only one of the studies in the sample used association rule mining and no studies used sequential covering, suggesting these approaches could certainly use more attention.

2.5.3 Datasets

The characteristics of the mined datasets are summarized in Table 4. Approximately half of the datasets analyzed were institutional data collected by universities, while about a quarter were survey results collected by researchers. The outcomes examined included proxies of educational achievement or retention (e.g., GPA, test scores, second year retention) and indicators of socioemotional wellness. Just two studies utilized large-scale US national datasets. The dataset sample size ranged from less than 100 to over 50,000 with half of the studies with sample sizes under 1000. The number of predictors ranged between 5 to a few hundred, with almost half of the studies examining 15 or fewer variables, and just five studies examining over 50 variables.

Table 4. Characteristics of datasets mined in orthodox education research

	Study	Dataset source	N	Outcome	# Pred
Higher education achievement and retention	Bailey, 2006	Information on IHEs from IPEDS	Up to 5771	Graduation rate	>1000
	Delen, 2012	Institutional data on students (US university)	6454	2nd fall registration Yes no 50-50	39
	Eykamp, 2006	Institutional data on students (US university)	9438	Time until degree completion	6
	Guruler, Istanbullu, & Karahasan, 2010	Institutional data on students (Turkish university)	3110	Dichotomized GPA (>=2.0, and >=3.0)	21
	Herzog, 2006	Institutional data on students (US university)	8081; 15457	1 year retention; Time to degree completion	40; 79
	Kopiez et al., 2006	Survey and test data on postsecondary piano students collected by researchers	52	Sight-reading proficiency	27
	Schumacher, Olinsky, Quinn, & Smith, 2010	Departmental data on students (business school actuarial department)	201	Retained in program vs changed major (~50-50)	5
	Vandamme, Meskens, & Superby, 2007	Institutional and survey data on students (Belgian medical school)	533	Risk of failure at university	375
	Willett & Hom, 2007	Institutional data on enrollments (US community college)	53753	Stay vs move (~1:2)	15
K12 achievement	Compton et al., 2006	Longitudinal (2-year), academic and demographic data collected by researchers	206	1st grade reading scores (dichotomized)	10
	Flores, Inan, & Lin	ELS:2002	2848	Probability of success in mathematics (high vs low)	~20
	X. Liu & Ruiz, 2008	NAEP (2005, 2000) and TIMMS (1995, 1999, 2003) data on student group performance	76	Student cohort performance on science items (satisfactory vs unsatisfactory)	5
	X. Liu & Whitford, 2011	PISA data (2006) on US students	5611	Science proficiency (dichotomized)	24
	Pai, Lyu, & Wang, 2010	Taiwan Educational Panel Survey 2004	500	Academic achievement	12
	Streifer & Schumann, 2005	District data on US middle school students	500	7th grade reading score	24
Non-cognitive	Faulkner Davidson, & McPherson, 2010	Survey data on Australian children's musical interests collected by researchers	139	1st year retention to instrument study (y/n); Practices regularly at same time (y/n)	75
	Kayri & Gunuc, 2010	Survey of randomly selected secondary school students in Turkey, collected by the researchers	754	Total score on the Internet Dependent Scale	7
	Y.-C. Liu & Hsu, 2013	High school counseling records housed in the	32908	n/a (association rule mining)	n/a (text mining)

Table 4 continued

	Study	Dataset source	N	Outcome	# Pred
		Education Administration of the Taiwan government			
Other	Horner, Fireman, & Wang, 2010	Survey and demographic data on US elementary school students collected by authors	1493	Presence of disciplinary actions (binary)	11
	Masunaga & Lewis, 2011	Survey on teacher disposition of US elementary school teacher candidates	277	Struggle with student teaching experience (binary)	15
	Weerts & Ronca, 2009	Institutional data on alumni of US institution	1441	Level of charitable contributions last year and lifetime	~250

2.5.4 Methods, validation, results and inferences

The data mining and validation approaches taken in these applied studies were somewhat varied, including omissions and errors that indicate that we are still in the early stages of adopting this methodology. More than half of the studies tended to use either a hold-out approach or n-fold cross validation, report confusion matrices, display trees and discuss the most predictive variables detected from the approach. However, over a third of the studies generated rules without mentioning how they validated the accuracy. Among those who used validation, a third did not discuss ruleset validity beyond overall accuracy or reported an erroneous conception of cross-validation. The validation of individual rules was infrequently reported or discussed, in terms of predictive accuracy or interestingness (exceptions included X. Liu & Ruiz, 2009, X. Liu & Whitford, 2011, and Weerts & Ronca, 2009). The imbalance in the outcome variables were often not taken into consideration, and unexpected rules were rarely if ever discovered most likely due to a combination of limitations in dataset richness and methodology (namely, screening for rules with high confidence and high support tend to eliminate unexpected rules).

2.6 SUMMARY AND IMPLICATIONS

As many have cautioned, we should be wary about blindly applying data mining methodologies to education research (Martin & Sherin, 2013; Reimann et al., 2014; Selwyn, 2015; Zhao & Luan, 2006). Particular attention must be paid to reducing threats to statistical and ecological validity of inferences from data mining. It is important to use multiple algorithms, cross-validate results and use alternative means to check the validity of any new inferences. It is also highly recommended to follow a data mining framework (Azevedo, 2008), to articulate background assumptions when preparing the data, and to not blindly search for "anything that sticks."

However, as many early-adopters of this methodology in education have pointed out, there also good reasons to believe that rule induction methods could improve educational research (Delen, 2012; Flores et al., 2013; Kopiez et al.; Vandamme et al., 2007). Rule induction approaches tend to be very flexible in terms of what kinds of variables they can include in their analysis. For example, most decision tree approaches can include continuous and discrete variables as predictors and outcome. Most algorithms can also accommodate cases with missing information. In addition, since rule induction approaches are non-parametric (i.e., does not require that we assume the data comes from a population that are distributed in a particular way), they tend to be robust to outliers, and the results tend to be easily understood, even to those who are not trained in statistics. It has also been suggested that the flexibility of decision tree methods tends to result in better models, particularly if the sample size is large enough such that the issue of model over-fitting can be avoided through a test and validation approach (Provost & Fawcett, 2013).

The biggest potential benefit to orthodox education research and practice appears to be their ability to uncover detect predictor-outcome relationships that could be unique to subgroups, and help discover new insights about relationships between variables. Factors that mediate student

learning are known to be numerous, and interrelated differently among different students. For example, how much a school or family emphasizes the importance of grades may have a different effect on student learning, depending on what specific messages are conveyed, how they are conveyed, and the students' temperament and past experiences with school. Discovering *this* type of complex relationship among variables is not straight-forwardly possible with more traditional statistical approaches such as regression, including all-possible-subsets regression.

However, this potential to uncover interesting subgroup characteristics within education datasets remains largely untapped. This is in part because of constraints in datasets that have been explored so far—with many of them being too small and/or not comprehensive enough to examine potentially interesting subgroup characteristics (Table 4, above). In addition, because of barriers to the adoption of data mining in education (Section 1.1.3), there has not been enough sustained thought on reasonable *methods* to find such subgroup characteristics, in a way that makes sense to the field. The following chapter attempts to contribute to the methodological front.

3.0 METHODS

The primary goal of the following empirical aspect of this project was to find illustrative example(s) in which rule induction methods, relative to regression approaches, (1) improved classification accuracy, and/or (2) offered new avenues of explanation through their unique ability to detect predictor-outcome relationships that could be unique to subgroups. A secondary goal was to identify some sound, practical and helpful way to incorporate rule induction into orthodox education research. As stated in the introduction, the intention was not to generalize from these cases that rule induction is always (or even often) useful to education research, but rather to be able to draw upon the experience and findings to deepen the discussion about its potential utility.

To accomplish these goals, I re-analyzed two regression studies on the National Educational Longitudinal Study of the Eight Grade Class of 1988 (NELS:88) dataset using rule induction approaches, and compared results across methods to identify whether, in what ways, and why the latter methodology might provide additional insights. I chose NELS:88 because it is a rich and varied dataset that has been analyzed thoroughly by other researchers, and because there are other similar datasets that can be used to examine the generalizability of inferences. The study on predictors of science achievement by Byrnes and Miller (2007) used hierarchical multiple regression, while the study on predictors of academic success of Black students by Thomas (2006) used logistic regression. The models and predictors used in each study are shown in Table 5 through Table 7. Both investigate research questions that are theoretically well grounded, but also exploratory in nature, which makes them conducive to inquiry through data mining. They also have a very different but relatively large sample size (15855 and 1176, respectively), and include many predictors. Each also includes a careful discussion of their final models.

Thomas explored the NELS:88 dataset using logistic regression, to better understand factors that were related to achievement differences among Black students. She operationalized "high achievers" as those whose 12th grade achievement test scores (average of math, reading, science, and history) were in the top quartile, among all the Black students in the sample. Similarly, those whose average score was in the bottom quartile, were considered "low achievers." Those who were missing on any of the test scores were considered to have earned the average score for the group. Drawing from a College Board report by a national task force on minority high achievement (Cota-Robles & Gordan, 1999), Thomas hypothesized that lower-achieving Black students would: (1) have fewer educational resources, including parental financial support; (2) have parents who are less involved in their schooling; (3) associate with 'bad" peers[8]; (4) participate in fewer cultural activities; and (5) attend schools that are less conducive to learning. She hypothesized that the converse would be true for high-achieving students.

To test these hypotheses, Thomas examined cross tabulations of achievement and related factors, and conducted logistic regressions, where low or high achievement was predicted using student factors, family factors, peer factors, community factors, or school factors. In addition, she modeled achievement using all the factors that were statistically significant in the first series of modeling. The latter process helped shed light to other research questions, namely, to understand the factors that "best predict the successful or unsuccessful adaptation to school for Black students" and to understand the extent to which the students' academic achievement can be explained by family, school, peer and community variables (p.117). For this project, I focus on re-visiting her analysis of predictors of high-achieving Black students.

[8] Thomas (2006) did not fully describe what a "bad" peer was, but operationalized it as 8th graders' peers who placed importance in parties, sex, drugs and alcohol. She operationalized "good" peers as those who placed importance in academics and educational attainment. See Appendix A for full description of the variables.

Byrnes and Miller (2007) examined the NELS:88 dataset to test hypotheses regarding predictors of math and science achievement. They were particularly motivated to examine a general and comprehensive theory of learning using a large number of variables, in contrast to most quantitative studies in education that tend to test a narrow and specific aspect of learning using very few variables (their review found that most studies included no more than 8 variables in their model). They hypothesized that there are two necessary conditions to student achievement: opportunities for the student to learn (including inside and outside of school), and propensity for the student to learn (including ability, willingness to learn and self-regulatory skills). In addition, they surmised that factors more distal to the learning experience, including socioeconomic status, self-and parental expectations regarding achievement, parental values and prior achievement can directly or indirectly affect achievement. Finally, Byrnes and Miller hypothesized that race and gender would not predict achievement when all factors in this model are controlled. The "opportunity-propensity model of achievement" is diagrammed in Figure 4.

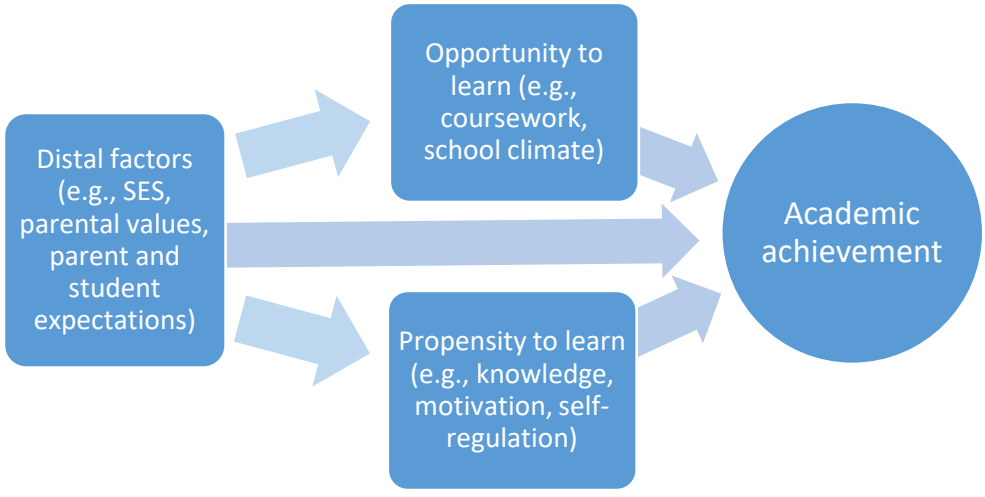


Figure 4. Opportunity-propensity model of achievement examined by Byrnes and Miller (2007)

Adapted from Byrnes & Miller (2007, p.602).

To understand the relative contribution of opportunity, propensity and distal factors to achievement, Byrnes and Miller conducted hierarchical regression that predicted 10th and 12th grade math and science achievement scores from 25 NELS variables. They also ran structural equation models to assess possible causal relationships between the cluster of factors. For this project, I focused on re-analyzing their examination of 12th grade math achievement using hierarchical regression.

Table 5. Models to be replicated

Thomas, 2006, Logistic regression model

- *High academic achievement* = $e^{(\beta'X)} / (1 + e^{(\beta'X)})$, where $\beta'X = \beta_0 + \beta_1$ (Parent education) + β_2 (1991 income) + β_3 (homework hours in school) + β_4 (homework hours out of school) + β_5 (household resources) + β_6 (religious school) + β_7 (parental involvement) + β_8 (parents expect college) + β_9 (good peers) + β_{10} (bad peers) + β_{11} (peers expect college) + β_{12} (cultural activities) + β_{13} (neighborhood diversity) + β_{14} (percent free lunch) + β_{15} (school climate) + β_{16} (feels unsafe in school) + β_{17} (disruptions in school) + β_{18} (number of Black teachers)+ β_{19} () + β_{20} ()+ ε_0.

Byrnes and Miller, 2007, Multiple linear regression (hierarchical) model

- *12^{th} grade math achievement* = $\beta_0 + \beta_{D1}$ (8^{th} grade SES) + β_{D2} (Parent expectations) + β_{D3} (Student expectations) + β_{D4} (Middle school GPA) + β_{O1} (Gen math .5yr) + β_{O2} (Gen math 1yr) + β_{O3} (Gen math 1.5-2yr) + β_{O4} (Geometry .5yr) + β_{O5} (Geometry 1yr) + β_{O6} (Geometry 1.5-2yr) + β_{O7} (Algebra II .5yr) + β_{O8} (Algebra II 1yr) + β_{O9} (Algebra II 1.5-2yr) + β_{O19} (Perception of math emphasis) + β_{O21} (Perception of teacher responsiveness .5yr) + β_{P1} (Pre-9^{th} math achievement) + β_{P2} (9^{th}/10^{th} math GPA) + β_{P5} (HS graduation efficacy) + β_{P6} (Plans to take SAT) + β_{P7} (math self-concept) + β_{Demo1} (Gender) + β_{Demo2} (Black) + β_{Demo3} (Hispanic) + β_{Demo4} (Asian) + β_{Demo5} (Native American) + ε_0.

Note: Byrnes and Miller included distal (D) variables first, followed by opportunity variables (O), propensity (P) and demographics (Demo).

Table 6. Variables used in Thomas, 2006

Outcome
High achievement (**F22XRC, F22XMC, F22XSC, F22XHC**)
Student characteristics
1. Sex (**F4SEX**) 2. Parental education (**BYPARED**) 3. Number of siblings 8th grader has (**BYP3A**) 4. Parents' marital status (**BYPARMAR**) 5. Single parent (**F2P7**) 6. Income from all sources 1991 (**F2P74**) 7. Hours of homework in school (**F2S25F1**) 8. Hours of homework out of school (**F2S25F2**)
Family variables
1. Household resources (**F2N12A, B, D, E, F, H, M, O**) 2. Parents pay for tutor (**BYP82D**) 3. Private school (**G12CTRL1**) 4. Religious school (**G12CTRL1**) 5. Autonomy (**F2S98A, B, C, D**) 6. Parental involvement in school (**BYP59A, B, C, D, E**) 7. Parents expect college (**F2S41A, F2S41B**)
Peer variables
1. Good peers (**F2S68A, B, D, F, H**) 2. Bad peers (**F2S68M, N, O, P**) 3. Person student admires most, among all personal acquaintances, is intelligent (**F1S71D**) 4. Peers expect most important thing for you to do after HS is college (**F2S41C**)
Community variables
1. Activities outside of school (**BYP60A, B, C, D, E, F, G, H**) 2. Student's cultural activities (**BYP61AB, BB, CB, DB, EB**) 3. Neighborhood safety (**F2P60**) 4. Neighborhood diversity (**F4JRDVA**)
School variables
1. Public school (**G8CTRL**) 2. Urbanicity of school (**G8URBAN**) 3. Percent minority in school (**G8MINOR**) 4. Percent receiving free lunch in school (**G8LUNCH**) 5. School climate (**BYS58A, B, C, D, E, F, G, H, I, J, K**) 6. Student assigned for racial/ethnic composition (**BYSC24C**) 7. Student feels unsafe in school (**BYS59K**) 8. Disruptions in school prevent learning (**BYS59L**) 9. Student-teacher ratio (**BYRATIO**) 10. Number of Black, non-Hispanic teacher (**BYSC20D**)

Table 7. Variables used in Byrnes & Miller, 2007

Sample selection flag and weight

BY, F1 and F2 participant, and non-dropout at F1 (**F2TRP1FL=1, F2F1QFLG=1, F2PNLWT)**

Outcome variables

12th grade math (**F22XMIRR**)

Distal factors (D)

1. 8th grade SES (**BYSES**)
2. Parent expectations for child in 8th grade (**BYP76**)
3. Student expectations in 8th grade (**BYS45**)
4. Middle school GPA (**BYGRADS**)

Opportunity factors (O)

1. General math courses .5 year (**F1S22A**)
2. General math courses 1 year (**F1S22A**)
3. General math courses 1.5-2 years (**F1S22A**)
4. Geometry courses .5 year (**F1S22D**)
5. Geometry courses 1 year (**F1S22D**)
6. Geometry courses 1.5-2 years (**F1S22D**)
7. Algebra II courses .5 year (**F1S22E**)
8. Algebra II courses 1 year (**F1S22E**)
9. Algebra II courses 1.5-2 years (**F1S22E**)
10. General science courses .5 year (**F1S23A**)
11. General science courses 1 year (**F1S23A**)
12. General science courses 1.5-2 years (**F1S23A**)
13. Biology courses .5 year (**F1S23C**)
14. Biology courses 1 year (**F1S23C**)
15. Biology courses 1.5-2 years (**F1S23C**)
16. Chemistry courses .5 year (**F1S23E**)
17. Chemistry courses 1 year (**F1S23E**)
18. Chemistry courses 1.5-2 years (**F1S23E**)
19. Student perception of math emphasis (**F1S31A, B, C, D, E**)
20. Student perception of science emphasis (**F1S30A, B, C, D, E**)
21. Student perception of teacher responsiveness (**F1S7A, D, G, H, I, J**, **L**)

Propensity factors (P)

1. Math achievement before start of 9th grade (**BYTXMIRR**)
2. Math GPA in 9th and 10th grades (**F1S39A**)
3. Science achievement before start of 9th grade (**BYTXSIRR**)
4. Science GPA in 9th and 10th grades (**F1S39D**)
5. Efficacy for graduating high school (**F1S18A**)
6. Plans to take SAT (**F1S50B**)
7. Math self-concept (**F1S63D, J, Q, S**)

Demographic factors (DEMO)

1. Gender (**SEX, F1SEX, F2SEX**)
2. Race/ethnicity dummy (Black) (**RACE, F1RACE, F2RACE1**)
3. Race/ethnicity dummy (Hispanic) (**same**)
4. Race/ethnicity dummy (Asian) (**same**)
5. Race/ethnicity dummy (Native American) (**same**)

Three types of analyses were conducted per study. ***(1) Replication.*** I first replicated their analyses using their respective methodologies (hierarchical regression, logistic regression). ***(2) Rule induction using only the study predictors.*** I then analyzed their research questions using several rule induction algorithms *including only the predictors in their final models*. ***(3) Rule induction using all reasonable predictors.*** Third, I analyzed their questions using the same rule induction algorithms and outcome variable, but by adhering to the CRISP-DM process of applying data mining (Chapman et al., 2000) and including all reasonable and available predictors as potential variables for the model.

I used the following ruleset induction approaches: RIPPER (Cohen, 1995), CBA (B. Liu et al., 1998; B. Liu, Ma, et al., 2000), PART (Frank & Witten, 1998), CART (classification and regression tree, Breiman et al., 1984), C4.5 (Quinlan, 1993) and See5/C5.0 (Quinlan, 2013) and QUEST (Quick Unbiased Efficient Statistical Tree, Loh & Shih, 1997). Apriori (Agrawal et al., 1993) was also used, adjusted so that it only generates rules with the outcome variable as a consequent. These algorithms were chosen because of their popularity, availability and diversity in the way they generate rules. Random forests, bagged version of CART and boosted version of C4.5 were included (although they do not result in rules), just as a way to further contextualize the predictive accuracies of the aforementioned algorithms.

Data had to be prepared slightly differently depending on what each algorithm accepts, as detailed in Section 3.2.3. Missing data in the outcome variable was dealt in accordance with each study. Data missing in the predictors were managed by relying on the missing data feature of each algorithm (e.g., CART uses a surrogate predictor to split the missing value), or if such features are not available, by adhering to the methodology of the studies. When possible, misclassification

costs were adjusted so that the cost of the false positive rate to false negative rate was proportional to the ratio of positive and negative sample sizes.

I used R Version 3.2.2 through 3.3.2 (R Core Team, 2016) and RStudio Version 0.99.896 through 1.0.136 (RStudio Team, 2015) for most of the data mining, and SAS software Version 9.4 for data cleaning and other analyses. R was chosen because it is widely used, open-source, free-of-charge, flexible and comprehensive in the analyses it allows. The specific methods and packages used are summarized in Table 18 in Section 3.2.3. The QUEST algorithm was not available as an R package so SPSS Version 22 was used. Most of the analysis was conducted on a SONY VAIO laptop v.1511, with an Intel(R) CORE^{TM}i7 processor (2GHz), 8GB RAM and 64-bit operating system. Analyses that required larger memory (namely, CBA using the small dataset for study 2, and random forest and C5.0/boosted C5.0 using the larger dataset for study 2) were conducted using the Bridges system of the Extreme Science and Engineering Discovery Environment (XSEDE) at the Pittsburgh Supercomputing Center (Nystrom, Levine, Roskies, & Scott, 2015; Towns et al., 2014). For study 2 using the large dataset, C5.0 converged with 1250GB of RAM with a 5-hour wall time, while the random forest converged with 24 hours wall time and 2000GB of RAM. A few algorithms including CBA in large datasets and bagged CART for the large dataset study 2 did not converge or threw an error, even with the supercomputer.

I used a hold-out approach to avoid model over-fit, using 70% of the data for model creation and the remaining for testing. I chose this over a *k*-fold cross-validation approach, since the sample size allowed for it, it was easier and faster to implement this approach consistently across all rule induction approaches, and because the methods and results are much more observable. The 70-30 random split was stratified across outcome, and the same split was used

across algorithms. Exploration of model parameters, when needed, were done using a holdout approach within the test set.

Across analyses of each study, I compared the predictive accuracy (confusion matrices, F-ratio), relative importance of predictors included in the model, and examined potentially interesting predictor-outcome relationships that could be unique among subgroups.

3.1 REPLICATION

The NELS:88 data and SAS formatting syntax were obtained from ICPSR in October 2015. Creation of the final datasets involved extracting and merging relevant information from the base year (United States Department of Education National Center for Education Statistics, 2006a), the first follow up (United States Department of Education National Center for Education Statistics, 1999), and the second follow up (United States Department of Education National Center for Education Statistics, 1995) studies. The specific datasets and variables needed for replication are described in Table 8 and Table 9. Datasets were merged by student ID and/or school ID.

Table 8. Datasets and variables required for replication of Thomas (2006)

Study	Datasets	Variables
Base year	Student, parent, school	STU_ID SCH_ID RACE SEX BYPARED BYP3A BYPARMAR BYP82D BYP59A BYP59B BYP59C BYP59D BYP59E BYP60A BYP60B BYP60C BYP60D BYP60E BYP60F BYP60G BYP60H BYP61AB BYP61BB BYP61CB BYP61DB BYP61EB G8CTRL G8URBAN G8MINOR G8LUNCH BYS58A BYS58B BYS58C BYS58D BYS58E BYS58F BYS58G BYS58H BYS58I BYS58J BYS58K BYS59K BYS59L BYPARTIC
First follow-up	Student	STU_ID F1RACE SEX F1S71D
Second follow-up	Student part 1, student part 2, parent	STU_ID F2RACE1 SEX F22XRTH F22XMTH F22XSTH F22XHTH F2P7 F2P74 F2S25F1 F2S25F2 F2N12A F2N12B F2N12D F2N12E F2N12F F2N12H F2N12M F2N12O G12CTRL1 F2S98A F2S98B F2S98C F2S98D F2S41A F2S41B F2S68A F2S68B F2S68D F2S68F F2S68H F2S68M F2S68N F2S68O F2S68P F2S41C F2P60

Table 9. Datasets and variables required for replication of Byrnes and Miller (2007)

Study	Datasets	Variables
Base year	Student, parent	STU_ID SCH_ID RACE SEX BYSES BYP76 BYS45 BYGRADS BYTXMIRR BYTXSIRR BYTXMFS BYTXSFS BYTXMIRS BYTXMSTD BYTXSIRS BYTXSSTD
First follow-up	Student	STU_ID F1SEX F1RACE F1S22A F1S22D F1S22E F1S23A F1S23C F1S23E F1S31A F1S31B F1S31C F1S31D F1S31E F1S30A F1S30B F1S30C F1S30D F1S30E F1S7A F1S7D F1S7G F1S7H F1S7I F1S7J F1S7L F1S39A F1S39D F1S18A F1S50B F1S63D F1S63J F1S63Q F1S63S F1TXMIRR F1TXSIRR
Second follow-up	Student part 1, student part 2	STU_ID F2SEX F2RACE1 F2TRP1FL F2F1QFLG F2BYQFLG F2F1QFLG F2PNLWT F22XMIRR F22XSIRR

Because I identified a few errors in Thomas' data selection and cleaning process, and I did not want to replicate those errors, I decided to select and clean the data in the way I thought was

best, see if it the replication still matched Thomas' main results, and proceed with data mining only if it did. The variable selection and transformation processes for replication of Thomas' (2006) study are detailed in Appendix A. My process departed from Thomas' in a few ways. The main difference was that my final sample size was 1223 Black students who had data on at least one of the four academic tests and had participated in all three studies, in contrast to Thomas whose sample of 1176 Black students included 385 (30%) who were missing the outcome variable, and students who had not participated in some of the relevant study waves. I obtained this sample by using the second follow-up or earlier datasets, rather than the fourth follow-up dataset as Thomas had done (the second follow-up dataset contained nearly all variables for her models, and had double the sample size). 96.5% of the students in my sample had all four test scores, while 2% were missing 1 score, and the remaining 1.5% were missing two or more scores. The smaller difference was that I used the raw IRT theta scores as the basis for the outcome variable, instead of the IRT centile scores, because the latter were not yet available in the F2 dataset. In addition, the racial composition of the students' neighborhood could not be included, since this variable was not available until the 4th follow-up.

To accommodate missing data on the 20 predictors, I used multiple imputation rather than using Thomas' method (of using the mean values of the male/female groups), so that the missing values may more closely represent the values that the students may actually have had. I assumed, as Thomas had to have done, that data was missing at random (i.e., the missingness to be independent of the missing values of other variables) and therefore suitable for multiple imputation. Five rounds of imputation were conducted using PROC MI. Missing variables were computed for each variable separately (i.e., a fully conditional specification imputation) using logistic or multinomial logistic regression for categorical variables, and linear regression for

continuous variables. The amount of missing values for each variable and descriptive statistics before and after imputation, are displayed in Table 10 and Table 11. Selected correlations between variables are presented in Table 12 through Table 14.

The imputed datasets were used to create composite variables per Thomas' specifications, and logistic regression parameters were estimated for each sample run, for each of her five models. The independent variables included for each model, including transformations, are described in Appendix A. These mirrored Thomas' approach, except that they attempt to remedy coding errors, excluded the racial composition of the students' neighborhood as explained above, and excluded one nominal variable (parents' marital status, 6 categories) that Thomas included in her first model as a numeric variable. I excluded this variable, rather than including a dummy-coded version, since the model already included a similar variable ("single parent") that accounted for much of the variance of the nominal variable. For each model, the estimates from the imputed samples were combined using PROC MIANALYE. The results from this were compared with that of Thomas, to see that there was a substantial amount of overlap in the general conclusions, to justify proceeding with the data mining analyses.

Table 10. Descriptive statistics of variables, by sex, after replication of Thomas (2006) protocol, without multiple imputation of missing data on predictors.

	Female (N = 641)					Male (N = 582)				
Variable	*M*	*SD*	*Min*	*Max*	*% Miss*	*M*	*SD*	*Min*	*Max*	*% Miss*
Mean centile score	47.27	8.38	30.19	72.29	0	47.28	8.01	24.46	71.26	0
High achieving	.26	.44	0	1	0	.24	.43	0	1	0
Parent's highest level of education	2.69	1.08	1	6	2	2.84	1.13	1	6	2
Number of siblings	2.71	1.84	0	6	9	2.54	1.75	0	6	10
Single parent	.49	.50	0	1	10	.46	.50	0	1	10
Total family income - all sources '91	8.37	2.86	1	15	14	8.77	2.76	1	15	13
Time spent on homework in school	3.01	1.98	0	8	13	3.09	2.03	0	8	15
Time on homework out of school	3.27	1.85	0	8	13	3.15	2.02	0	8	15
Household resources	5.29	1.61	0	8	2	5.19	1.80	0	8	3
Parents pay for tutor	.02	.16	0	1	0	.03	.17	0	1	0
Private school	.17	.38	0	1	0	.15	.36	0	1	0
Religious school	.08	.28	0	1	0	.05	.21	0	1	0
Student autonomy	3.34	.94	1	5	20	3.58	1.00	1	5	27
Parent involvement in school	2.99	1.27	0	6	12	2.93	1.26	0	6	12
At least 1 parent expects college	.75	.43	0	1	0	.64	.48	0	1	0
Good peers	2.62	.44	1	3	15	2.43	.50	1	3	22
Bad peers	1.47	.44	1	3	15	1.75	.47	1	3	23
Person student admires is intelligent	.78	.41	0	1	12	.71	.46	0	1	20
Peers expect college	.57	.50	0	1	10	.53	.50	0	1	12
Activities outside of school in 8th gr	1.17	1.40	0	8	10	.87	1.12	0	6	10
Student's cultural activities in 8th gr	2.63	1.68	0	5	11	2.58	1.73	0	5	11
How safe is neighborhood	1.74	.69	1	4	10	1.72	.74	1	4	10
Public school	.88	.32	0	1	0	.91	.29	0	1	0
Urban school	.39	.49	0	1	0	.37	.48	0	1	0
% Minority in school	5.02	1.60	0	8	0	4.94	1.66	0	8	0
% Free lunch in school	4.34	1.97	0	8	0	4.19	2.07	0	8	0
School climate	3.59	3.34	0	11	0	2.98	3.28	0	11	0
Assigned for racial/ethnic composition	.20	.40	0	1	0	.19	.39	0	1	0

Table 10 continued

Variable	Female (N = 641)					Male (N = 582)				
	M	*SD*	*Min*	*Max*	*% Miss*	*M*	*SD*	*Min*	*Max*	*% Miss*
Student feels unsafe in school	.15	.36	0	1	5	.15	.36	0	1	7
Disruptions in school	.51	.50	0	1	0	.49	.50	0	1	0
Student-teacher ratio	18.10	4.31	10	30	1	17.47	4.12	10	30	1
Black/Hispanic teachers	3.77	1.88	0	6	3	3.74	1.85	0	6	3

%Miss = % of sample that were missing values

Table 11. Descriptive statistics of variables, by sex, after replication of Thomas (2006) protocol, after multiple imputation of missing data on predictors.

	Female (N = 641 x 5)				Male (N = 582 x 5)			
Variable	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Mean centile score	47.27	8.38	3.19	72.29	47.28	8.00	24.46	71.26
High achieving	.26	.44	0	1	.24	.43	0	1
Parent's highest level of education	2.69	1.08	1	6	2.84	1.13	1	6
Number of siblings	2.74	1.84	0	6	2.56	1.74	0	6
Parent's marital status	4.48	1.90	1	6	4.46	1.89	1	6
Single parent	.49	.50	0	1	.47	.50	0	1
Total family income-all sources '91	8.39	2.87	1	15	8.74	2.77	1	15
Time spent on homework in school	3.04	2.02	0	8	3.10	2.07	0	8
Time on homework out of school	3.30	1.97	0	8	3.18	2.09	0	8
Household resources	5.39	1.56	0	8	5.39	1.65	0	8
Parents pay for tutor	.04	.20	0	1	.06	.24	0	1
Private school	.17	.38	0	1	.15	.36	0	1
Religious school	.08	.28	0	1	.05	.21	0	1
Student autonomy	3.34	.90	1	5	3.54	.94	1	5
Parent involvement in school	2.85	1.36	0	6	2.78	1.34	0	6
At least 1 parent expects college	.80	.40	0	1	.71	.45	0	1
Good peers	2.57	.49	1	3	2.35	.56	1	3
Bad peers	1.53	.48	1	3	1.81	.50	1	3
Person student admires is intelligent	.77	.42	0	1	.69	.46	0	1
Peers expect college	.54	.50	0	1	.51	.50	0	1
Activities outside of school in 8th gr	1.62	1.87	0	8	1.26	1.65	0	8
Student's cultural activities in 8th gr	2.76	1.67	0	5	2.71	1.73	0	5
How safe is neighborhood	1.74	.69	1	4	1.72	.74	1	4
Public school	.88	.32	0	1	.91	.29	0	1
Urban school	.39	.49	0	1	.37	.48	0	1
% Minority in school	5.02	1.60	0	8	4.94	1.66	0	8
% Free lunch in school	4.34	1.96	0	8	4.19	2.07	0	8
School climate	3.76	3.29	0	11	3.25	3.24	0	11
Assigned for racial/ethnic composition	.20	.40	0	1	.20	.40	0	1
Student feels unsafe in school	.16	.36	0	1	.16	.36	0	1
Disruptions in school prevent learning	.54	.50	0	1	.53	.50	0	1
Student-teacher ratio	18.10	4.31	10	30	17.47	4.10	10	30
Black/Hispanic teachers	3.76	1.88	0	6	3.74	1.86	0	6

Table 12. Intercorrelations of Thomas (2007) variables part 1

Variable	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*	*(10)*	*(11)*
(1) Mean centile score	—	.79	.00	.39	-.09	-.10	.30	.16	.24	.24	.04
(2) High achieving		—	.02	.33	-.07	-.10	.27	.12	.21	.19	.02
(3) Female			—	-.07	.05	.02	-.07	-.02	.03	.03	-.02
(4) Parent's education				—	-.19	-.15	.48	.06	.16	.30	.12
(5) Number of siblings					—	-.01	-.22	.02	-.01	-.05	.00
(6) Single parent						—	-.45	.02	.03	-.15	-.03
(7) Total family income							—	-.01	.09	.31	.09
(8) Time on hw in school								—	.47	.03	.04
(9) Time on hw out of school									—	.13	.05
(10) Household resources										—	.04
(11) Parents pay for tutor											—
(12) Private school	.02	.05	.02	.10	-.03	.07	.04	.01	.15	.03	.02
(13) Religious school	.14	.14	.07	.18	-.08	.01	.12	-.01	.08	.11	.04
(14) Student autonomy	.02	.04	-.12	-.01	.04	.07	-.02	.00	-.04	-.01	-.03
(15) Parent sch involvement	.19	.15	.02	.30	-.17	-.13	.27	.01	.07	.20	.06
(16) Parent expects college	.29	.19	.11	.24	-.10	-.03	.19	.10	.11	.17	.04
(17) Good peers	.05	.02	.19	.03	.01	-.04	.03	.01	.06	.08	-.02
(18) Bad peers	-.08	-.05	-.29	-.01	-.05	.00	.03	-.03	-.05	.00	.05
(19) Person S admires is intelligent	.12	.09	.09	.10	-.12	.02	.08	.07	.11	.13	-.01
(20) Peers expect college	.19	.13	.04	.11	-.03	-.02	.09	.08	.15	.05	-.01
(21) Activities outside of school in 8th gr	.21	.19	.12	.26	-.03	-.01	.19	.06	.12	.20	.12
(22) S's cultural activities gr8	.26	.22	.01	.36	-.11	-.06	.30	.06	.11	.27	.08
(23) How safe is neighborhood	-.03	.01	.01	-.11	.06	.10	-.09	.00	-.02	-.09	-.05
(24) Public school	-.25	-.21	-.04	-.28	.13	-.01	-.20	-.01	-.13	-.16	-.05
(25) Urban school	.03	.04	.02	.10	-.07	.02	.07	.04	.05	-.02	.04
(26) % Minority in school	-.13	-.14	.02	-.11	.03	.07	-.07	-.06	-.07	-.07	.01
(27) % Free lunch in school	-.25	-.23	.04	-.25	.14	.08	-.25	-.05	-.13	-.16	-.06
(28) School climate	.06	.03	.09	.04	.06	-.05	.01	.04	-.02	.02	.00
(29) Assigned for racial/ethnic composition	-.02	-.05	.01	-.01	-.01	.02	.03	.06	.01	-.01	-.01
(30) S feels unsafe in school	-.21	-.14	.00	-.09	.04	-.03	-.05	-.06	-.08	-.01	.02
(31) Disruptions	-.17	-.14	.02	-.12	.05	-.01	-.08	-.06	-.09	-.05	-.04
(32) Student-teacher ratio	-.04	-.01	.07	-.02	-.03	.03	-.04	-.03	-.01	.02	.01
(33) Black/Hispanic teachers	-.18	-.18	.01	-.16	.05	.01	-.13	-.06	-.11	-.11	-.01

Table 13. Intercorrelations of Thomas (2007) variables part 2

Variable	*(12)*	*(13)*	*(14)*	*(15)*	*(16)*	*(17)*	*(18)*	*(19)*	*(20)*	*(21)*	*(22)*
(1) Mean centile score	.02	.14	.02	.19	.29	.05	-.08	.12	.19	.21	.26
(2) High achieving	.05	.14	.04	.15	.19	.02	-.05	.09	.13	.19	.22
(3) Female	.02	.07	-.12	.02	.11	.19	-.29	.09	.04	.12	.01
(4) Parent's education	.10	.18	-.01	.30	.24	.03	-.01	.10	.11	.26	.36
(5) Number of siblings	-.03	-.08	.04	-.17	-.10	.01	-.05	-.12	-.03	-.03	-.11
(6) Single parent	.07	.01	.07	-.13	-.03	-.04	.00	.02	-.02	-.01	-.06
(7) Total family income	.04	.12	-.02	.27	.19	.03	.03	.08	.09	.19	.30
(8) Time on hw in school	.01	-.01	.00	.01	.10	.01	-.03	.07	.08	.06	.06
(9) Time on hw out of school	.15	.08	-.04	.07	.11	.06	-.05	.11	.15	.12	.11
(10) Household resources	.03	.11	-.01	.20	.17	.08	.00	.13	.05	.20	.27
(11) Parents pay for tutor	.02	.04	-.03	.06	.04	-.02	.05	-.01	-.01	.12	.08
(12) Private school	—	.60	-.07	.08	-.20	.04	-.02	.05	.10	.06	.09
(13) Religious school		—	-.08	.16	.15	.06	-.02	.07	.09	.12	.16
(14) Student autonomy			—	-.08	-.08	-.09	.11	.06	-.04	-.03	-.01
(15) Parent sch involvement				—	.11	.10	-.03	.11	.08	.22	.37
(16) Parent expects college					—	.09	-.07	.11	.38	.17	.19
(17) Good peers						—	-.24	.14	.21	.12	.07
(18) Bad peers							—	-.09	-.09	-.11	-.03
(19) Person S admires is intelligent								—	.10	.14	.09
(20) Peers expect college									—	.08	.13
(21) Activities outside of school in 8th gr										—	.37
(22) S's cultural activities gr8											
(23) How safe is neighborhood	.09	.01	.02	-.04	-.05	.02	-.02	-.01	-.02	-.04	-.06
(24) Public school	-.57	-.73	.04	-.23	-.17	-.05	-.01	-.09	-.11	-.18	-.24
(25) Urban school	.10	.02	-.06	.07	.04	.02	-.06	.06	.07	.09	.18
(26) % Minority in school	.06	.10	-.06	-.01	-.01	.08	-.07	.01	-.02	-.02	-.03
(27) % Free lunch in school	-.23	-.29	-.02	-.16	-.11	.04	-.08	-.07	-.11	-.15	-.19
(28) School climate	-.07	-.06	.08	-.04	.06	-.06	.02	.02	.00	.03	.00
(29) Assigned for racial/ethnic composition	-.06	-.13	.02	.02	-.03	.01	-.02	.04	.02	.02	-.01
(30) S feels unsafe in school	-.02	-.05	.06	-.05	-.11	.01	.06	-.12	-.08	-.04	-.03
(31) Disruptions	-.04	-.05	-.06	-.04	-.02	.00	-.05	-.03	.02	-.01	-.04
(32) Student-teacher ratio	.14	.30	-.10	.02	.05	.03	.02	.03	.02	.00	-.01
(33) Black/Hispanic teachers	-.19	-.21	-.01	-.07	-.05	.06	-.06	.01	-.06	-.05	-.08

Table 14. Intercorrelations of Thomas (2007) variables part 3

Variable	*(23)*	*(24)*	*(25)*	*(26)*	*(27)*	*(28)*	*(29)*	*(30)*	*(31)*	*(32)*	*(33)*
(1) Mean centile score	-.25	.03	-.13	-.25	.06	-.02	-.21	-.17	-.04	-.18	-.25
(2) High achieving	-.21	.04	-.14	-.23	.03	-.05	-.14	-.14	-.01	-.18	-.21
(3) Female	-.04	.02	.02	.04	.09	.01	.00	.02	.07	.01	-.04
(4) Parent's education	-.28	.10	-.11	-.25	.04	-.01	-.09	-.12	-.02	-.16	-.28
(5) Number of siblings	.13	-.07	.03	.14	.06	-.01	.04	.05	-.03	.05	.13
(6) Single parent	-.01	.02	.07	.08	-.05	.02	-.03	-.01	.03	.01	-.01
(7) Total family income	-.20	.07	-.07	-.25	.01	.03	-.05	-.08	-.04	-.13	-.20
(8) Time on hw in school	-.01	.04	-.06	-.05	.04	.06	-.06	-.06	-.03	-.06	-.01
(9) Time on hw out of school	-.13	.05	-.07	-.13	-.02	.01	-.08	-.09	-.01	-.11	-.13
(10) Household resources	-.16	-.02	-.07	-.16	.02	-.01	-.01	-.05	.02	-.11	-.16
(11) Parents pay for tutor	-.05	.04	.01	-.06	.00	-.01	.02	-.04	.01	-.01	-.05
(12) Private school	-.57	.10	.06	-.23	-.07	-.06	-.02	-.04	.14	-.19	-.57
(13) Religious school	-.73	.02	.10	-.29	-.06	-.13	-.05	-.05	.30	-.21	-.73
(14) Student autonomy	.04	-.06	-.06	-.02	.08	.02	.06	-.06	-.10	-.01	.04
(15) Parent sch involvement	-.23	.07	-.01	-.16	-.04	.02	-.05	-.04	.02	-.07	-.23
(16) Parent expects college	-.17	.04	-.01	-.11	.06	-.03	-.11	-.02	.05	-.05	-.17
(17) Good peers	-.05	.02	.08	.04	-.06	.01	.01	.00	.03	.06	-.05
(18) Bad peers	-.01	-.06	-.07	-.08	.02	-.02	.06	-.05	.02	-.06	-.01
(19) Person S admires is intelligent	-.09	.06	.01	-.07	.02	.04	-.12	-.03	.03	.01	-.09
(20) Peers expect college	-.11	.07	-.02	-.11	.00	.02	-.08	.02	.02	-.06	-.11
(21) Activities outside of school in 8th gr	-.18	.09	-.02	-.15	.03	.02	-.04	-.01	.00	-.05	-.18
(22) S's cultural activities gr8	-.24	.18	-.03	-.19	.00	-.01	-.03	-.04	-.01	-.08	-.24
(23) How safe is neighborhood	.01	.19	.19	.06	.03	.10	.02	-.06	.03	.17	.01
(24) Public school	—	-.07	.00	.40	.08	.17	.06	.08	-.21	.37	—
(25) Urban school		—	.33	.15	.05	.17	.03	-.03	.09	.23	-.07
(26) % Minority in school			—	.47	.05	.09	.04	.01	.21	.59	.00
(27) % Free lunch in school				—	.03	.12	.10	.11	-.01	.38	.40
(28) School climate					—	.05	.08	.11	-.02	.10	.08
(29) Assigned for racial/ethnic composition						—	.00	-.03	-.04	.18	.17
(30) S feels unsafe in school							—	.18	-.01	.06	.06
(31) Disruptions								—	.01	.06	.08
(32) Student-teacher ratio									—	.00	-.21
(33) Black/Hispanic teachers										—	.37

For the replication of Byrnes and Miller's (2007) study, I report weighted statistics on a sample of 15,855 students who participated in the base year and first and second follow-ups, and had not dropped out of high school by the second follow-up. Transformations conducted to each of the variables are described in Appendix B, and consisted mainly of specifying missing values, dummy coding, creating scale scores from a set of items, and collapsing categories. The descriptive statistics and correlations of the variables were nearly identical to B&M's as indicated in Table 15 and Table 16.

I used the PROC SURVEY command in SAS to conduct hierarchical regression with weights. Assumptions for regression were checked. Multicollinearity and linearity were met, normality was not met (Kolmogorov-Smirnov D = .019, p <.01), homoscedasticity was also not met (Breusch-Pagan statistic (13) = 212.1, p <.001) so heteroscedasticity consistent standard errors were consulted. 17 outliers and influential cases were detected, but retained since the results did not differ with or without them.

Table 15. Weighted descriptive statistics of Byrnes and Miller (2007) variables

Variable	*M*	*SD*	*Min*	*Max*	*Comparison with Byrnes and Miller*
12th grade math	48.2	13.6	16.8	78.1	Exact match on *M*, range
12th grade science	23.4	9.0	10.0	36.0	Exact match on *M*, range
8th grade SES	-.1	3.2	-3.0	2.6	*M*=0, Range = -2.93 to 2.75
Parent expectations in 8th grade	-	-	-	-	Exact match on distribution (0=11.9%, 1=27.3%, 2=39.6%, 3=21.3%)
Student expectations in 8th grade	-	-	-	-	Exact match on distribution (0=10.0%, 1=22.1%, 2=44.7%, 3=23.3%)
Middle school GPA	2.9	3.1	.5	4.0	Exact match on *M*, range
General math courses	-	-	-	-	Comparative information not available (0yr=70.7%, .5yr=2.9%, 1yr=16.9%, >1yr=9.5%)
Geometry courses	-	-	-	-	Comparative information not available (0yr=51.2%, .5yr=4.88%, 1yr=45.1%, >1yr=1.32%)
Algebra II courses	-	-	-	-	Comparative information not available (0yr=72.9%, .5yr=4.43%, 1yr=.22%, >1yr=.7%)
General science courses	-	-	-	-	Comparative information not available (0yr=72.8%, .5yr=2.9%, 1yr=20.4%, >1yr=3.9%)
Biology courses	-	-	-	-	Comparative information not available (0yr=14.3%, .5yr=5.7%, 1yr=76.7%, >1yr=3.3%)
Chemistry courses	-	-	-	-	Comparative information not available (0yr=82.9%, .5yr=2.3%, 1yr=14.4%, >1yr=.4%)
Student perception of math emphasis	2.7	4.1	0	5	Exact match on *M*, range
Student perception of science emphasis	2.1	3.8	0	0	Exact match on *M*, range
Student perception of teacher responsiveness	4.7	4	0	7.8	*M*=9.94, Range = 1.65 to 15.54
Math achievement before start of 9th grade	22.6	10.6	7.3	39.9	*M*=35.88
Math GPA in gr 9 & 10	2.8	3.5	.5	4.0	Exact match on *M*, range
Science achievement before start of 9th grade	13.8	7.6	5.2	24.9	*M*=18.72
Science GPA in gr 9 & 10	2.8	3.5	.5	4.0	Exact match on *M*, range
Efficacy for graduating HS	2.83	2.5	0	3.0	Match on *M* (2.84), range
Plans to take SAT	.61		0	1	Exact match on *M*, range
Math self-concept	6.1	6.2	0.0	9.9	*M*=8.90, Range = -1.70 to 15.50
Female	.50		0	1	Exact match on *M*, range
Asian, Pacific Islander	.04		0	1	Exact match on *M*, range

Table 15 continued

Variable	*M*	*SD*	*Min*	*Max*	*Comparison with Byrnes and Miller*
Black	.13		0	1	Exact match on *M*, range
Hispanic	.10		0	1	Exact match on *M*, range
Native Am, Alaskan Native	.01		0	1	Exact match on *M*, range
White	.72		0	1	Exact match on *M*, range

Table 16. Selected intercorrelations of Byrnes and Miller (2007) variables

Note: Byrnes and Miller's statistics, if different from the replication set, are underlined and indicated underneath.

Variable						
Within distal	*(1)*	*(2)*	*(3)*	*(4)*		
(1) SES	—	.42	.38	.29		
(2) Parent expectations		—	.50	.38		
(3) Student expectations			—	.44		
(4) Middle school GPA				—		
Within opportunity	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	
(1) General math (1 yr)	—	.26	-.14	-.11	-.08	
			<u>-.15</u>	<u>-.10</u>		
(2) Geometry (1 yr)		—	.25	.18	.12	
					<u>.13</u>	
(3) Algebra II (1 yr)			—	.11	.09	
					<u>.10</u>	
(4) Math emphasis				—	.23	
					<u>.24</u>	
(5) Teacher responsiveness					—	
Within opportunity (science)	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	
(1) General science (1 yr)	—	-.09	-.11	-.05	-.08	
(2) Biology (1 yr)		—	.14	.07	.07	
				<u>.08</u>		
(3) Chemistry (1 yr)			—	.15	.08	
(4) Science emphasis				—	.20	
(5) Teacher responsiveness					—	
Within propensity	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
(1) Math achv prior to 9th gr	—	.72	.37	.38	.22	.35
			<u>.38</u>	<u>.39</u>	<u>.23</u>	
(2) Sci achv prior to 9th gr		—	.25	.34	.19	.30
				<u>.35</u>	<u>.20</u>	
(3) Math GPA gr9&10			—	.46	.22	.22
					<u>.24</u>	
(4) Sci GPA gr9&10				—	.25	.30
					<u>.28</u>	
(5) Efficacy for graduating					—	.21
						<u>.23</u>
(6) Plans to take SAT						—
Distal with opportunity (math)	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*	
(1) SES	-.19	.31	.18	.10	.06	
				<u>.11</u>		
(2) Parent expectations	-.21	.33	.22	.15	.14	
		<u>.34</u>				

Table 16 continued

Variable							
(3) Student expectations	-.21	.33	.21	.14	.14		
					.15		
(4) Middle school GPA	-.28	.45	.32	.20	.17		
(5) General math (1 yr)							
(6) Geometry (1 yr)							
(7) Algebra II (1 yr)							
(8) Math emphasis							
(9) Teacher responsiveness							
Distal with opportunity (science)	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*		
(1) SES	-.13	.18	.18	.06	.06		
				.07			
(2) Parent expectations	-.16	.18	.18	.10	.14		
(3) Student expectations	-.15	.18	.20	.11	.14		
					.15		
(4) Middle school GPA	-.20	.27	.26	.12	.17		
(5) General science (1 yr)							
(6) Biology (1 yr)							
(7) Chemistry (1 yr)							
(8) Science emphasis							
(9) Teacher responsiveness							
Distal with propensity	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*	*(10)*	*(11)*
(1) SES	.43	.38	.16	.21	.18	.30	.09
					.19		
(2) Parent expectations	.39	.33	.20	.25	.17	.35	.14
	.40	.34			.19		
(3) Student expectations	.38	.34	.21	.30	.21	.37	.16
					.23		
(4) Middle school GPA	.53	.44	.42	.51	.25	.37	.28
					.27		
(5) Math achv prior to 9th gr							
(6) Sci achv prior to 9th gr							
(7) Math GPA gr9&10							
(8) Sci GPA gr9&10							
(9) Efficacy for graduating							
(10) Plans to take SAT							
(11) Math self-concept							
Opportunity with propensity (math)	*(6)*	*(7)*	*(8)*	*(9)*	*(10)*		
(1) General math (1 yr)	-.32	-.12	-.12	-.18	-.10		
	-.31		-.13				
(2) Geometry (1 yr)	.50	.25	.20	.31	.20		
			.22				
(3) Algebra II (1 yr)	.36	.21	.12	.22	.20		
	.38		.13				

Table 16 continued

Variable					
(4) Math emphasis	.20	.19	.12	.16	.20
			.13		
(5) Teacher responsiveness	.12	.20	.14	.16	.20
		.21	.16	.17	.21
(6) Math achv prior to 9th gr					
(7) Math GPA gr9&10					
(8) Efficacy for graduating					
(9) Plans to take SAT					
(10) Math self-concept					
Opportunity with propensity (science)	*(6)*	*(7)*	*(8)*	*(9)*	
(1) General science (1 yr)	-.18	-.14	-.10	-.14	
			-.11		
(2) Biology (1 yr)	.21	.15	.16	.19	
	.20		.17		
(3) Chemistry (1 yr)	.22	.17	.08	.18	
	.23		.09		
(4) Science emphasis	.12	.12	.09	.14	
		.13			
(5) Teacher responsiveness	.09	.21	.14	.16	
	.10		.16	.17	
(6) Sci achv prior to 9th gr					
(7) Sci GPA gr9&10					
(8) Efficacy for graduating					
(9) Plans to take SAT					

3.2 DATA MINING

I used the CRISP-DM model for data mining as a guide, modifying some of the details to be more compatible with educational research. I supplemented the first step—business understanding, or in my case, understanding of the research context and question—with a process inspired by Kane's (2006) argument-based approach to validity in the field of educational and psychological measurement. Specifically, I tried to identify the type of interpretations or claims that I hoped to make using data mining, and to articulate the specific types of evidence that would be needed to support each of these claims. I added this step to help focus and direct the subsequent process of exploration.

3.2.1 Problem understanding, desired inferences and validity evidence

The objective of rule and ruleset induction process was to supplement Thomas' inquiry on what factors matter for student achievement of Black students, and to what extent they matter, and to supplement Byrnes and Miller's inquiry on predictors of high school mathematics achievement. More specifically, the purpose was to examine (1) whether mined rulesets highlight important variables and relationships (including potentially interesting predictor-outcome relationships that could be unique to subgroups) not found from regression, (2) whether rulesets have higher predictive accuracy than regression, and (3) whether interesting rules (relevant to the field and not discovered by regression) are mined. The desired inferences and the types of evidence I could collect to support each inference are summarized in Table 17.

Table 17. Desired inferences and associated validity evidence for rule and ruleset induction

Desired inferences	Supportive evidence
Mined ruleset(s) have higher predictive accuracy than regression.	Ruleset mining approaches (including ensemble approaches) have higher overall accuracy, and small- and large- group accuracies, relative to regression.
Mined rulesets highlight important variables and relationships not found in regression.	Relationships found in rulesets fulfill the following criteria: ✓ Not discovered by regression approaches ✓ Unlikely to be an artifact* ✓ Has potential implications for practice/research
Interesting rules were discovered, and these rules were not found by regression.	Rules fulfill at least some of the following criteria: ✓ Contradicts main trends discovered by regression approaches ✓ Unlikely to be an artifact* ✓ Has potential implications for practice/research

*As explained in Section 2.4.1, a ruleset or rule may be an artifact of sampling, algorithmic bias, data representation, or data inadequacy.

3.2.2 Data understanding

Data understanding was acquired through reviewing the studies and documentation about the dataset, examining descriptive statistics and correlation matrices, replicating their analyses, and reflecting on the following key questions: (1) Who is the sample? How representative is the sample? (2) Which variables appear to be related? (3) In what ways might the data be limited for the purposes at hand (e.g., missing, erroneous, irrelevant, illegitimately correlated with the outcome)? (4) Are there any other (good) ways for key variables to be represented?

3.2.3 Data preparation

Data prepared for the replication (Section 3.1) were used for the first round of data mining for each study. For the second round of data mining for each study, the datasets were expanded to include all available variables in the NELS:88-92 datasets. The data preparation steps are outlined in Appendices G and H. Only potentially relevant variables were included. This meant excluding flags, weights, variables that are highly correlated with the outcome. It also meant including only base year variables as predictors for Thomas' re-analysis, while for Byrnes and Miller's re-analysis including first and second follow-up variables only if they had to do with opportunities provided by the school or community, and not "propensities" of students (assessment on teachers' expectation for S not included, but S assessment on whether teachers and students get along in general in the school is included). Depending on the algorithm, missing data were retained, or substituted with the variable mean or median. Numeric data were collapsed into 4 categories if algorithms required. The outcome variable for Byrnes and Miller was dichotomized into at/above and below median achievement based on the weighted median, when the algorithm did not allow for a numeric outcome, since evaluation of multi-class outcomes would add complexity beyond the scope of this research project. Variables with more than 95% missing were eliminated. Categorical and ordinal variables with over 10 levels were examined individually and typically re-coded to reduce the number of levels. A few variables that appeared redundant—either conceptually or statistically—were also deleted to improve processing speed and strain on memory. Summary of additional data preparation for each algorithm is presented in Table 18.

For re-analysis of Byrnes and Miller, because the data mining algorithms generally do not have ways to incorporate sample weights, the cleaned dataset was expanded using the sample

weights. I multiplied each case by the weight divided by either 4 or 40, and rounded to the nearest whole number (and rounded up to 1 if less than 1).

There were in fact two possible datasets that were appropriate for the first round of Thomas' reanalysis: the smaller dataset that included the 19 predictors that Thomas used in her final model, and a larger dataset of that included 12 additional predictors that Thomas considered but did not include in her final model due to their weak relationship with the outcome. I prepared and analyzed both, but discuss only the results of the smaller dataset because it seemed like a fairer comparison with Thomas' original results, and because the 12 additional predictors barely changed the results.

Table 18. Rule induction methods and settings

Analysis method	R packages used	Data preparation	Algorithm settings	Validation
Apriori (association rule)	*arules* (Hahsler, Buchta, Gruen, & Hornik, 2016; Hahsler, Chelluboina, Hornik, & Buchta, 2011; Hahsler, Gruen, & Hornik, 2005; Hahsler, Grün, Hornik, & Buchta, 2009)	Missing treated as own category; Numeric data collapsed into categorical. Data transformed into transactional form.	For Thomas reanalysis, only rules that predict high achievers were generated. For Byrnes and Miller reanalysis, only rules that predicted the highest two quintiles were generated (separately for each quintile). Minimum support = .015, lift = 2 (i.e., confidence = .5 for Thomas reanalysis, confidence = .4 for Byrnes and Miller reanalysis). Rules had to be robust within the training set, and include at least one unexpected variable (see text for details).	30% holdout; Within the training set, use 50-50 split to identify rules.
CBA (covering)	*arules* (see above for references) and *rCBA* (Kuchar, 2015)	Used dataset prepared for association rules, except eliminated ">" and ",".	Used Apriori (arules) minimum support = .02, confidence = .98. Rules only predicted high achieving students. rCBA method = m2cba (default).	30% holdout. Within the training set, used 50-50 split to identify Apriori parameters.
RIPPER (covering)	*RWeka* (Hornik, Buchta, & Zeileis, 2009; Witten & Frank, 2005)	Missing numeric data were substituted with the sample mean. Missing categorical data were categorized as a new level.	JRip algorithm with default settings, except minimum cases per node set at 2% of sample. For Thomas reanalysis, cost of misclassifying high achievers set to be 3 times more than misclassifying non-high achievers.	30% holdout.
PART (covering)	*RWeka* (Hornik et al., 2009; Witten & Frank, 2005)	Missing numeric data were substituted with the sample mean. Missing categorical data were categorized as a new level.	PART algorithm with default settings, except minimum cases per node set at 2% of sample. For Thomas reanalysis, cost of misclassifying high achievers set to be 3 times more than misclassifying non-high achievers.	30% holdout.

Table 18 continued

Analysis method	R packages used	Data preparation	Algorithm settings	Validation
CART (tree) and bagging CART (ensemble)	*rpart* (Therneau et al., 2015) and *ipred* (Peters & Hothorn, 2015)	Categorical variables set as factors or ordered. Missing values retained (algorithm default, per recommendation of Breiman et al. (1984), uses surrogates to decide the split).	For standard CART: Gini splitting function. Tree was over-grown (complexity parameter=0, no stopping criteria) and pruned based on 10-fold cross-validated error to the most parsimonious sub-tree where the cross-validated error was within one standard deviation of the best model. Cost of misclassifying high achievers set to be 3 times more than misclassifying non-high achievers. For bagging CART, default settings were used (cost adjustment was not available).	30% holdout.
C5.0 (tree) and boosted C5.0 (ensemble)	*C50* (Kuhn, Weston, Coulter, Quinlan, & Culp., 2015)	Categorical variables set as factors or ordered. Missing values retained. Reclassify any levels with no name.	Cost of misclassifying high achievers set to be 3 times more than misclassifying non-high achievers. For non-boosted model, default settings except CF=.01. For boosted model, default settings except trials=20, CF=.01, fuzzyThreshold = TRUE.	30% holdout. Within the training set, used 50-50 holdout method to decide on C50 parameters including CF, fuzzyTheshold and winnow.
C4.5 (tree)	*RWeka* (Hornik et al., 2009; Witten & Frank, 2005)	Missing numeric data were substituted with the sample mean. Missing categorical data were categorized as a new level.	J48 algorithm with default settings, except minimum cases per node set at 2% of sample. For Thomas reanalysis, cost of misclassifying high achievers set to be 3 times more than misclassifying non-high achievers.	30% holdout.
QUEST (tree)	Not available in R (used SPSS v22)	Exported training and test sets from R using *haven* package. Missing variables were retained.	QUEST algorithm. Pruning enabled. For Thomas reanalysis, cost of misclassifying high achievers set to be 3 times more than misclassifying non-high achievers. Minimum terminal node set at 2 for Thomas reanalysis, and 2% of sample (12320) for Byrnes & Miller reanalysis.	30% holdout.

Table 18 continued

Analysis method	**R packages used**	**Data preparation**	**Algorithm settings**	**Validation**
Random forests (ensemble)	*randomForest* (Liaw & Wiener, 2002)	Missing numeric data were substituted with the sample mean. Missing categorical data were categorized as a new level.	Default settings were used.	30% holdout.

3.2.4 Modeling

A variety of data mining analyses were conducted, relying on R Version 3.2.2 through 3.3.2 (R Core Team, 2016), and RStudio Version 0.99.896 through 1.0.136 (RStudio Team, 2015). The QUEST algorithm was not available as an R package so SPSS Version 22 was used. The analysis flow and codes are provided in Appendices G and H.

Classification Based on Associations (CBA). CBA was conducted using the algorithm instantiated in the R-package *rCBA* (Kuchar, 2015) after rules were mined using Apriori instantiated in the *arules* package (Hahsler et al., 2016; Hahsler et al., 2011; Hahsler et al., 2005; Hahsler et al., 2009). The dataset was split into 30% holdout sample and a 70% training sample. The training sample was further split into half to identify the best association rule generation algorithm parameter settings. I had to use the supercomputer for both, and only the small datasets converged. For both studies, I used a minimum support of .02, confidence .98, minimum length of 2, maximum length of 6 and maximum time of 200 seconds, analyzing with 10 nodes. I automated the process of applying the CBA ruleset to new data and evaluating the outcome by creating an R program provided in Appendix G. These rulesets were validated on the holdout sample.

RIPPER. RIPPER, a sequential covering modeling, was conducted using the JRip algorithm in the *RWeka* package (Hornik et al., 2009; Witten & Frank, 2005). Missing numeric data were substituted by the sample mean, while missing categorical data were combined into a new category titled "missing." The cost of misclassifying high achievers was set to be three times more than misclassifying non-high achievers. The ruleset was derived using a 70% training sample (the same sample used for the other algorithms), and validated on the remaining holdout sample.

PART. PART, a sequential covering modeling, was conducted using the PART algorithm in the *RWeka* package (Hornik et al., 2009; Witten & Frank, 2005). Missing numeric data were substituted by the sample mean, while missing categorical data were combined into a new category titled "missing." The cost of misclassifying high achievers was set to be three times more than misclassifying non-high achievers. The ruleset was derived using a 70% training sample (the same sample used for the other algorithms), and validated on the remaining holdout sample.

Classification and Regression Tree (CART). CART was conducted using the rpart (recursive partitioning) algorithm in the *rpart* package (Therneau et al., 2015). Ordinal variables were converted into ordinal factors, while other categorical variables were designated to be non-ordered factors. Surrogate variables were used to determine the class when a case was missing on a predictor variable. An overgrown classification tree was created and pruned using 70% of the sample. The initial tree was grown without any stopping rule using the Gini splitting criteria and 10-fold cross-validation. The most parsimonious subtree that had an error rate within one standard deviation of the error rate of the model with the lowest cross-validated error was chosen as the final model. This final, pruned tree was validated on the holdout sample.

Bagging CART was conducted using the *ipred* package (Peters & Hothorn, 2015), which uses *rpart* for model construction. The variables prepared for standard CART were used to construct the model, and the model was tested in the same 30% holdout sample as the other trials. Default settings were used for bagging, and the misclassification costs were not adjusted (feature was unavailable).

C5.0. C5.0 was conducted using the *C50* package (Kuhn et al., 2015). Ordinal variables were converted into ordinal factors, while other categorical variables were designated as factors. Missing data were retained. The cost of misclassifying high achievers was set to be three times

more than misclassifying non-high achievers. 70% of the sample was used to determine the parameter settings for one C5.0 tree and one boosted C5.0 tree. The parameters settings that were varied included the confidence factor, whether there was advanced pruning of predictors ("winnowing"), whether possible advanced splits were evaluated ("fuzzy threshold"). The confidence factor, which indicates the pruning severity was varied between .25 (default) and .05 (very severe), in .05 increments; .01 and .001 were also tested. The maximum boosting iteration was set to 20. Both the non-boosted and boosted trees were created using half of the training set and tested on the remaining half. The best performing non-boosted tree based on the confusion matrix, Kappa statistic, F-measure, and tree morphology and simplicity, was where the confidence factor was .01, and default settings were used (i.e., no winnowing, no fuzzy threshold). The best performing boosted tree was when the confidence factor was .01, and fuzzy threshold was enabled.

C4.5. C4.5 was conducted using the J48 algorithm in the *RWeka* package (Hornik et al., 2009; Witten & Frank, 2005). Missing numeric data were substituted by the sample mean, while missing categorical data were combined into a new category titled "missing." The cost of misclassifying high achievers was set to be three times more than misclassifying non-high achievers. The ruleset was derived using a 70% training sample (the same sample used for the other algorithms), and validated on the remaining holdout sample.

QUEST. QUEST was conducted using SPSS Version 22. Missing values were retained. The cost of misclassifying high achievers was set to be three times more than misclassifying non-high achievers. Pruning was enabled using default settings. The ruleset was derived using a 70% training sample (the same sample used for the other algorithms), and validated on the remaining holdout sample.

Random Forest. Random Forest mining was conducted using the randomForest algorithm in the *randomForest* package (Liaw & Wiener, 2002). Missing numeric data were substituted by the sample mean, while missing categorical data were combined into a new category titled "missing." (The imputation function within Random Forest was also explored, but produced very similar results.) Default settings were used. The ruleset was derived using a 70% training sample (the same sample used for the other algorithms), and validated on the remaining holdout sample.

Logistic regression. To generate a confusion matrix for the logistic regression model that is comparable with those generated by the data mining methods, logistic regression was conducted on the same (70%) training sample as the other algorithms, and tested on the holdout sample. Youden's index was used to determine a cutoff value that provides the best balance between the sensitivity and specificity. SPSS was used for this analysis.

Association rule mining. Association rule mining was conducted using the Apriori algorithm instantiated in the R-package *arules* (Hahsler et al., 2016; Hahsler et al., 2011; Hahsler et al., 2005; Hahsler et al., 2009). Initially, I attempted to generate all rules describing high achievers with sufficient generality and accuracy, and screen among those for rules that included attribute-value pairs that were unexpected. However, this turned out to be computationally intractable, even with a supercomputer, particularly when the number of variables were increased. Thus, I used a more targeted approach to rule mining, looking for association rules among predefined subgroups where (I believed) there was a higher possibility for interesting rules to be found. In short, I conducted rule mining on subgroups that were likely to be similar in their outcome, per regression and ruleset mining. Because logistic regression and ruleset mining for the re-analysis for Thomas (2006) showed that higher family income and higher parental education were positively associated with high achievement, I investigated what commonalities there might

be (if any) among high achievers with *lower* family income or *lower* parental education. For study 2, because regression and ruleset mining showed that 8th grade math achievement had a strong and positive relationship with 12th grade math achievement, I split the sample into 4 according to their 8th grade math score and searched within each group for commonalities among those who scored over 10 points than their score predicted by CART. Those with residuals between 7.6 and 10 were excluded from consideration so that any differences between the higher and lower achieving groups (i.e., higher and lower residual group) would not be diluted by those whose outcomes were only somewhat higher than expected.

The relevant datasets were split into a 30% holdout sample and 70% training sample. To reduce the possibility of detecting artifacts, half of the training sample was used to generate the rules (I called this the "generation sample"), and the remaining half was used to screen out inaccurate rules (I called this the "screening sample"). I used educated guesses and trial-and-error to optimize the minimum support and rule length when generating the rules. The goal was to generate rules that were specific enough to include surprising ones, but not too specific to reduce artifacts. Once the algorithm settings were decided and initial set of rules generated from the “generation set”, the rules were tested on the "screening set" and checked for accuracy. Only rules that had the minimum support used for the generation set were preserved.

I generated the coverage of the short list of rules for the entire training set, for the high and lower achievers separately, and used these values to calculate the relative probabilities for each rule (i.e., probability that the rule applies to the high achievers divided by the probability that the rule applies to the lower achievers). I sorted the list of promising rules in descending order by the relative probability, and more carefully examined the meaning of the top rules. I only looked at

rules with length 2 and 3 (i.e., with 1 and 2 conditions in the antecedent) because generating and examining longer rules proved to be too time-consuming to be worthwhile.

3.2.5 Evaluation and model deployment

The rules and rulesets were evaluated based on the validity criteria articulated in the initial stage of the data mining process (Section 3.2.1, particularly Table 17). To assess whether mined rulesets have a higher predictive accuracy than regression, the test set confusion matrices (including the Kappa statistic and F-measure) of 10 ruleset generation approaches were compared with that of logistic regression.

To assess whether the rule and ruleset induction approaches highlighted important variables and relationships not found in regression, rule meaning and their accuracy measures were reviewed. Rules within rulesets generally required brute-force calculations of their coverage and confidence for the training and test sets. This was not a big problem when the number of rules was small, but not practically feasible for CBA (which generated a ruleset including 67 rules for Thomas (2006) re-analysis, and 832 rules for Byrnes & Miller (2007) re-analysis). I examined each rule and their accuracy measures in the training set to identify a list of rules that seemed interesting (at minimum they had to tell me something that regression did not tell me), and documented reasons I found them interesting. Then, I examined the performance on the test set to see whether the generality and accuracy still held. For study 2, because the data lent itself, I represented each of the rules and accuracy measures using mosaic diagrams, which made them helpful to compare.

4.0 RESULTS

I first present results from replication and rule induction re-analyses of Thomas (2006), which I call "Study 1." Results from Byrnes and Miller's (2007) replication and re-analyses ("Study 2") follows.

4.1 RESULTS FOR STUDY 1 (THOMAS, 2006)

4.1.1 Replication

A side-by-side comparison of the replication results, and Thomas' (2006) results, are shown in Table 19 through Table 21. The directions and magnitude of the coefficient estimates were quite similar between the two, although there were also some differences, presumably due to some of the changes in the methodology and sample noted above. Clear patterns observed in both Thomas' and my replication results were as follows:

1. Higher parental education is associated with high achievement. All else being equal, a student whose parent's highest level of education is one level higher has approximately a 51% (replication) or 22% (Thomas, 2006) greater likelihood of being in the high achieving group.
2. Higher family income is associated with high achievement. All else being equal, a student whose family income is one level higher is 7% (replication) or 8% (Thomas, 2006) more

likely to be in the high achieving group. However, this association seem to not exist for males.

3. Greater hours of homework outside of school is associated with high achievement (+14% (replication) or +29% (Thomas, 2006) per level).
4. The percentage of students receiving free lunch in school is negatively associated with achievement, particularly for males.
5. Students, particularly females, who report feeling unsafe in school, or that disruptions in school prevent their learning, are less likely to be high achieving.
6. The number of Black teachers in students' school is negatively associated with high achievement, particularly for females.[9]

Some patterns observed in both analyses were less clear:

7. Attending a private school is negatively associated with being a high achiever (likelihood is reduced to 51% (replication) or 25% (Thomas, 2006)). However, this effect may not be very strong, as it seems to disappear when males and females are examined separately.
8. Having "bad" peers may be negatively associated with high achievement, particularly for females.
9. Having peers who expect the student to go to college is positively associated with high achievement. However, this effect may not be very strong, as it seems to disappear when males and females are examined separately.

[9] A reminder that this association does *not* imply that Black teachers *cause* low student achievement. It is very likely that in the late 1980s when the data were collected, the number of Black teachers reflected school neighborhood average socioeconomic status, and therefore school resources, which impacts students' opportunity to learn and achievement.

10. For males, participating in a greater number of cultural activities may be positively associated with high achievement.

In addition, both analyses found that hours of homework conducted in school, household resources, parental involvement in school, activities outside of school, are not associated with high achievement, after controlling for other factors.

There were only three main inconsistencies observed between Thomas' and replication analyses. Only Thomas' analyses but, not replication, suggested that:

11. Parental expectation of college was positively associated with high achievement for females.
12. Having "good" peers was negatively associated with high achievement, particularly for males.
13. A positive school climate may be positively associated with achievement, particularly for males.

These three inconsistencies could be due to differences in sample and data cleaning methodology. There was no instance where Thomas' results and replication results showed effects in opposite directions. The overall similarity between the results two sets of analyses suggest that the replication was conducted successfully and provide confidence that most of the findings above—particularly results 1 through 6—are indeed patterns in the data that can be detected by logistic regression.

Table 19. Prediction of Black student achievement with NELS:88, with Thomas' (2006) final variables

Variable	Replication		Thomas (2006)	
	B	*Exp(B)*	*B*	*Exp(B)*
Intercept	-2.622***	.073	-1.344***	.261
Parental education (BYPARED)	.415***	1.514	.201**	1.223
Income from all sources 1991 (F2P74)	.067*	1.069	.083**	1.087
Hours of homework in school (F2S25F1)	.044	1.045	.039	1.040
Hours of homework out of school (F2S25F2)	.128**	1.137	.255***	1.290
Household resources (hhressc)	.029	1.029	.038	1.039
Private school (privsch)	-.668+	.512	-1.4**	.247
Religious school (religsch)	.677	1.969	1.775**	5.900
Parental involvement in school (pinvolve)	.050	1.051	.069	1.071
Parents expect college (pexpcol)	.081	1.084	.82***	2.270
Good peers (goodpeer)	-.078	.925	-.549**	.578
Bad peers (badpeer)	-.313+	.731	-.774***	.461
Peers expect college (peerexcl)	.342*	1.408	.357*	1.429
Activities outside of school (activity)	.016	1.016	.055	1.057
Student's cultural activities (sculture)	.131*	1.140	.086	1.090
Percent receiving free lunch in school (G8LUNCH)	-.097*	.907	-.051	.950
School climate (climate)	.026	1.026	.07**	1.073
Student feels unsafe in school (unsafe)	-.799**	.450	-.801**	.449
Disruptions in school prevent learning (disrupt)	-.573***	.564	-.421*	.656
Number of Black, non-Hispanic teachers (BYSC20D)	-.114**	.892	-.116**	.890

Table 20. Prediction of Black female student achievement with NELS:88, with Thomas' (2006) final variables

Variable	Replication		Thomas (2006)	
	B	*Exp(B)*	*B*	*Exp(B)*
Intercept	-3.207**	.040	-2.844*	.058
Parental education (BYPARED)	.488***	1.629	.111	1.117
Income from all sources 1991 (F2P74)	.114*	1.121	.139**	1.149
Hours of homework in school (F2S25F1)	.041	1.042	.048	1.049
Hours of homework out of school (F2S25F2)	.182**	1.199	.379***	1.461
Household resources (hhressc)	-.029	.972	.052	1.053
Private school (privsch)	-.772	.462	-1.907*	.149
Religious school (religsch)	.891	2.438	2.024*	7.569
Parental involvement in school (pinvolve)	.054	1.055	.105	1.111
Parents expect college (pexpcol)	.519	1.680	1.267***	3.550
Good peers (goodpeer)	-.076	.926	-.359	.698
Bad peers (badpeer)	-.513+	.599	-.848**	.428
Peers expect college (peerexcl)	.315	1.370	.416+	1.516
Activities outside of school (activity)	.059	1.061	.091	1.095
Student's cultural activities (sculture)	.121	1.129	.05	1.051
Percent receiving free lunch in school (G8LUNCH)	-.052	.950	-.02	.980
School climate (climate)	.047	1.048	.064+	1.066
Student feels unsafe in school (unsafe)	-1.047*	.351	-1.206**	.299
Disruptions in school prevent learning (disrupt)	-.981***	.375	-.573*	.564
Number of Black, non-Hispanic teachers (BYSC20D)	-.146*	.864	-.157*	.855

Table 21. Prediction of Black male student achievement with NELS:88, with Thomas' (2006) final variables

Variable	Replication		Thomas (2006)	
	B	*Exp(B)*	*B*	*Exp(B)*
Intercept	-2.250*	.105	.288	1.334
Parental education (BYPARED)	.387**	1.472	.27*	1.310
Income from all sources 1991 (F2P74)	.003	1.003	-.001	.999
Hours of homework in school (F2S25F1)	.048	1.049	.051	1.052
Hours of homework out of school (F2S25F2)	.073	1.076	.13	1.139
Household resources (hhressc)	.077	1.081	.024	1.024
Private school (privsch)	-.548	.578	-.946+	.388
Religious school (religsch)	.420	1.522	1.785*	5.960
Parental involvement in school (pinvolve)	.064	1.066	.017	1.017
Parents expect college (pexpcol)	-.117	.890	.5+	1.649
Good peers (goodpeer)	-.095	.910	-.668*	.513
Bad peers (badpeer)	-.137	.872	-.761*	.467
Peers expect college (peerexcl)	.416+	1.516	.215	1.240
Activities outside of school (activity)	-.047	.954	-.058	.944
Student's cultural activities (sculture)	.129+	1.138	.168*	1.183
Percent receiving free lunch in school (G8LUNCH)	-.160**	.852	-.137*	.872
School climate (climate)	.008	1.008	.082*	1.085
Student feels unsafe in school (unsafe)	-.515	.598	-.352	.703
Disruptions in school prevent learning (disrupt)	-.219	.803	-.291	.748
Number of Black, non-Hispanic teachers (BYSC20D)	-.065	.937	-.058	.944

4.1.2 Results from rule induction

I first present predictive accuracies across approaches, followed by model predictors and their importance, interesting rules and their accuracies, and results from association rule mining.

4.1.2.1 Predictive accuracies of ruleset induction

The confusion matrices and associated predictive accuracy measures for ruleset induction for Study 1, using 19 and 1372 predictors, are presented in tables Table 22 and Table 23, respectively. The relative performance of each algorithm in terms of the F-measure and the Kappa statistic are illustrated in Figure 5. The overall accuracy for logistic regression using only the variables Thomas used for her study (presented at the bottom of Table 22) was 70%, with 67% of the high achieving group correctly classified (recall), and 43% of those who were classified as high achieving being correctly classified (precision). The F-measure, or the harmonic mean between precision and recall, was .525. The Kappa statistic was .316, indicating that if predicting as well as random chance were 0 and making a perfect prediction were 100, logistic regression performed at about 32.

With just 19 variables used in the logistic regression analysis, the rule induction classifiers performed comparably or slightly worse, depending on the accuracy measure, with overall accuracy ranging from 78% to 53%, the F-measure ranging from .359 to .509, and the Kappa statistic between .117 and .317. PART performed the closest to logistic regression in terms of the F-measure and Kappa statistic, followed by bagging CART and C5.0. PART and bagging CART tended to be more conservative than logistic regression in categorizing students as "high achieving", while C5.0 tended to be more liberal. The worst performers in terms of the F-measure was Random Forest, which classified the fewest students as high achieving (but was correct

approximately 2/3 of the time). The worst performer in terms of the Kappa statistic were CART and RIPPER. Their predictions of higher and lower achievers were less accurate relative to logistic regression.

With an additional 1354 predictors from which to build a model, the rule induction methods generally improved in their predictive accuracies. QUEST surpassed logistic regression in terms of both F-measure and Kappa statistics, while two ensemble methods—Random Forest and boosted C5.0—performed better than logistic regression in terms of Kappa but worse in their F-measures. CART, RIPPER, PART and C5.0 performed comparably to logistic regression in at least one measure and was slightly worse on the other, while C4.5 and bagging CART performed quite a bit worse than logistic regression according to both measures.

Table 22. Confusion matrices for ruleset mining (Study 1, 19 possible predictors)

		Prediction on test set				
		High achieving	Not high achieving	% Correct	F-measure	Kappa
CBA	High achieving	66	26	71.7%	.496	.250
	Not high achieving	108	167	60.7%		
	% Correct	37.9%	86.5%	63.5%		
RIPPER	High achieving	52	40	56.5%	.421	.155
	Not high achieving	103	172	62.5%		
	% Correct	33.5%	81.1%	61.0%		
PART	High achieving	56	36	60.9%	.509	.307
	Not high achieving	72	203	73.8%		
	% Correct	43.8%	84.9%	70.6%		
CART	High achieving	63	29	68.5%	.423	.117
	Not high achieving	143	132	48.0%		
	% Correct	30.6%	82.0%	53.1%		
C5.0	High achieving	71	21	77.1%	.503	.250
	Not high achieving	119	156	56.7%		
	% Correct	37.4%	88.1%	61.9%		
C4.5	High achieving	62	30	67.4%	.482	.237
	Not high achieving	103	172	62.5%		
	% Correct	37.6%	85.1%	63.8%		
QUEST	High achieving	55	37	59.8%	.451	.201
	Not high achieving	97	178	64.7%		
	% Correct	36.2%	82.8%	63.5%		
Bagging CART	High achieving	38	54	41.3%	.466	.317
	Not high achieving	33	242	88.0%		
	% Correct	53.5%	81.8%	76.3%		
Boosted C5.0	High achieving	28	64	54.3%	.400	.275
	Not high achieving	20	255	70.5%		
	% Correct	38.1%	82.2%	66.5%		
Random forest	High achieving	23	69	25.0%	.359	.254
	Not high achieving	13	262	95.3%		
	% Correct	63.4%	79.1%	77.7%		
Logistic regression	High achieving	62	30	67.4%	.525	.316
	Not high achieving	82	193	70.2%		
	% Correct	43.1%	86.7%	69.5%		

Table 23. Confusion matrices for ruleset mining (Study 1, 1372 possible predictors)

		Prediction on test set				
		High achieving	Not high achieving	% Correct	F-measure	Kappa
CBA	High achieving					
	Not high achieving		*Results Not*	*Attained*[1]		
	% Correct					
RIPPER	High achieving	59	33	64.1%	.500	.280
	Not high achieving	85	190	69.1%		
	% Correct	41.0%	85.2%	67.8%		
PART	High achieving	72	20	78.3%	.509	.258
	Not high achieving	119	156	56.7%		
	% Correct	37.7%	88.6%	62.1%		
CART	High achieving	38	54	44.6%	.472	.328
	Not high achieving	31	244	79.6%		
	% Correct	42.3%	81.1%	70.8%		
C5.0	High achieving	60	32	65.2%	.500	.276
	Not high achieving	88	187	68.0%		
	% Correct	40.5%	85.4%	67.3%		
C4.5	High achieving	32	60	34.8%	.390	.218
	Not high achieving	40	235	85.5%		
	% Correct	44.4%	79.7%	72.8%		
QUEST	High achieving	72	20	78.3%	.593	.408
	Not high achieving	79	196	71.3%		
	% Correct	46.7%	90.7%	73.0%		
Bagging CART	High achieving	92	0	100%	.401	0
	Not high achieving	275	0	0%		
	% Correct	25.1%	-	25.1%		
Boosted C5.0	High achieving	31	61	33.7%	.423	.363
	Not high achieving	11	264	96.0%		
	% Correct	73.8%	81.2%	80.4%		
Random forest	High achieving	31	61	33.7%	.470	.375
	Not high achieving	9	266	96.7%		
	% Correct	77.5%	81.3%	80.9%		

[1]Possibly due to the large number of predictors, CBA algorithm did not run to convergence.

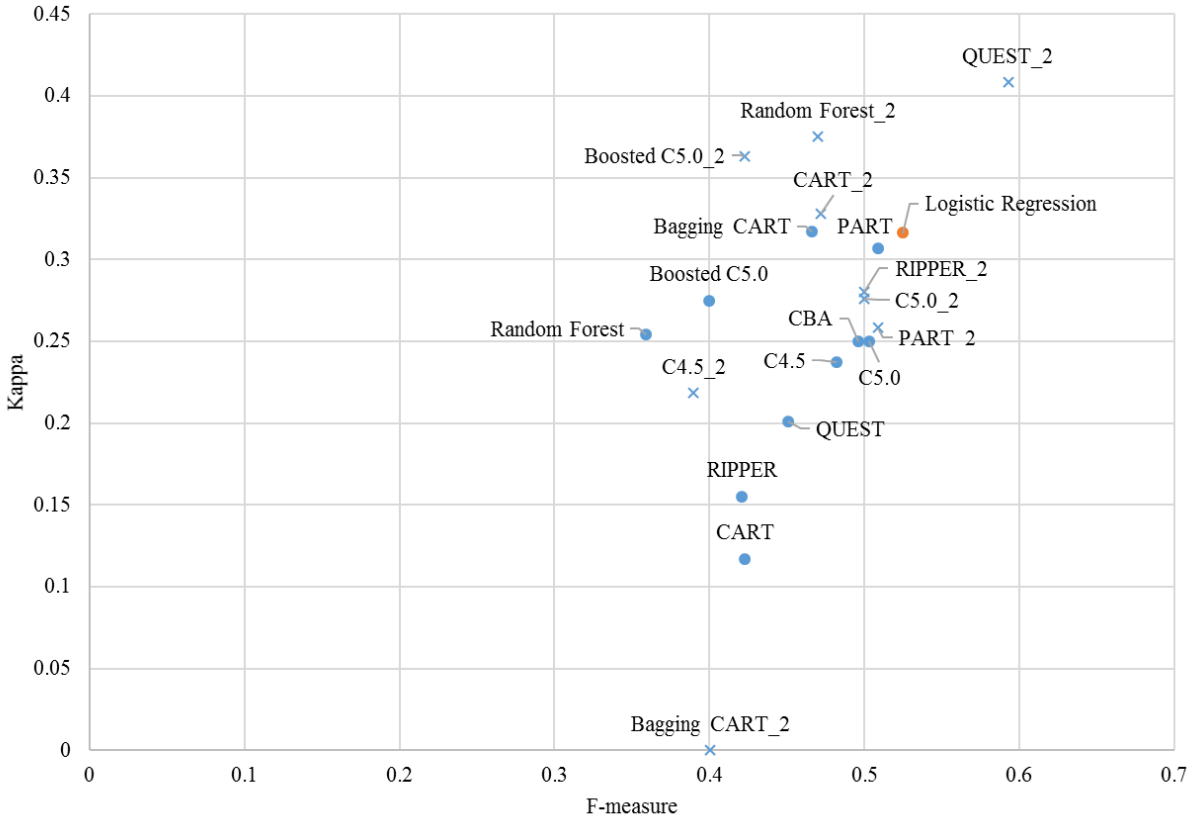


Figure 5. F-measure and Kappa statistics of logistic regression vs rule induction (Study 1)[1]

[1]The orange marker represents result from logistic regression, while the blue markers indicate results from rule induction approaches. Circle markers indicate results using Thomas' (2006) 19 possible predictors, while "X"s indicate results using 1372 possible predictors.

4.1.2.2 Model predictors and their importance

Figure 6 and Figure 7 illustrate the relative importance of the predictors or predictor sets that were most frequently included across eight of the ten ruleset induction models for which such information was available. The criteria for "model importance" varied slightly across models, depending on the information that was available for each algorithm, and are indicated as notes attached to each figure. The figures indicate that generally, across algorithms, parent educational attainment was most important among 19 variables that Thomas had considered in her final model, and that when 1372 variables were considered, SES, academics (e.g., whether student is in gifted

classes, or had ever been held back, see Appendix C, Table 56), parent expectation and behavior were at least as important as parent educational attainment. The figures also indicate that predictor (or predictor set) importance varies widely across rule induction approaches.

Figure 8 and Figure 9 better illustrate how predictor or predictor set importance varies across rule induction approaches. The heights of the predictors are proportional to their importance. When only 19 predictors were available, the single tree inducers and PART strongly relied on parental educational attainment to classify Black students. RIPPER, and the two ensemble methods relied on many more variables about as much as they did on parental education. CART, C4.5 and QUEST relied on just 1 to 3 predictors, while the other approaches relied on at least 7 predictors.

Results across algorithms varied even more when 1372 predictors were made available for rule induction. Tree approaches relied on the fewest set of predictors (3 to 7), with C5.0 relying heavily on SES, CART relying heavily on parental expectations, and C4.5 relying heavily on a combination of academics, behavior and SES. In addition to SES, academics and parent expectations, QUEST utilized school type and a variable on the number of hours the 8th grader reads on their own. As might be expected, the ensemble methods utilized many more variables, as did sequential covering models. All four approaches utilized academics more than other types of predictors. RIPPER also relied heavily on SES, academic expectation, behavior and locus of control. Random Forest emphasized locus of control, SES and several parent factors (expectation, education, employment), while boosted C5.0 emphasized student factors (locus of control, behavior), parent factor (education), and school factors (type, structure/policy).

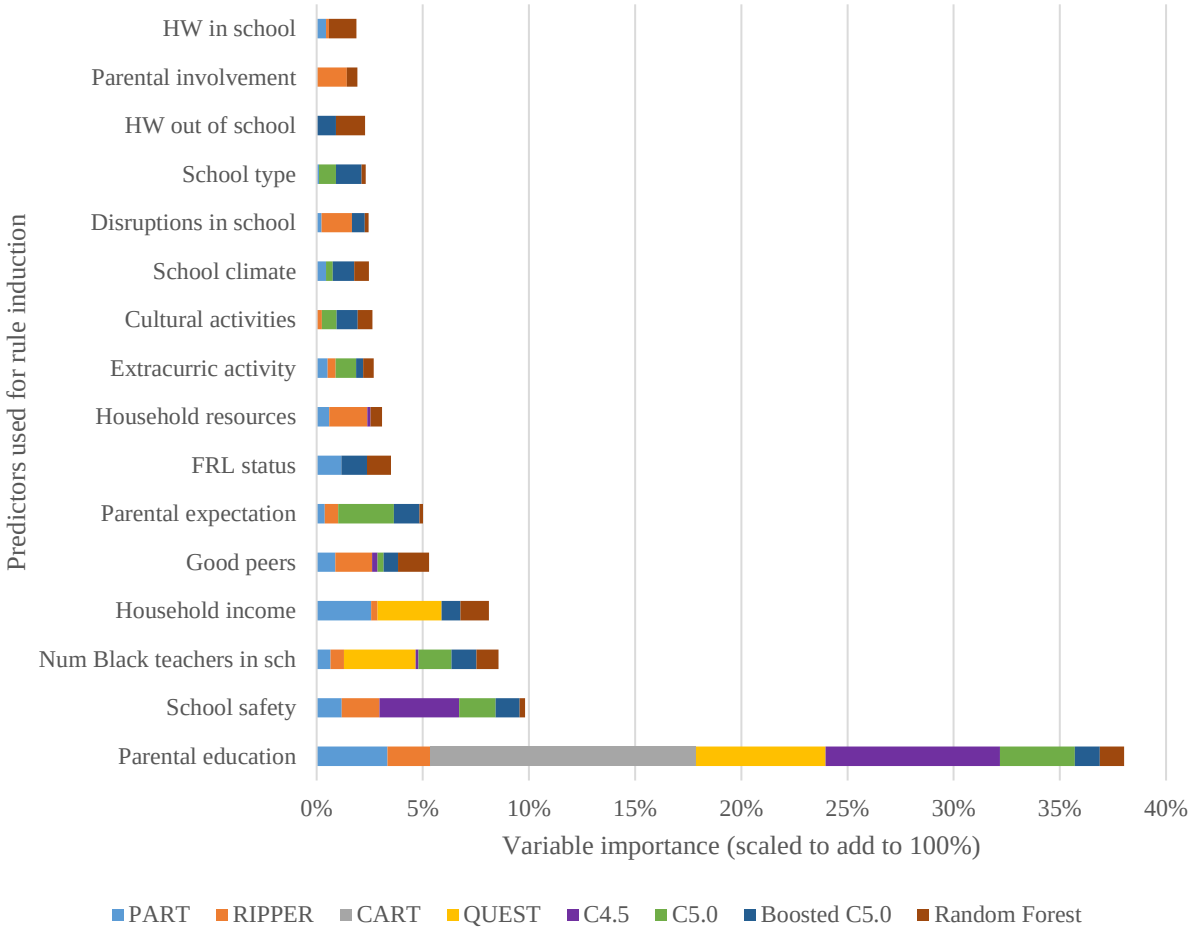


Figure 6. Predictor importance by algorithm (Study 1, 19 possible predictors)

Note: Predictor importance for CART and Random Forest were calculated as the extent to which the variable reduced the Gini index, as these were automatically generated. For the rest of the algorithms where that measure was not available, attribute usage was used. Attribute usage was calculated as the number of participants that the variable sorted. The predictor importance for each algorithm was scaled to total 100. Private school and religious school attendance was grouped into "school type", and three peer-related predictors were combined as "good peers."

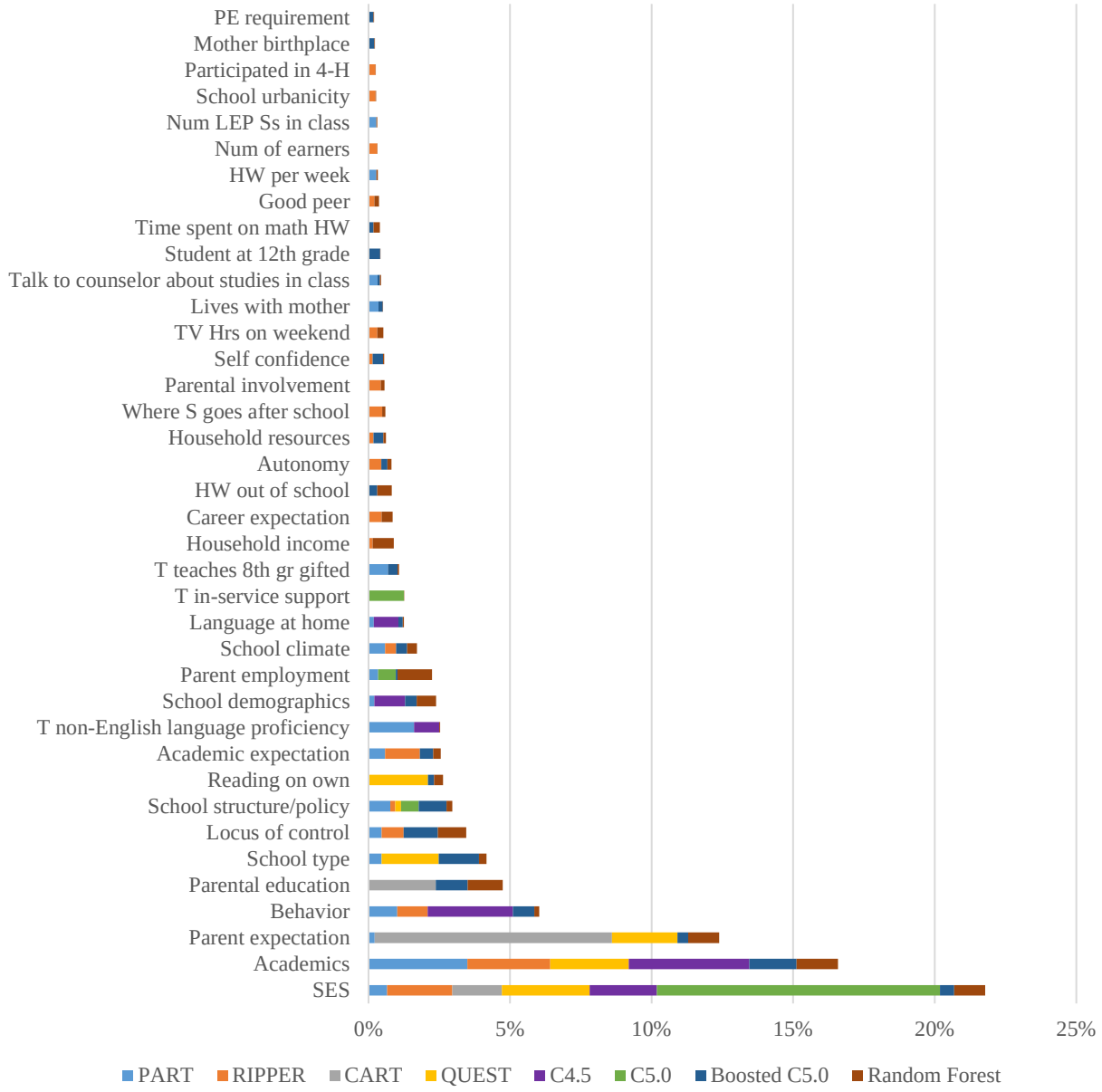


Figure 7. Predictor importance by algorithm (Study 1, 1372 possible predictors)

Note: Predictor importance for CART and Random Forest were calculated as the extent to which the variable reduced the Gini index, as these were automatically generated. For the rest of the algorithms where that measure was not available, attribute usage was used. Attribute usage was calculated as the number of participants that the variable sorted, where participants were double-counted as being sorted by that variable if, in tree-induction, the variable was used twice to sort the same person and if there was at least one other variable that sorted that person between the first and second sorting. The predictor importance for each algorithm was scaled to total 100, predictors that contributed less than 1 in all of the non-ensemble algorithms were excluded from consideration, and remaining predictors were grouped into the categories in the figure. Refer to Appendix C, Table 56 for the groupings.

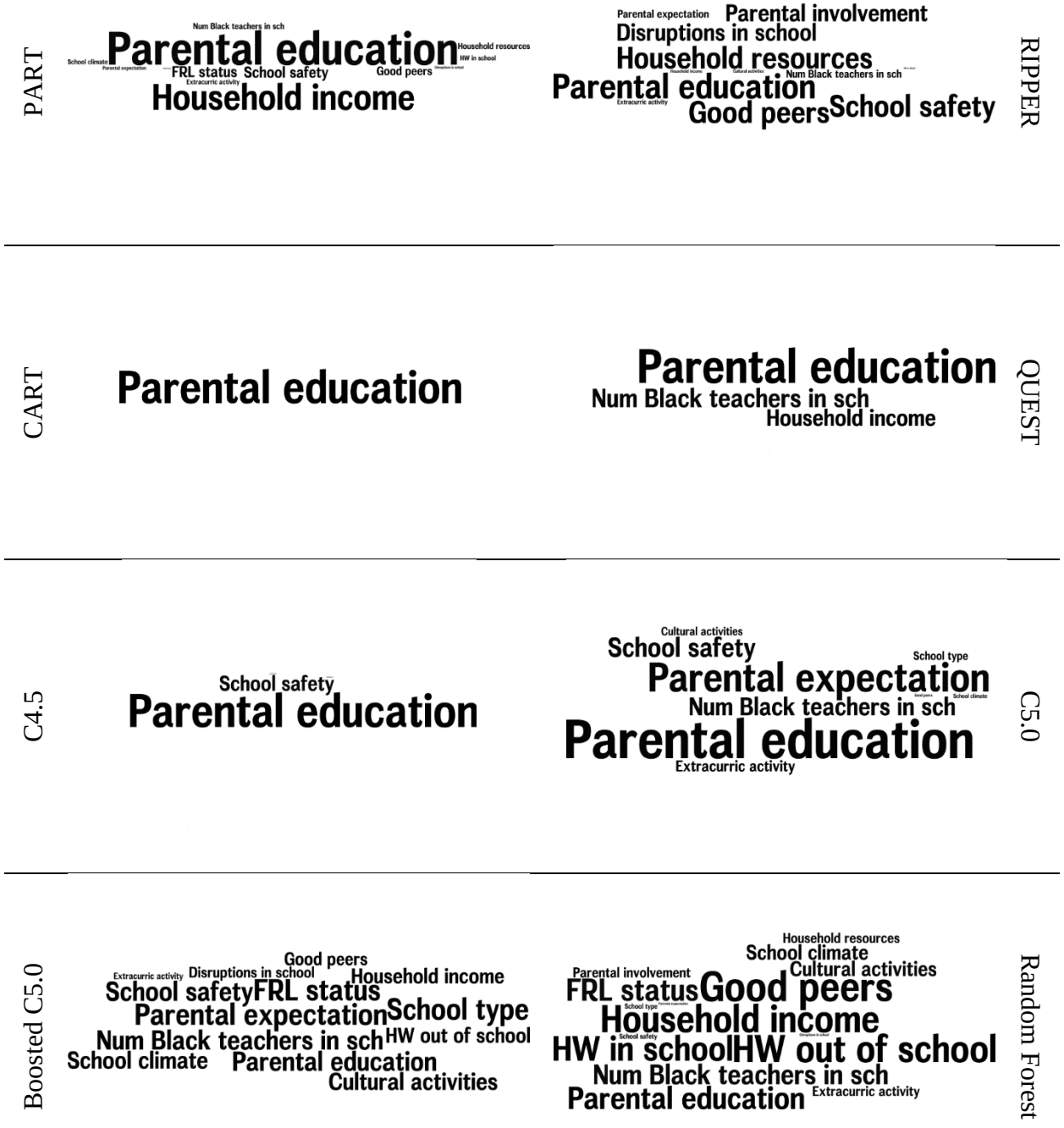


Figure 8. Predictors included in model, sized proportionally to importance (Study 1, 19 possible predictors)

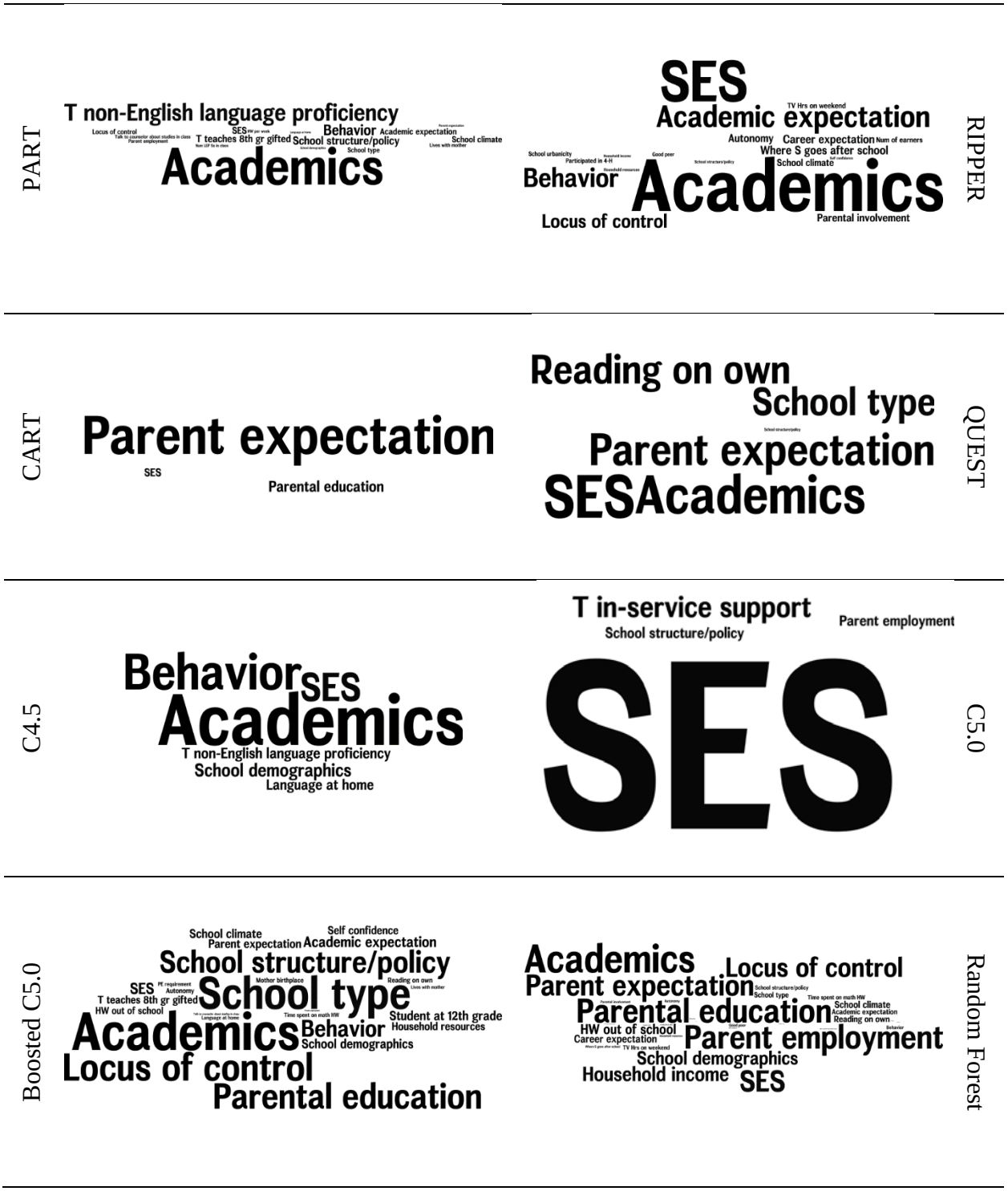


Figure 9. Predictors included in model, sized proportionally to importance (Study 1, 1372 possible predictors)

4.1.2.3 Interesting rules within rulesets and their accuracies

The rulesets and trees that were mined for each algorithm, including their confidence (proportion the rule was correct, given the antecedent applied) and coverage (proportion of the sample to which the rule applied) are presented in Appendix C. Based on rule meanings and training set performance, 16 out of 213 rules appeared to provide additional information over and above what was discovered through logistic regression. However, further inspection of the coverage and confidence in the test set suggested that about a third of those were likely to be statistical artifacts since the rules were not as accurate or prevalent in the test set. Table 24 summarizes the number of rules, interesting rules and false alarms (i.e., rules that initially seemed interesting but did not perform well in the test set). Table 25 presents the 11 rules that were not discovered by logistic regression results, and seemed relatively less likely to be a design or statistical artifact based on their predictive accuracy measures and rule semantics. Table 26 describes five rules that initially appeared to be interesting, but upon further inspection seemed to be false alarms. For some of the rules, I conducted further analyses to identify the relationship of subsets of attribute-values and the outcome.

The rule induction including only 19 predictors highlighted potentially interesting *combinations* of attribute-values that are predictive of the outcome that were not observable through regression. These combinations may have implications for research or practice because of the variables they do and do not include. For example, RIPPER suggested that the combination of positive peer expectation, no classroom disruptions and higher parental involvement were predictive of high achievement. This combination described over 10% of the sample, and predicted high achievers at over twice the rate of chance. This combination of attribute-values is noteworthy because it does *not* include parental education—a predictor that was prominent in regression and

was included in many of the rules generated by rule induction. All three predictors that were selected are also easier for schools and policies to intervene on relative to parental education.

The rule induction with the larger dataset identified additional variables that Thomas (2006) may have found relevant to include in her model. Experience of being held back in elementary or middle school is a student factor (and arguably also a school and family factor) that predicted lower achievement.

Several rules suggested implications for educational theory and practice. For example, the C4.5 rule induced from the large dataset, which indicated that those who were ever held back in elementary or middle school being are associated with lower achievement, raises several questions: (1) why is the variable so predictive (and predictive over others), (2) why are there so many students (13%) who were held back, (3) methodologically, is there value to including variables like this one, that are not purely school, student or family factor, but a hybrid of these factors? The first rule induced by RIPPER from the small dataset (high achievement's association with: peers expect college, no disruptions and parental involvement is 68th percentile or above), suggest the importance for schools to foster college expectations, involve parents and promote a positive (non-disruptive) learning environment, and that may be at least as important as demographic factors that are difficult to intervene on—e.g., parental education or household income.

Table 24. Number of rules, interesting rules, and false alarms discovered by algorithm (Study 1)

	Variables considered	Rules	Interesting	False alarm
CBA	19	67	*	*
RIPPER	19	7	1	
	1372	18	3	1
PART	19	31		
	1372	16	2	
C4.5	19	8		1
	1372	23	1	1
CART	19	2		
	1372	2		
C5.0	19	12	1	
	1372	14	2	2
QUEST	19	5		
	1372	8	1	
Total		213	11	5

* Note: As mentioned in section 3.2.5, interestingness analysis was not conducted for CBA due to intractable number of rules.

Table 25. Interesting rules discovered by ruleset induction (Study 1)

Set of predictors & rule origin	**Consequent**	**Reasons for interest**	**Reason for disinterest or caution in interpretation**
Been held back in ES or MS (C4.5, large, R21 & R22)	Low achieving	Predictor not considered by Thomas. Applies quite widely (13%/15%)* and is fairly accurate in classifying those whom the rule applies (95%/93%, where random chance would predict 75%). First predictor to be selected by algorithm, indicating strength of prediction of low achievers relative to others. Same predictor was selected across multiple algorithms. Potentially significant implications for educational theory and practice.	Might already be well-known by the field.
Peers expect college, no disruptions and parental involvement is 68th percentile or above (RIPPER, small, R1)	High achieving	Specific combination of predictors not highlighted by Thomas. Applies quite widely (11%/12%)* and is fairly accurate in classifying those whom the rule applies (50%/56%, where random chance would predict 25%). First rule to be selected by algorithm, indicating strength of prediction relative to others. Potentially significant implications for educational theory and practice.	Might already be well-known by the field.

Table 25 continued

Set of predictors & rule origin	Consequent	Reasons for interest	Reason for disinterest or caution in interpretation
SES is less than 89th percentile, not taking higher level courses in 8th grade, and parents are not doctors, and (but) read on own 6hr+/wk in 8th grade. (QUEST, large, R2)	High achieving	Reading and course-taking predictors were not considered by Thomas. Predictive validity and generality are reasonable (see next column). Potentially significant implications for educational theory and practice. Follow-up analysis found that the voracious reader condition applied to 5%/6% of those whose SES is less than 89th percentile, not taking higher level courses in 8th grade, and parents are not doctors, and when applied, was correct 36% (vs 14%/13% chance).	Rule does not apply very widely (3%), and is not extremely accurate (42%/33%, where random chance would predict 25%).
SES is higher (64th-93rd percentile), has been counseled about drug/alcohol and sent to the office more than twice in 8th grade. (C5.0, large, R9)	Not high achieving	Behavioral factors were not considered by Thomas. High predictive accuracy (100%)	Applies to a small group of students (2%/0.8%) Might be well-known in field.
If SES is higher (64th+ percentile), and behavior is good (haven't spoken to counselor about drug/alcohol abuse and have not been sent to office more than twice), and parent had thought about S's test scores being probably not good enough to qualify for loan/scholarship (i.e., selected "true" or "false" rather than "haven't thought about it") (C5.0, large, R11)	High achieving	Behavioral factors and 8th graders' parents' beliefs about college financial aid were not considered by Thomas. Applies quite widely (21%/22%)* and is fairly accurate in classifying those whom the rule applies (59%/50%, where random chance would predict 25%).	[Follow-up analysis] The contributions of the behavior and financial aid variables to the prediction are somewhat small (coverage was 71% and lift was 1.32/1.18 when examining only those whose SES is 64th+ percentile)

Table 25 continued

Set of predictors & rule origin	Consequent	Reasons for interest	Reason for disinterest or caution in interpretation
Parental educational attainment is more than high school and less than a 4-year college degree, parent expects college, student does not feel unsafe in school and fewer than 10 Black teachers in school. (C5.0, small, R8)	High achieving	Specific combination of predictors not highlighted by Thomas. Applies quite widely (22%/21%)* and somewhat accurate in classifying those to whom the rule applies (34%/35%, where random chance would predict 25%). [Follow-up analysis] The rule was even more predictive when examining only those whose parental educational attainment was more than high school and less than college degree, with coverage of 33%/32% and confidence of 34%/35% where random would predict 22%. Potentially significant implications for educational theory and practice.	Predictive validity is not extremely high. May not have practically significant implications for research and practice.
[Excluding 12% of respondents who were predicted to be high achieving by the first rule based on a combination of SES, academic, family and school demographic factors] If student was not held back, does not have a math teacher who teaches gifted & talented program, does not attend a religious school, and locus of control is at or lower than 37th percentile. (PART, large, R2)	Low achieving	Locus of control and whether math teacher teaches gifted/talented program were not considered by Thomas. Applies quite widely (28%/24%)* and somewhat accurate in classifying those to whom the rule applies (93%/87%, where random chance would predict 71%/78%). Potentially significant implications for educational theory and practice.	Predictive accuracy is not as compelling in the test set (lift – 1.10, as opposed to 1.30 in training set). Might be well-known in field. May not have practically significant implications for research and practice.

Table 25 continued

Set of predictors & rule origin	Consequent	Reasons for interest	Reason for disinterest or caution in interpretation
[Excluding 23% of respondents who were predicted to be low achieving by the first 4 rules based mainly on sense of safety and parent education] Doing 7-12 hours of homework outside of school, and peers expect college (PART, large, R5 & R6)	High achieving	Specific combination of predictors not identified by Thomas as being more predictive than others [Follow-up analysis] Applies reasonably widely (17%/14% for just the homework condition, 14%/9% for both homework and peer college expectation conditions)* and somewhat accurate in classifying those to whom the rule applies (51%/45% for just the homework condition where random chance would predict 35%/32%, and 55%/53% for both conditions where random chance would predict 30%/28%). Potentially significant implications for educational theory and practice.	Reduced generality and accuracy in test set. Might be well-known in field. May not have practically significant implications for research and practice.
SES is 58th percentile or higher, in higher achieving classes and never sent to office for misbehavior (RIPPER, large, R1)	High achieving	Includes predictors not considered by Thomas. Specific combination of predictors not identified by Thomas as being more predictive than others Applies somewhat widely (7%) and quite accurate in classifying those to whom the rule applies (87%/67%, where random chance would predict 25%).	Reduced generality and accuracy in test set. Might be well-known in field. May not have practically significant implications for research and practice.

Table 25 continued

Set of predictors & rule origin	Consequent	Reasons for interest	Reason for disinterest or caution in interpretation
[Not in above high achieving group] SES is 36th percentile or higher, parent has expectations for PSE and parent does not think child's test scores will be too low to qualify for college financial aid. (RIPPER, large, R2)	High achieving	Includes predictors not considered by Thomas. Applies somewhat widely (10%/12%) and quite accurate in classifying those to whom the rule applies (71%/49%, where random chance would predict 25%). Potentially significant implications for educational theory and practice.	Reduced generality and accuracy in test set. Might be well-known in field. May not have practically significant implications for research and practice.
[Not in above two high achieving groups] Locus of control is 54th percentile or above, autonomy is 56th percentile or higher, watches over 5 hours of TV a day in the weekend, student goes home after school, and household has one income earner. (RIPPER, large, R3)	High achieving	Includes predictors not considered by Thomas. Applies somewhat widely (2%/4%) and quite accurate in classifying those to whom the rule applies (79%/47%, where random chance would predict 25%).	Reduced generality and accuracy in test set. Might be well-known in field. May not have practically significant implications for research and practice.

*"Small" and "large" refers to dataset with 19 and 1372 predictors, respectively. "R##" refers to the rule number indicated in Appendix C.

**Percentages in parentheses indicate predictive validity for the training set (first percentage) and test set (second percentage). Only one number (the more conservative of the two) is indicated if the two were within a percentage point difference.

Table 26. Rules that initially seemed interesting but were not interesting after further investigation (Study 1)

Set of predictors	Consequent	Reason for initial interest	Reason for retraction in interest
Student's SES is above the 64th percentile, attends a school with few ELL students (<=10%), has generally good behavior (never been held back, cut classes, spoken to counselor about drug/alcohol abuse), neither parent nor social studies teacher is not proficient in a non-English language, and parent does not believe that their student's test scores will not be good enough to qualify for college financial aid (C4.5, large, R8)	High achieving	Applies quite widely (16%)* and is very accurate in classifying those whom the rule applies (69%/53%, where random chance would predict 25%). Includes variables not considered by Thomas.	Variable about social studies teacher's English fluency seems somewhat arbitrary given that only a subset of students had a social studies teacher respond to the survey. Accuracy in test set is higher than chance, but considerably lower than the test set (see column to the left). Difficult to think of clear implications for theory and practice. [Follow-up analysis] Financial aid condition applies to 88%/83% of the subgroup to which the rest of the conditions apply, and has a lift of (only) 1.10/1.07.
Parental educational attainment is more than high school and less than a 4-year college degree and student does not feel unsafe in school (C4.5, small, R1)	High achieving	Specific combination of predictors not highlighted by Thomas. Applies quite widely (37%/35%)* and seemed reasonably better than chance (25%) in classifying those whom the rule applies in the training set (31%). Sense of school safety was selected as the first predictor for the students whose parents' educational attainment was as indicated to the left. Potentially significant implications for educational theory and practice.	Worse-than-chance prediction in the test set (22%).

Table 26 continued

Set of predictors	Consequent	Reason for initial interest	Reason for retraction in interest
SES is lower than 64th percentile (but) score academically high in 12th grade, have parents who have a higher degree (MA or higher) and have a not-very-low score on locus of control (16th percentile or higher). (C5.0, large, R4)	High achieving	Included variable not considered by Thomas (locus of control). Specific combination of predictors not highlighted by Thomas (low-SES and high parent education). Applies reasonably widely (8%/10%)* and was very accurate in classifying those whom the rule applies in the training set (46%, where random chance would predict 25%).	Much lower (and near-chance) prediction in the test set (29%).
Shows that among SES is at least 64th percentile, have spoken to a counselor about drug/alcohol abuse, and teachers have spoken individually with at least 10 students' parents about student performance. (C5.0, large, R8)	Low achieving	Included variables not considered by Thomas. Applied reasonably widely (4%)* and was very accurate in classifying those whom the rule applies in the training set (100%, where random chance would predict 75%).	Lower-than-chance prediction in the test set (71%, where chance is 75%).
[Not in first three high achieving groups identified by RIPPER] Never held back in school, disagrees that chance and luck are important in life, and expects to do professional, business, or managerial work. (RIPPER, large, R4)	High achieving	Included variables not considered by Thomas. Applied reasonably widely (4%)* and was very accurate in classifying those whom the rule applies in the training set (51%, where random chance would predict 25%).	Lower-than-chance prediction in the test set (23%, where chance is 25%).

*"Small" and "large" refers to dataset with 29 and 1372 predictors, respectively. "R##" refers to the rule number indicated in Appendix C.

**Percentages in parentheses indicate predictive validity for the training set (first percentage) and test set (second percentage). Only one number (the more conservative of the two) is indicated if the two were within a percentage point difference.

4.1.2.4 Results from association rule mining

The number of rules generated by association rule mining for each of the subgroups is described in Table 27. For each subgroup, between 300,000 to 400,000 rules of lengths 2 to 3 described at least 25% of the high achievers in the generation set, and of those, approximately 250,000 rules described at least 25% of the high achievers in the screening set. Among those rules, just a few hundred had a positive likelihood ratio of four or greater on the training set, and fewer than 200 met that cut-off in the test set. There were very few length-two rules with high positive likelihood ratios.

Table 27. Number of association rules generated by subgroup (Study 1)

Subgroup & parameters	# of rules generated	# of rules after screening	# of rules with PLR >= 4	# of rules with length 2 with PLR >= 1.5
Lower income TL income <$25,000 (104/603 are high achieving, min sup .25, max len = 3)	400,545 (1346 length 2)	259,103 (1124 length 2)	434/140	107/100
Higher income TL income >=$25,000 (103/303 are high achieving, min sup .25, max len = 3)	336,526 (1278 length 2)	262,879 (1127 length 2)	1043/194	75/54
Lower parental education TL parental education is less than a 4yr degree (103/999 are high achieving, min sup .25, max len = 3)	338,966 (1284 length 2)	267,847 (1148 length 2)	249/44	87/77
Higher parental education TL parental education is at least a 4yr degree (89/196 are high achieving, min sup .25, max len = 3)	339,065 (1262 length 2)	242,505 (1104 length 2)	528/70	29/13

Note. In the last two columns, the first number represents the number of rules in the training set, while the second represents that in the test set.

Table 28 summarizes the association rules with length 2 (i.e., only one condition in the antecedent) that were found among the four subsamples of parental education and income (see

Appendix E, Table 71 to Table 74 for individual rules). Perhaps not surprisingly, attending higher-level classes in 8th grade was associated with higher achievement in 12th grade, regardless of income or parental education. However, most variables were associated with high achievement within only some subgroups and not others. For example, enrollment in gifted/talented programs, sense of school safety, and study of music were associated with 12th grade achievement only among students with lower parental education or who are from lower income households. Students in these subgroups who were enrolled in gifted/talented programs in 8th grade were 2-5 times more likely to be higher achieving in 12th grade than those who were not. Those who studied music in 8th grade were 1.5 to 2 times more likely to be higher achieving in 12th grade than those who did not. Similarly, not working for pay was associated with higher 12th grade achievement for those from higher income households, but not among others. There were variables that were associated with the outcome in the training set, but not in the test set. Of 298 rules that applied at least 1.5 times more to the high achievers than low achievers in the training set, 172 applied with at least the same relative probability to the test set. Examples of rules that appeared promising but did not perform well on the test set are provided in Table 29.

Table 28. Factors associated with high achievement among parental education and income subgroups (Study 1)

Category	**Details**	**Groups to which rules applied** (PLR=relative probability)
[Student] Enrolled in gifted/talented program in 8th grade	Enrolled in gifted/talented program in 8th grade according to parent or student. Participated in academic honors society.	**Low parental education** (TPR .32-.26; PLR=2.2-3.5/2.1-4.8) **Low income** (TPR .25-.37/.29-.25; PLR=2.3-3.5/2.9-5.6) Note: Parent report more predictive than student report for gifted/talented program participation.
[Student] Attends higher level classes in 8th grade	Taking algebra, in higher ability group for math and/or English. Does not attend regular math.	**Low parental education** (TPR=.41-.53/.48-.68; PLR=1.9-3.0/1.7-2.8) **High parental education** (TPR=.49-.53/.28-.47; PLR=1.8-2.3/1.5-1.7) **Low income** (TPR .41-.52; PLR=2.0-3.0/2.9-4.0) **High income** (TPR .43-.57/.32-.51; PLR=2.2-3.5/1.5-4.5) Note: For high parental education and high-income groups, math only (not English) was predictive. Prediction for high parental education group was not as strong relative to other groups.
[Student] T believes Ss class is higher achieving than average	T considers achievement of Ss class to be higher achieving relative to average, according to English, Science and average of all 4 core subject area teachers.	**Low parental education** (TPR=.21-.4/.12-.36; PLR=3.3-5.5/1.7-2.8) **Low income** (TPR .25-.42/.35-.48; PLR=3.9-6.3/2.6-6.7)
[Student] High locus of control in 8th grade	Highest quartile on 2 types of locus of control composite. Disagrees that chance/luck is important in life.	**Low parental education** (TPR=.44-.62/.43-.64; PLR=1.9-2.7 /1.9-2.1) **Low income** (TPR=.29-.52/.32-.64; PLR=1.6-3.0 / 1.7-2.3) **High income** (TPR=.47-.57/.45-.68; PLR=1.7-2.4 / 1.7-1.8)
[Student] 8th grader expects postsecondary education after high school	Expects to go to school after high school. Expects to attend a college prep program after high school.	**Low parental education** (TPR=.39-.49/.38-.45; PLR=2.0-2.4/2.1-2.3) **Low income** (TPR=.44-.47/.42-.45; PLR=2.1-2.5/1.9-2.5) **High income** (TPR=.46-.47/.55-.61; PLR=2.4-3.1/2.2-2.6)
[Student/Family] Parents expect 8th grader to get more schooling after college	Mother expects higher school after college. Father expects higher school after college.	**High income** (TPR=.38-.50/.48-.52; PLR=.17-.16/.19)

Table 28 continued

[Student/Family] Parent expects child's HS test score to not be bad	Parent does *not* believe that their child's test score would not be good enough for child to qualify for financial aid	**Low parental education** (TPR = .70/.78; PLR=1.7/1.6) **Low income** (TPR = .78/.74; PLR=1.8/1.5) **High income** (TPR = .78/.68; PLR=1.7/1.5)
[Student] 8th grader studies music	Attends music at least once a week. Participated in band or orchestra. Child studies music outside regular school.	**Low parental education** (TPR=.53/.57; PLR=1.6/2.1) **Low income** (TPR=.27-52; PLR=1.7-2.5/1.5-2.0)
[Student] 8th grader studies foreign language	Enrollment/attendance in foreign language course.	**Low parental education** (TPR=.25-.27/.19-.21; 2.0-2.5/1.7-1.8) **High income** (TPR=.40/.49; PLR=2.6/2.4)
[Student/family] 8th grader does not work for pay	8th grader does not work for pay	**High income** (TPR = .39/.42; PLR=1.8/2.5)
[Student] 8th grader expects a professional, managerial or business occupation	8th grader expects a professional, managerial or business occupation at age 30	**Low income (**TPR = .40/.42; PLR=1.9/1.6) **High income** (TPR = .40/.42; PLR = 1.7/1.7)
[School] 8th grader's school has moderate attendance issues and minor other behavioral issues	Student tardiness and absenteeism are considered a "moderate" problem (in a scale of "serious", "moderate", "minor" and "not a problem") by student. Robbery/theft, and verbal abuse of teachers at school are considered either a "minor" and/or "moderate" problem by student. Class cutting is considered "not a problem" by teacher or student.	**Low parental education** (TPR=.20-.47/.26-.36; PLR=1.6-2.1/1.5-2.2) **Low income** (TPR=.26-.37/.19-.42; PLR=1.6-4.1/1.4-3.3) Note: Class cutting rule applied to high income group as well.
[School] 8th grader's math class emphasizes Algebra	Algebra is a major topic in student's math class.	**Low income** (TPR=.34/42; PLR=1.5)
[School] 8th grader's school has formal admissions procedures	School has formal admissions procedures.	**Low parent education** (TPR=.23/.24; PLR=2.0) **High income** (TPR= .38/.39; PLR=1.9/2.3)
[School] 8th grader is challenged at school	Parent "strongly agrees" that their 8th grader is challenged at school.	**Low parent education** (TPR=.32/.31; PLR=2.1/2.6) **Low income** (TPR= .32/.39; PLR=2.1/3.6)
[School] Parent believes 8th grader's school sets realistic standards	Parent "strongly agrees" that school sets realistic standards.	**Low parent education** (TPR=.27/.29; PLR=1.6/2.0)
[School] 8th grader feels safe at school	8th grader "strongly disagrees" that they do not feel safe at their school	**Low parent education** (TPR=.44/.52; PLR=1.5/1.8)

Table 28 continued

		Low income (TPR= .49/.55; PLR=1.7/1.8)
[School] Parent is very satisfied with 8th grader's education	Parent "very satisfied" with education child has received	**High parent education** (TPR=.68/.59; PLR=2.1/1.6)
[Family] Parent of 8th grader goes to museums	Parent reports going to history, art, and/or science museums	**Low parent education** (TPR=.26-.27/.33-.34; PLR=1.6/1.7-1.9) **Low income** (TPR=.26-.32/.23-.32; PLR=1.6-2.4/1.7-3.7)
[Student] 8th grader goes to history museums	Parent reports that 8th grader goes to history museums	**Low parent education** (TPR=.43/.47; PLR=1.6/1.9) **Low income** (TPR=.47/.35; PLR=2.7/1.7)

Table 29. Examples of variables that were associated with high achievement in the training set, but not test set (Study 1)

Category	**Details**	**Groups to which rules applied** (RP=relative probability)
High self-concept	Highest quartile on a self-concept composite. Believes they are a person of worth.	**High income** (RP=1.6-1.7 /1.2-1.4)
Coming prepared for class	Seldom come to class without pen/pencil, homework or books.	**Low parental education** (RP=1.6/1.3-1.5) **High parental education** (RP=1.5/.92) **Low income** (RP=1.5-1.9/1.3-1.4) **High income** (RP=2.1-2.3 /1.2)
Watches over 4 hours of TV every day	Watches 4-5hrs of TV on weekdays and/or 5+ hours on weekends.	**Low income** (RP=1.7-3.1/1.2-1.4) **High income** (RP=1.7/1.5)
8th grader goes to art museums	Parent reports that their 8th grader goes to art museums	**High parental education** (RP=1.6/1.1) **Low income** (RP = 1.9/1.2) **High income** (RP=1.5/0.9)
Participation in science fair	8th grader reports participating in science fair	**Low parental education** (RP = 1.5/1.3)
Participation in vaguely defined enrichment groups	"Child ever involved in community group" "Child studies other skills outside regular school day"	[Community group] **Low parental education** (RP=1.5/.86) [Other skills] **Low income** (RP=2.6/.85)
Female	Student is female	**High parental education** (RP=1.8/.84)

Rules of length 3 (i.e., two conditions in the antecedent) were difficult to summarize cleanly and comprehensively, since there were so many of them, and were interrelated in overlapping ways within and across categories to varying degrees. In the following two subsections, I provide two sets of findings that illustrate a possible way to summarize association rules for descriptive education research.

How 12th grade achievement relates to household income, parental education and participation in gifted/honors programs in 8th grade

Table 30 and Table 31 provide an analysis of rules that included gifted/honors program participation as a condition. As mentioned in Table 28, and detailed in Table 29, enrollment in a gifted/talented program in 8th grade (BYS68A, BYP51), and/or participation in academic honors society (BYS82O) in 8th grade was associated with high achievement in 12th grade only **among those whose parents did not have a college degree, and those from low income households.** For these subgroups, high achievers were 2-4 times more likely to have been in a gifted and talented program than lower achievers. About 26-38% of the high achievers tended to be gifted/honors while this was true only for about 7-15% of the lower achievers. 20-46% of those who were in a gifted, talented or honors program in 8th grade became high achieving in 12th grade, which is 2.3-3.6 times the rate of high achievers among students who were not in gifted programs. Furthermore, **among students from low income households,** parent satisfaction of their 8th grader's education and belief in their future test scores and grades in increased the association between gifted/honors participation and high achievement. High achievers were 4-6 times more likely to have both been in a gifted/talented program and have at least one of the parent conditions. Also, if these conditions were met, the student was 3.6-4.2 times more likely to be high achieving than not.

Among those whose parents had at least a college degree, being in a gifted/honors program in 8th grade, by itself, was not associated with high achievement in 12th grade. Approximately 60% of students of students in this group were in a gifted/honors program. However, their 12th grade achievement score tends to be higher if they were gifted, talented and/or honors student in 8th grade, *and* at least <u>one</u> of the following were also true:

In 8th grade, the student…

- Did well/fine academically
- Was part of a quite selective gifted/talented program
- Had good behavior
- Had few behavioral and academic issues
- Did not participate in speech/language-related programs
- Was not involved in neighborhood clubs, and/or band/orchestra
- Did not attend home economics and/or consumer ed at least once a week
- Went home directly after school
- Did not often count on parents to solve problems

And/or the student's…

- Household spoke only English
- Household never had adult neighbors at home when student returned from school
- Parent knew the parent of the child's third friend

And/or, the student's school…

- Attended a school that was departmentalized
- Had 5.5-7 class periods in a school day

High achievers were at least 5 times more likely to be in a gifted/honor program and to have any of one of these characteristics than lower achievers. About 25-30% of the high achievers tended to have these 8th grade characteristics, while only 0-5% of the lower achievers tended to have these characteristics. 82-91% of those who were in a gifted, talented or honors program and had at least one of the above characteristics in 8th grade was high achieving in 12th grade, which is 2.1-2.4 times the rate of high achievers among students who did not have these characteristics.

Among those who were from higher income households, being in a gifted/honors program in 8th grade (approximately 60% of students), by itself, was not associated with high achievement in 12th grade. However, their 12th grade achievement score tends to be higher if they were gifted, talented and/or honors student in 8th grade, *and* one of the following were also true:

In 8th grade…

- Student attended algebra at least once a week
- Student agreed that discipline is fair
- Friends neither encouraged or discouraged student from taking algebra
- Parent did not expect child to be able to earn money for postsecondary education

High achievers were at least 5 times more likely to have any of these combinations of characteristics than lower achievers. About 21-28% of the high achievers tended to have these 8th grade characteristics, while only 0-6% of the lower achievers tended to have these characteristics. 70-96% of those who were in a gifted, talented or honors program and had at least one of the above characteristics in 8th grade was high achieving in 12th grade, which is 2.5-3.3 times the rate of high achievers among students who did not have these characteristics.

Table 30. Association between 12th grade achievement and participation in 8th grade gifted/honors program by income and parental education subgroups (Study 1)

Group	Var name	Proportion of gifted students among high achievers (TPR)	Proportion of gifted students among lower achievers (FPR)	TPR / FPR (PLR)	Proportion of high achievers among gifted (Precision)	Proportion of high achievers among not-gifted (FOR)	Precision / FOR (RP)
Low income	BYS68A	.38	.15	2.53	.35	.13	2.62
	BYS82O	.26	.08	3.25	.4	.14	2.81
	BYP51	.33	.08	4.13	.46	.13	3.51
Low parental educ	BYS68A	.32	.15	2.13	.2	.08	2.34
	BYP51	.26	.07	3.71	.3	.08	3.57

TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving. For high parental education group, TPR .61-.62; PLR=.73-.87. For high income group, TPR .57-.63; PLR=.66-.91.

Table 31. Additional conditions that increase associations between 12th grade achievement and participation in 8th grade gifted/honors program by income and parental education subgroups (Study 1)

Grp	Additional condition	TPR	FPR	PLR	Prec	FOR	RP
LI	Parent believes S's grades will be good enough to qualify for college financial aid.	.3	.06	4.8	.5	.13	3.71
LI	Parent believes S's test scores will be good enough to qualify for college financial aid.	.34	.08	4.31	.47	.13	3.63
LI	Parent very satisfied with education 8th grader has received	.29	.05	6.26	.57	.13	4.21
LI	Watches over 5 hrs of TV a day on weekends	.24	.04	5.99	.56	.14	3.92
HP	5.5-7 class periods in a school day	.24	.04	5.4	.82	.4	2.06
HP	8th grader comes home directly after school	.3	.04	6.65	.85	.38	2.23
HP	8th grader did not attend consumer ed at least 1/wk	.28	.04	6.23	.84	.39	2.18
HP	8th grader did not attend home economics at least 1/wk	.28	.03	8.28	.87	.38	2.28
HP	8th grader did not participate in band/orchestra	.24	.03	7.17	.86	.39	2.17
HP	8th grader did not participate in debate/speech team	.29	.04	6.44	.84	.38	2.2
HP	8th grader did not participate in foreign language club	.26	.04	5.81	.83	.39	2.12
HP	8th grader did not participate in neighborhood clubs/programs	.27	.01	23.99	.95	.38	2.5
HP	8th grader did not talk to counselor about discipline problems	.29	.03	8.6	.88	.38	2.31
HP	8th grader did not talk to other adult about drug/alcohol abuse	.28	.02	12.41	.91	.38	2.4
HP	8th grader does not often count on parents to solve problems	.27	.01	23.99	.95	.38	2.5
HP	8th grader has never been in a fight	.27	.03	8.04	.87	.39	2.26
HP	8th grader has good behavior (doesn't skip classes)	.31	.04	6.85	.85	.38	2.26
HP	8th grader never offered drugs for sale	.31	.04	6.85	.85	.38	2.26
HP	8th grader never sent to office	.24	.03	7.21	.86	.39	2.17
HP	8th grader never sent to office with school work problems	.31	.04	6.85	.85	.38	2.26
HP	8th grader not involved in bilingual program	.28	.04	6.23	.84	.39	2.18
HP	8th grader not involved in boys'/girls' club	.28	.03	8.32	.87	.38	2.28
HP	8th grader's parent knows parent of child's third friend	.24	.02	1.84	.9	.39	2.3
HP	8th grader's school has no/low attrition	.24	.04	5.4	.82	.4	2.06
HP	8th grader's school is departmentalized	.29	.03	8.6	.88	.38	2.31
HP	Adult neighbor never at home when R returns from school	.25	.02	11.26	.9	.39	2.32
HP	English is only language spoken at 8th grader's home	.28	.04	6.23	.84	.39	2.18
HP	In advanced, enriched accelerated math	.29	.03	8.6	.88	.38	2.31

Table 31 continued

HP	Low percentage of students in gifted and talented	.27	.02	12.09	.91	.38	2.38
HP	Low percentage of students in remedial reading	.29	.04	6.44	.84	.38	2.2
HP	Parents trust 8th grader to do what they expect	.24	.04	5.4	.82	.4	2.06
HP	Physical abuse of teachers is not a problem	.26	.03	7.76	.87	.39	2.23
HI	Agrees that discipline is fair	.21	.01	21.37	.92	.29	3.16
HI	Does not expect child to be able to earn money for postsecondary education	.21	.01	42.74	.96	.29	3.31
HI	Enrolled in gifted/talented program	.23	.03	7.77	.8	.29	2.77
HI	Friends don't impact algebra	.28	.06	4.69	.71	.28	2.5

TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving.

How 12th grade achievement relates to household income, parental education and math course-taking in 8th grade

As detailed in Table 32, for all 4 groups, being in a higher-level math class or not being in "regular" math in 8th grade was associated with high achievement in 12th grade. High achievers were 2-4 times more likely to have been in a higher-level math class than lower achievers. About 28-67% of the high achievers tended to have these 8th grade characteristics, while only about 16-28% of the lower achievers tended to have these characteristics. 25-58% of those who attended higher level math class and/or were not in regular math in 8th grade became high achieving in 12th grade, which is 1.5-4 times the rate of high achievers among students who were not in gifted programs. This rate was higher for the lower parental education and lower household income groups (3-4) than for the higher parental education and higher household income groups (1.5-2.5).

Among those whose **parents had no college degree**, 12th grade achievement score tended to be higher if they were studying higher math or not taking regular math in 8th grade, *and* one of the following were also true:

In 8th grade, the student…

- Had high locus of control
- Did academically well/fine
- Had good behavior in school
- Had college expectations
- Felt safe in school

And/or the student's…

- Household had no adult relatives or older sibling at home when student came home
- Household had no specific place to study

Among those from a **low-income household,** 12th grade achievement score tended to be higher if they were studying higher math or not taking regular math in 8th grade, *and* one of the following were also true:

In 8th grade, the student…

- Had high self-concept
- Had high locus of control
- Did academically well/fine
- Had good behavior in school
- Felt safe in school / attended a school with few safety issues

And/or the student's…

- Parent regularly talked to child about school experiences
- Family had rules about programs that the child may watch
- Household never had adult neighbor or younger sibling at home when student came home

- Household had no specific place to study
- Parent did not think student will be able to earn their own money for college, or that relatives would pay for it
- Friends neither encouraged/discouraged student from taking algebra

And/or the student's school

- Had few safety issues
- Placed high priority on learning (as rated by parents)
- Has a school newspaper
- Involves parents in some school decisions
- Had 24-33 students enrolled in class

Among those from a **high-income household**, 12th grade achievement score tended to be higher if they were studying higher math or not taking regular math in 8th grade, *and* one of the following were also true:

In 8th grade, the student…

- Had high self-concept
- Had high locus of control
- Did academically well/fine
- Had good behavior in school
- Communicates with parents and teacher about school work and/or school activities
- Has NOT participated in chorus, scouting, neighborhood clubs/programs, summer programs, and/or varsity sport.

- Has NOT attended biology, earth science, art, computer ed, or sex ed at least once a week.
- Expects college prep program in high school

And/or, the student's

- Family included father and mother (vs missing one/both, or having a guardian for one/both)
- Family had high socioeconomic status and/or electric dishwasher
- Parent did not think student would be able to earn their own money for postsecondary education, that they can pay for college without assistance
- Parent believed that getting financial aid for college is a feasible option for their child
- Household never had adult neighbor, other adult relative and/or younger sibling home when student comes home from school
- Parents/guardians wanted 8th grader to take algebra
- Closest two peers attended the same school
- Friends neither encouraged/discouraged student from taking algebra

And/or the student's school…

- Had a safe/positive environment
- Low percentage of students in remedial reading and/or special education
- Required a full year of PE, and/or did not require a specific amount of instructional time for family life/ sex ed.
- Offered drama club, math club, science fair, student newspaper and/or foreign language course

- Did NOT offer religious organization, and/or debate/speech
- Suspended students in for the first occurrence of alcohol possession, weapon possession, alcohol use, illegal drug use, or injury to other student, and/or repeated occurrence of smoking.
- Teachers spent no/minimal time outside of school hours for record keeping and/or coordinating curriculum
- Teachers had a lot of influence in assigning high school courses
- Teachers who filled out the survey were White, non-Hispanic, and/or did not have a BA in Education

High achievers were at least 4 times more likely to have any of these combinations of characteristics than lower achievers. 25-45% of the high achievers tended to have these 8th grade characteristics, while only 2-14% of the lower achievers did. 32-83% of those who were in high level math and had at least one of the above characteristics in 8th grade was high achieving in 12th grade, which was 2.4-5.6 times the rate of high achievers among students who did not have these characteristics. More detailed data tables are provided in Appendix E.

Table 32. Association between 12th grade achievement and 8th grade math course-taking by income and parental education subgroups (Study 1)

Grp	Sub-category	Var name	Propn of higher lv math takers among high achievers (TPR)	Propn of higher lv math takers among lower achievers (FPR)	TPR/ FPR (PLR)	Pron. of high achievers who took higher lv math (Precision)	Propn of low achievers who took higher lv math (FOR)	Precision/ FOR (RP)
LP	Not in regular math	BYS67B	.4	.16	2.5	.22	.08	2.94
	Studies higher math	BYP53	.48	.21	2.29	.21	.07	2.96
		BYS67C	.5	.17	2.94	.25	.06	3.9
HP	Not in regular math	BYS67B	.43	.25	1.72	.59	.39	1.52
	Studies higher math	BYS67C	.55	.3	1.83	.6	.35	1.73
LI	Not in regular math	BYS67B	.45	.17	2.65	.36	.12	2.93
	Studies higher math	BYP53	.52	.2	2.6	.35	.11	3.16
		BYS67C	.57	.18	3.17	.4	.1	4.04
HI	Not in regular math	BYS67B	.40	.19	2.11	.52	.28	1.88
	Studies higher math	BYP53	.52	.23	2.26	.54	.24	2.21
		BYS67C	.54	.21	2.57	.57	.23	2.47

TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving.

4.1.3 Summary of Study 1 results

Key findings from Study 1 are summarized in Table 33. Ruleset mining and regression with 19 variables emphasized similar kinds of family and school-level variables such as parental education, school safety, socioeconomic status and number of Black teachers at the school. It was immediately clear with ruleset models, and not as much with regression, that parental education was sufficient to explain most of the explainable variance. With an expanded dataset, instead of relying primarily on parental education, ruleset models relied on factors such as 8th grader's academic achievement and behavior, family socio-economic status, and parent expectations. However, the overall model accuracies remained roughly the same when the dataset was expanded.

The consistency in model accuracy across datasets and modelling approaches suggests that each of the models are highlighting different relationships within the data, that there are likely multiple adequate ways to model the data, and that other criteria (e.g., usefulness) may need to be considered to understand the overall value of the model to the researcher. For example, for those who are seeking ways to improve Black student achievement through public schooling, the ruleset models that highlight the potential importance of 8^{th} grade achievement, behavior and parent expectations may be more useful than the models that account for most of the outcome variance using parental education.

While ruleset induction approaches provided several alternative ways to understand the relationship between predictors and the outcome, and had accuracy comparable to that of logistic regression, the individual rules with each ruleset tended to be difficult to understand and less reliable. Just 11 of the 213 rules provided information that was not already known through regression and performed reasonably well in the test set. Some rules were interesting because they included variables that regression did not include (e.g., being held back in ES/MS; amount of

reading in 8th grade; behavioral factors; parents' attitude towards financial aid in 8th grade; locus of control), while other rules were interesting because the set of variables identified as predictive did NOT include variables that were considered highly predictive by regression (e.g., relationship between high achievement with peer expectations of college, no disruption and medium to high parental involvement did not include parental education and socioeconomic status). Furthermore, one of the rules within the ruleset pointed out a relationship that comes somewhat close to contradicting a general trend found in regression. Despite the general strong positive relationship between parental educational achievement and high school achievement, a rule by C5.0 using the small dataset predicted that if the educational attainment is not particularly high but also not too low (more than high school and less than a 4-year college degree), achievement tends to be high if parent expects college, student does not feel unsafe in school and fewer than 10 Black teachers in school.

Association rule mining was a more direct and comprehensive approach to identifying factors that may predict high achievement among those with lower parental education, or lower-household income. Participation in honors or gifted and talented programs, attending higher-than-average level classes, parents' strong agreement that student is challenged at school, school safety, and study of music were more strongly associated (positive likelihood radio >= 2.5) with higher achievement in 12th grade for lower income students and students whose parents did not have a college degree. These factors alone were not as predictive of achievement among those in the other subgroups, and that was relatively easy to discern through combining and sorting the outputs. I currently cannot imagine a regression-based approach that would be as straight-forward and analogous to this process.

Table 33. Key findings from Study 1

Method	Findings
Logistic regression	Factors positively associated with 12th grade achievement included: Higher parental education, more homework out of school, higher family income (females), lower percentage of students receiving free/reduced lunch, fewer disruptions in school, greater sense of safety in school, fewer Black teachers. And also in Thomas's analysis: Parental expectation for college, having "good" peers, positive school climate.
Ruleset mining with 19 predictors	Parental education sufficient to determine much of the explainable variance. After parental education, school safety, number of Black teachers in school, and household income tend to explain much of the remaining explainable variance.
Ruleset mining with 1372 predictors	Higher SES, higher 8th grade academic achievement, higher parent expectation, and fewer behavioral issues in 8th grade account for much of the explainable variance in the outcome. Parental education, school type, locus of control, school structure/policy, reading on own and students' academic expectations explained much of the remaining explainable variance.
Examination of rules within rulesets	16 of 213 rules appeared to provide additional information to regression in the training set, but 5 seemed to be false alarms based on test set. Combinations of predictors predicted outcome well for some students. Factors not selected by Thomas were included in the expanded model. Was practically not feasible to calculate validity metrics for CBA rules.
Association rules—rules for household income (high, low) and parental education (high, low) subgroups	Participation in honors or gifted and talented programs, attending higher-than-average level classes, parents' strong agreement that student is challenged at school, school safety, and study of music were more strongly associated (PLR >= 2.5) with higher achievement in 12th grade for lower income students and students whose parents did not have a college degree. For those whose parents had a college degree and those from higher income households, participation in honors/gifted programs was more strongly associated with 12th grade achievement (PLR >=4) when in conjunction with some other factor. For the higher parental education group, there were many such factors, which tended to pertain to attendance in stable/positive schools and programs, or having good behavior. For those from higher income households, there were only three such factors, which seemed conceptually unrelated to one another (e.g., discipline is fair, friends do not impact students' decision to take algebra). If the student had taken a higher-level math class in 8th grade and/or not taken a remedial math, their likelihood of being in the high achieving group was 40-50%, which was 1.7 to 3.2 times the likelihood of the remaining sample being high achieving. The positive likelihood ratio was increased to at least 4 if another factor—such as school safety, high locus of control, and good behavior, parental expectation that students' test scores would be good enough for college financial aid—also applied.

Table 33 continued

Method	Findings
	Parents' assessment of the school seemed to be a greater factor for students from lower income families. Students' perception of their educational environment (of teachers, policies, usefulness of schoolwork) and SES seemed to be a greater factor for students from higher income families. Direct question to parent about expectation for college-after-high-school was associated with 12th grade achievement only for students from higher household income, while indirect question to parent and direct question to student were associated with outcome regardless of income. Not working for pay associated with higher achievement only for students from higher income families.

4.2 RESULTS FOR STUDY 2 (BYRNES & MILLER, 2007)

4.2.1 Replication

A side-by-side comparison of the replication and Byrnes & Miller's (2007) results, is shown in Table 34 and Table 35. The former shows the results of the hierarchical regression, while the latter display the zero-order and partial correlation (i.e., the proportion of variance in the outcome explained by the predictor after excluding the variance accounted for by other predictors). The directions and magnitude of the coefficient estimates, as well as the relative proportion of variance explained by the four factors, were almost identical, indicating a successful replication. Clear patterns observed in both Byrnes & Miller's and my replication results were as follows:

1. All factors together accounted for 76 percent of the variance in the outcome.
2. Distal factors including SES, parent expectations, student expectations and middle school GPA were each a significant predictor, together explaining 43% of variance in 12th grade math achievement when they were the only factors considered.
3. Among opportunity factors, math course-taking were significant predictors of 12th grade math achievement, explaining an additional 11 percent of variance beyond distal factors. Student perceptions of math emphasis and teacher responsiveness were not as predictive. Opportunity factors taken together, without accounting for any other factors, explained 45 percent of the outcome variance.
4. Propensity factors, including math achievement before 9th grade, explained an additional 22 percent of variance. Propensity factors taken together, without accounting for any other factors, explained 73 percent of the outcome variance.

5. Gender and race/ethnicity factors explained less than one percent of additional variance when all other factors were controlled. These factors taken together, without accounting for any other factors, explained 9 percent of the outcome variance.

In addition, both Byrnes and Miller and I found that the following variables were most correlated with the outcome: Each of the four distal factors such as 8th grade SES and parent expectations (correlation between .42 and .56), 1 year of general math (-.37), 1 year of geometry (.53), and 1 year of Algebra II (.37), math achievement before the start of 8th grade (.84), GPA in grade 9 and 10 (.44), and math self-concept (.40). Particularly noteworthy is that:

6. Math achievement before 9th grade had an extremely high correlation with the outcome, explaining approximately 70 percent (.84* .84) of the variance in the outcome variable, which is over 92% (70/76) of the total variance explained by the model.

The squared semi-partial correlation (the proportion of variance in the outcome that is explained uniquely by a predictor, which was not reported by Byrnes & Miller) was less than 1 percent for all variables except for math achievement before 9th grade (17%) and 1 year of geometry (1%). This means that deleting any one of the other variables besides the two just mentioned would not have made any difference to the total variance explained by the model.

Table 34. Prediction of 12th grade math achievement with NELS:88, with Byrnes & Miller's (2007) variables

Variable	Replication		Byrnes & Miller	
	Coef	ΔR^2	Coef	ΔR^2
Distal factors		.428		.430
SES in 8th gr (BYSES)	.821***		.790***	
Parent expectations in 8th gr (Pexp)	.445***		.478***	
Student expectations in 8th gr (Sexp)	.445**		.409**	
Middle school GPA (BYGRADS)	1.182***		1.215***	
Opportunity factors		.112		.112
General math ½ yr (gm_half)	-2.401**		-2.826**	
General math 1yr (gm_1)	-3.052***		-3.338***	
General math 1.5-2yrs (gm_2)	-3.078***		-3.477***	
Geometry ½ yr (geo_half)	1.452***		1.620***	
Geometry 1 yr (geo_1)	2.544***		2.660***	
Geometry 1.5-2yrs (geo_2)	2.156		2.237	
Algebra II ½ yr (al2_half)	1.591**		1.374**	
Algebra II 1yr (al2_1)	1.207***		.858**	
Algebra II 1.5-2yrs (al2_2)	.446		.261	
S perception of math emphasis (emph_m)	.111		.157*	
S perception of T responsiveness (t_rspnsv)	.232*		.132*	
Propensity factors		.224		.219
Math achiev before 9th gr (BYTXMIRR)	.975***		.678***	
Math GPA in 9th & 10th gr (GPA910_m)	.978***		.943***	
Efficacy for graduating HS (grad_eff)	.704**		.915**	
Plans to take SAT (SATplan)	1.161***		1.180***	
Math self-concept (m_selfcpt)	.310***		.175***	
Demographic factors		<.01		<.01
Female	-1.822***		-1.725***	
Black	-1.845***		-2.184***	
Hispanic	-.594		-.761*	
Asian	.877		.779	
Native American	-1.740		-2.078	

*$p < .05$, **$p < .01$, ***$p < .001$; N=8976 for replication, while 8969 for Byrnes & Miller. ΔR^2 result from a hierarchical regression where the distal factors are entered, followed by opportunity factors, propensity factors and demographic factors. The coefficient estimates are of the model that includes all predictors. The opportunity factors alone accounted for 45.1% (45.2% per Byrnes and Miller) of the outcome variance, propensity factors accounted for 73.1% (72.4%), and gender and race/ethnicity factors 9.1% (9.3%).

Table 35. Correlation of 12th grade math achievement with Byrnes & Miller's (2007) predictors

Note: Byrnes and Miller's statistics, if different from the replication set by over 2 percentage points, are underlined and indicated underneath.

Variable	Zero-order	Partial	Semi-partial
Distal factors			
SES in 8th gr (BYSES)	.45***	.08 ***	.04
	.42		.02
Parent expectations in 8th gr (Pexp)	.43***	.05 ***	.02
Student expectations in 8th gr (Sexp)	.42***	.05 ***	.05
Middle school GPA (BYGRADS)	.56***	.09 ***	.04
Opportunity factors			
General math ½ yr (gm_half)	-.11***	-.05 ***	.02
General math 1yr (gm_1)	-.37***	-.14 ***	.07
General math 1.5-2yrs (gm_2)	-.20***	-.12 ***	.06
Geometry ½ yr (geo_half)	.02*	.04 ***	.02
Geometry 1 yr (geo_1)	.53***	.15 ***	.07
Geometry 1.5-2yrs (geo_2)	-.01	.03	.02
Algebra II ½ yr (al2_half)	.10***	.05 ***	.02
Algebra II 1yr (al2_1)	.38***	.07 ***	.03
Algebra II 1.5-2yrs (al2_2)	-.01	.005	.00
S perception of math emphasis (emph_m)	.22***	.02	.01
S perception of T responsiveness (t_rspnsv)	.17***	.04 ***	.02
Propensity factors			
Math achiev before 9th gr (BYTXMIRR)	.84***	.65 ***	.41
Math GPA in 9th & 10th gr (GPA910_m)	.44***	.09 ***	.04
Efficacy for graduating HS (grad_eff)	.26***	.04 ***	.02
Plans to take SAT (SATplan)	.38***	.07 ***	.03
Math self-concept (m_selfcpt)	.40***	.10 ***	.05
Demographic factors			
Female	-.05***	-.13 ***	.06
Black	-.23***	-.08 ***	.04
Hispanic	-.14***	-.02 *	.01
Asian	.08***	.03 *	.01
Native American	-.06***	-.02 *	.01

*$p < .05$, **$p < .01$, ***$p < .001$; N=8976 for replication, while 8969 for Byrnes & Miller. Byrnes & Miller did not report semi-partial correlations. Significance testing is not reported for semi-partial correlations. Only two predictors with squared semi-partial correlation greater than 1% was Geometry 1 year (1%) and math achievement before 9th grade (17%).

4.2.2 Results from rule induction using study variables

Following the same format as Study 1, I first present predictive accuracies across approaches, followed by model predictors and their importance, interesting rules and their accuracies, and results from association rule mining.

4.2.2.1 Predictive accuracies of ruleset induction

The confusion matrices and associated predictive accuracy measures for ruleset induction for Study 2, using 29 and 1933 possible predictors are presented in Table 36 and Table 37, respectively. The relative performance of each algorithm in terms of F-measure and Kappa statistic is illustrated in Figure 10. The overall accuracy for multiple linear regression using only the variables Byrnes and Miller used in their study was 86%, with 88% of the high achieving group correctly classified (recall), and 85% of those who were classified as high achieving being correctly classified (precision). The F-measure, or the harmonic mean between precision and recall, was .860. The Kappa statistic was .725, indicating that if predicting as well as random chance were 0 and making a perfect prediction were 100, logistic regression performed at about 73.

With just 29 possible variables to choose from, the rule induction classifiers performed comparably. One exception was CBA, which performed *much* worse than regression under all measures examined, with an overall accuracy of 72%. Aside from that, the rule induction classifiers' overall accuracy ranged from 84% to 87%, F-measure ranged from .838 to .872, and the Kappa statistic between .685 and .744. Notably, the two that performed better than regression were ensemble classifiers (C5.0 and Random Forest). An additional 1904 predictors from which

to build a model did not change the predictive accuracies of the rule induction classifiers by very much.

Table 36. Confusion matrices for ruleset mining (Study 2, 29 possible predictors)

		Prediction on test set				
		High achieving	Not high achieving	% Correct	F-measure	Kappa
CBA	High achieving	130110	1955	98.5%	.781	.446
	Not high achieving	71097	60844	46.1%		
	% Correct	64.7%	96.9%	72.3%		
RIPPER	High achieving	111508	20557	84.4%	.853	.708
	Not high achieving	17980	113961	86.3%		
	% Correct	86.1%	84.7%	85.4%		
PART	High achieving	110900	21165	84.0%	.850	.702
	Not high achieving	18117	113824	86.3%		
	% Correct	86.0%	84.3%	85.1%		
CART	High achieving	107004	15760	87.2%	.838	.688
	Not high achieving	25490	115752	82.0%		
	% Correct	80.8%	88.0%	84.4%		
C5.0	High achieving	109170	22895	82.7%	.849	.706
	Not high achieving	15927	116014	87.9%		
	% Correct	87.3%	83.5%	85.3%		
C4.5	High achieving	114463	17602	86.7%	.853	.700
	Not high achieving	21971	109970	83.3%		
	% Correct	83.9%	86.2%	85.0%		
QUEST	High achieving	112237	20257	84.7%	.844	.685
	Not high achieving	21268	110244	83.8%		
	% Correct	84.1%	84.4%	84.3%		
Bagging CART	High achieving	112340	19391	85.3%	.850	.700
	Not high achieving	20154	112121	84.8%		
	% Correct	84.8%	85.3%	85.0%		
Boosted C5.0	High achieving	115023	17042	87.1%	.872	.744
	Not high achieving	16654	115287	87.4%		
	% Correct	87.4%	87.1%	87.2%		
Random forest	High achieving	111339	14921	88.2%	.862	.730
	Not high achieving	20726	117020	85.0%		
	% Correct	84.3%	88.7%	86.5%		
Regression	High achieving	111746	15963	87.5%	.860	.725
	Not high achieving	20319	115978	85.1%		
	% Correct	84.6%	87.9%	86.3%		

Table 37. Confusion matrices for ruleset mining (Study 2, 1933 possible predictors)

		Prediction on test set				
		High achieving	Not high achieving	% Correct	F-measure	Kappa
CBA	High achieving					
	Not high achieving		*Results Not Attained*[1]			
	% Correct					
RIPPER	High achieving	11566	1662	87.4%	.852	.697
	Not high achieving	2348	10856	82.2%		
	% Correct	83.1%	86.7%	84.8%		
PART	High achieving	11517	1711	87.1%	.851	.695
	Not high achieving	2319	10885	82.4%		
	% Correct	83.2%	86.4%	84.7%		
CART	High achieving	12059	2699	81.7%	.851	.680
	Not high achieving	1515	10159	87.0%		
	% Correct	88.8%	81.1%	84.0%		
C5.0	High achieving	10877	2351	82.2%	.844	.697
	Not high achieving	1660	11544	87.4%		
	% Correct	86.7%	85.4%	84.8%		
C4.5	High achieving	11006	2222	83.2%	.846	.697
	Not high achieving	1784	11420	86.5%		
	% Correct	86.1%	83.7%	84.8%		
QUEST	High achieving	11261	2069	84.5%	.847	.693
	Not high achieving	1987	11115	84.8%		
	% Correct	85.0%	84.3%	84.7%		
Bagging CART	High achieving					
	Not high achieving		*Results Not Attained*[1]			
	% Correct					
Boosted C5.0	High achieving	11613	1615	87.8%	.883	.767
	Not high achieving	1458	11746	89.0%		
	% Correct	88.8%	87.9%	88.3%		
Random forest	High achieving	11318	1453	88.6%	.871	.746
	Not high achieving	1904	11757	86.1%		
	% Correct	85.6%	89.0%	87.1%		

[1]Possibly due to the large number of predictors, CBA and Bagging CART did not run to convergence.

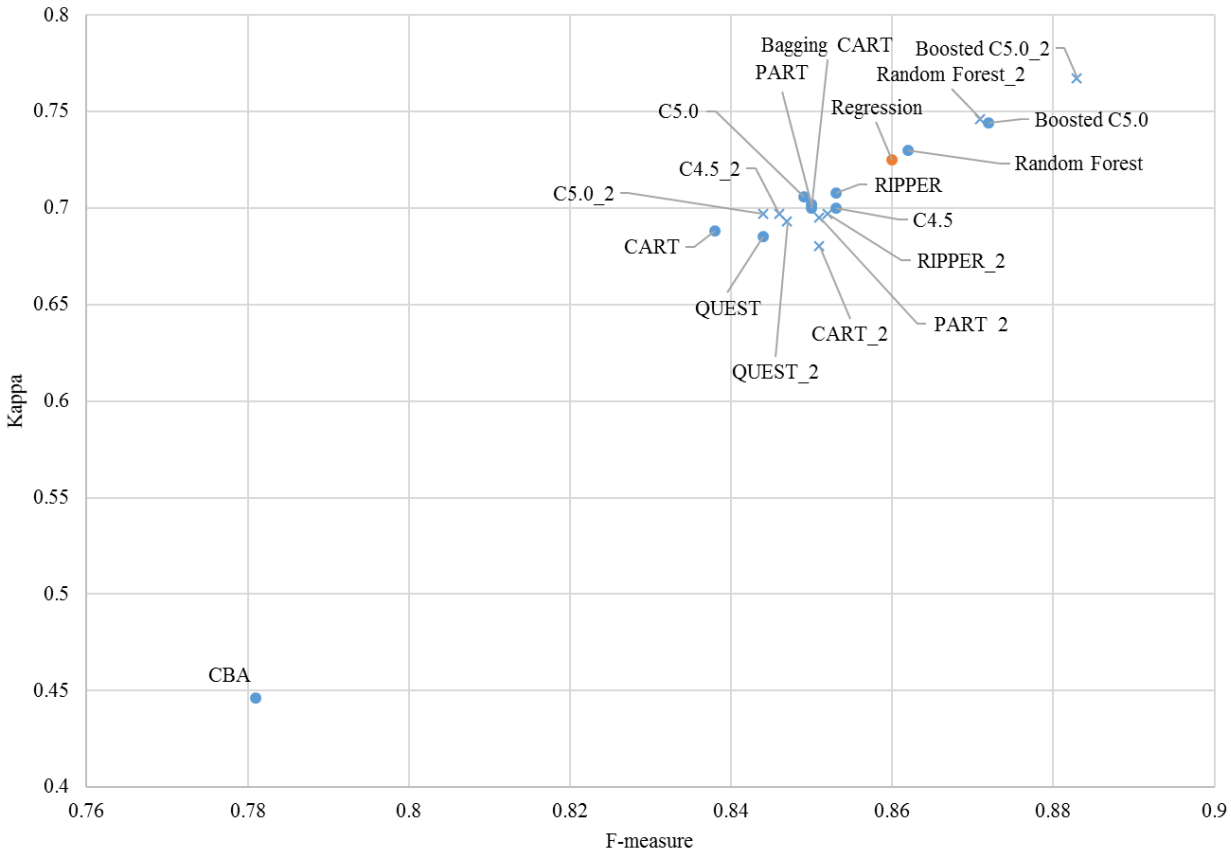


Figure 10. F-measure and Kappa statistics of logistic regression vs rule induction (Study 2)[1]

[1]The orange marker represents result from regression, while the blue markers indicate results from rule induction approaches. Circle markers indicate results using Byrnes & Miller's (2007) 29 possible predictors, while "X"s indicate results using 1933 possible predictors.

4.2.2.2 Model predictors and their importance

Figure 11 and Figure 12 illustrate the relative importance of the predictors or predictor sets that were most frequently included across the ruleset induction models for which such information was available. As with Study 1, the criteria for "model importance" varied slightly across models, depending on the information that was available for each algorithm, and are indicated as notes attached to each figure. The figures indicate that generally, across algorithms, 8th grade math score was most important among 29 variables that Byrnes and Miller had considered in their final model

followed by geometry course-taking, and that this was the same even when 1933 variables were considered. The figures also indicate that many of the other predictors are considered important only by some approaches and not others.

Figure 13 and Figure 14 better illustrate how predictor or predictor set importance varies across rule induction approaches. Again, the heights of the predictors are proportional to their importance. When only 29 predictors were available, PART, RIPPER, CART, C4.5 and C5.0 relied strongly on 8th grade math scores to predict 12th grade math scores. QUEST relied on more on math course-taking, although it also considered 8th grade math scores. The two ensemble approaches, boosted C5.0 and Random Forest, considered many other predictors aside from 8th grade math scores and math course-taking, although math scores and course-taking were still prominent for Random Forest.

Results did not appear to change very much even when an additional 1904 predictors were made available for rule induction. All algorithms relied heavily on 8th grade math scores for predicting the outcome, except for boosted C5.0, which relied equally on a very large number of variables, and more on math course-taking than others. Geometry, 8th grade GPA, and math self-concept were considered important by several algorithms.

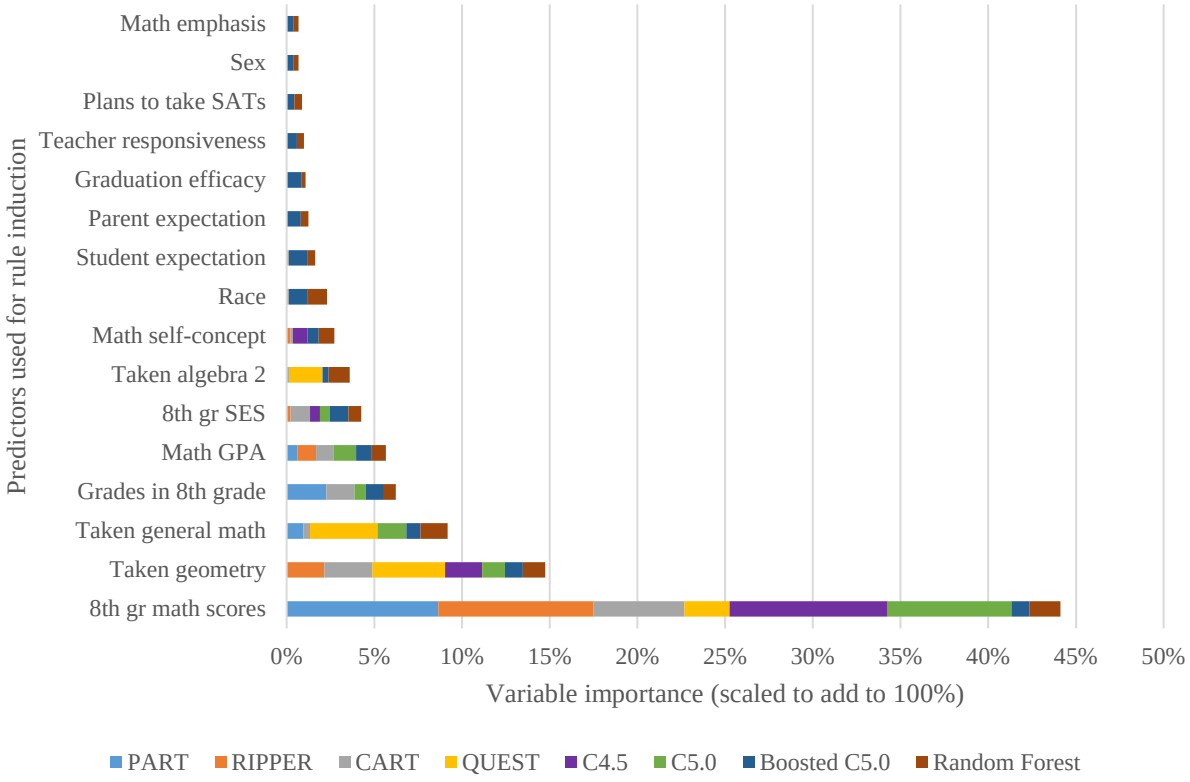


Figure 11. Predictor importance by algorithm (Study 2, 29 possible predictors)

Note: Predictor importance for CART and Random Forest were calculated as the extent to which the variable reduced the Gini index or mean squared error, respectively, as these were automatically generated. For the rest of the algorithms where that measure was not available, attribute usage was used. Attribute usage was calculated as the number of participants that the variable sorted. The predictor importance for each algorithm was scaled to total 100. Subcategories of math course-taking and race were combined.

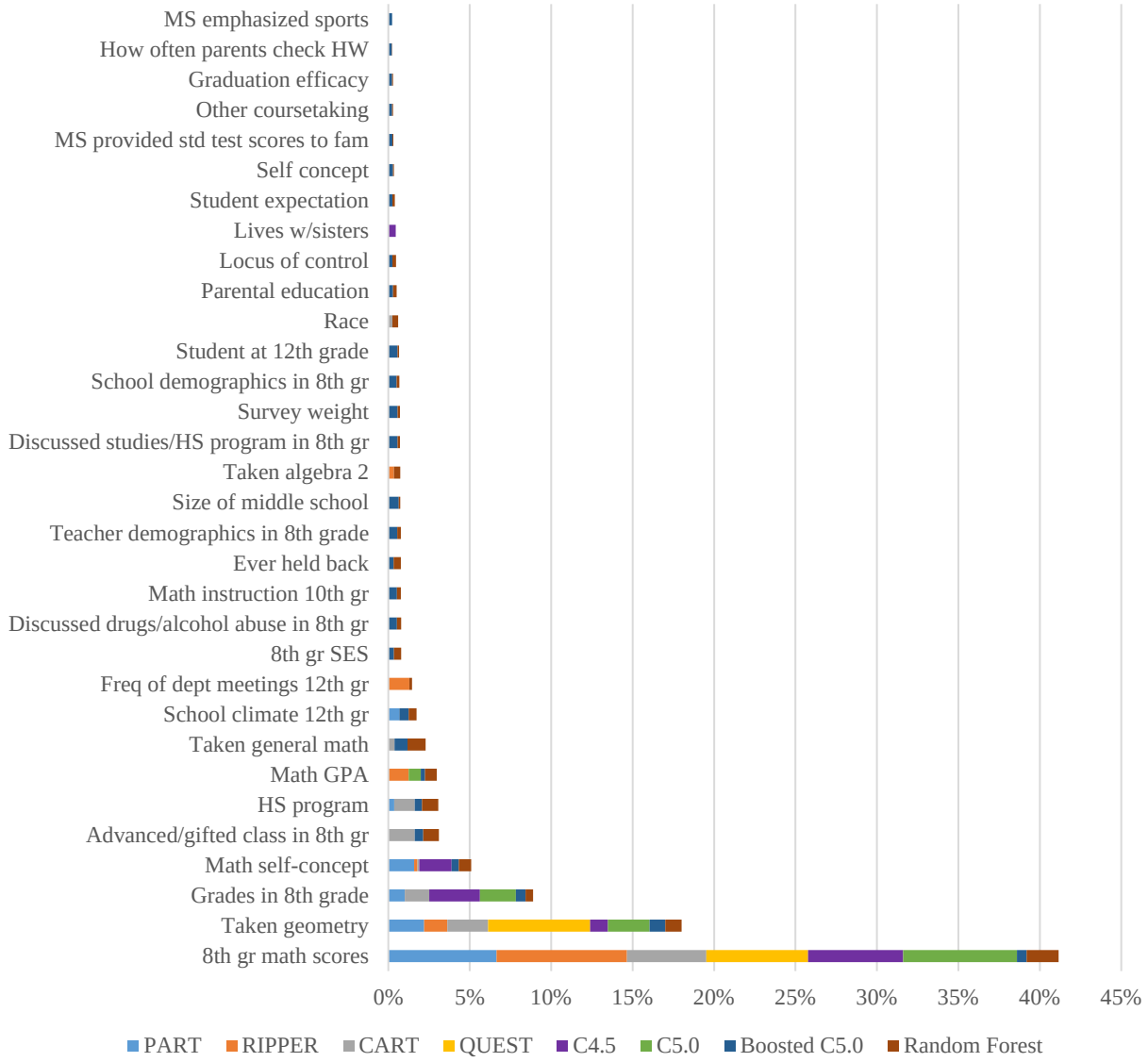


Figure 12. Predictor importance by algorithm (Study 2, 1933 possible predictors)

Note: Predictor importance for CART and Random Forest were calculated as the extent to which the variable reduced the Gini index or mean squared error, respectively, as these were automatically generated. For the rest of the algorithms where that measure was not available, attribute usage was used. Attribute usage was calculated as the number of participants that the variable sorted, where participants were double-counted as being sorted by that variable if, in tree-induction, the variable was used twice to sort the same person and if there was at least one other variable that sorted that person between the first and second sorting. The predictor importance for each algorithm was scaled to total 100, predictors that contributed less than 1 in all of the non-ensemble algorithms were excluded from consideration, and remaining predictors were grouped into the categories in the figure. Refer to Appendix D, Table 70 for the groupings.

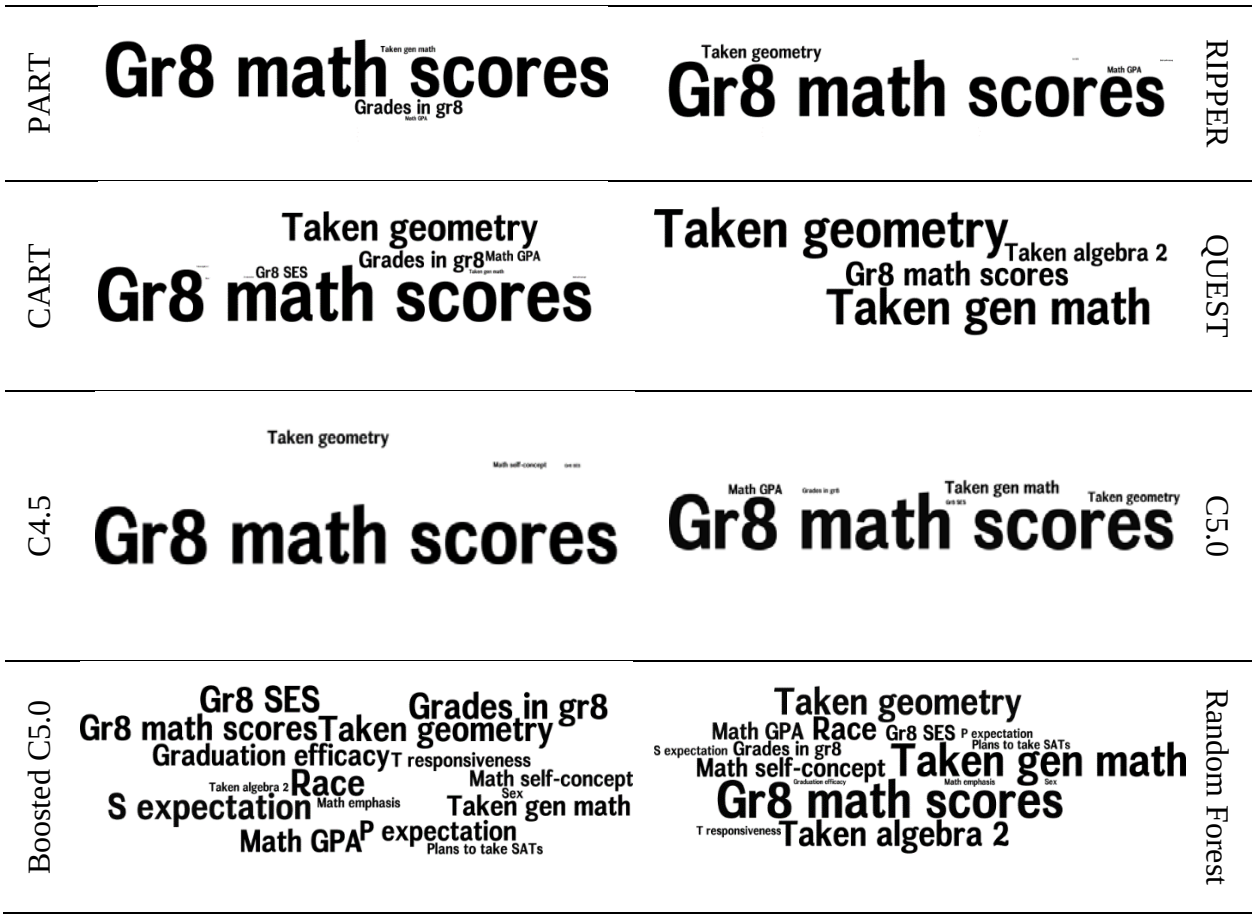


Figure 13. Predictors included in model, sized proportionally to importance (Study 2, 29 possible predictors)

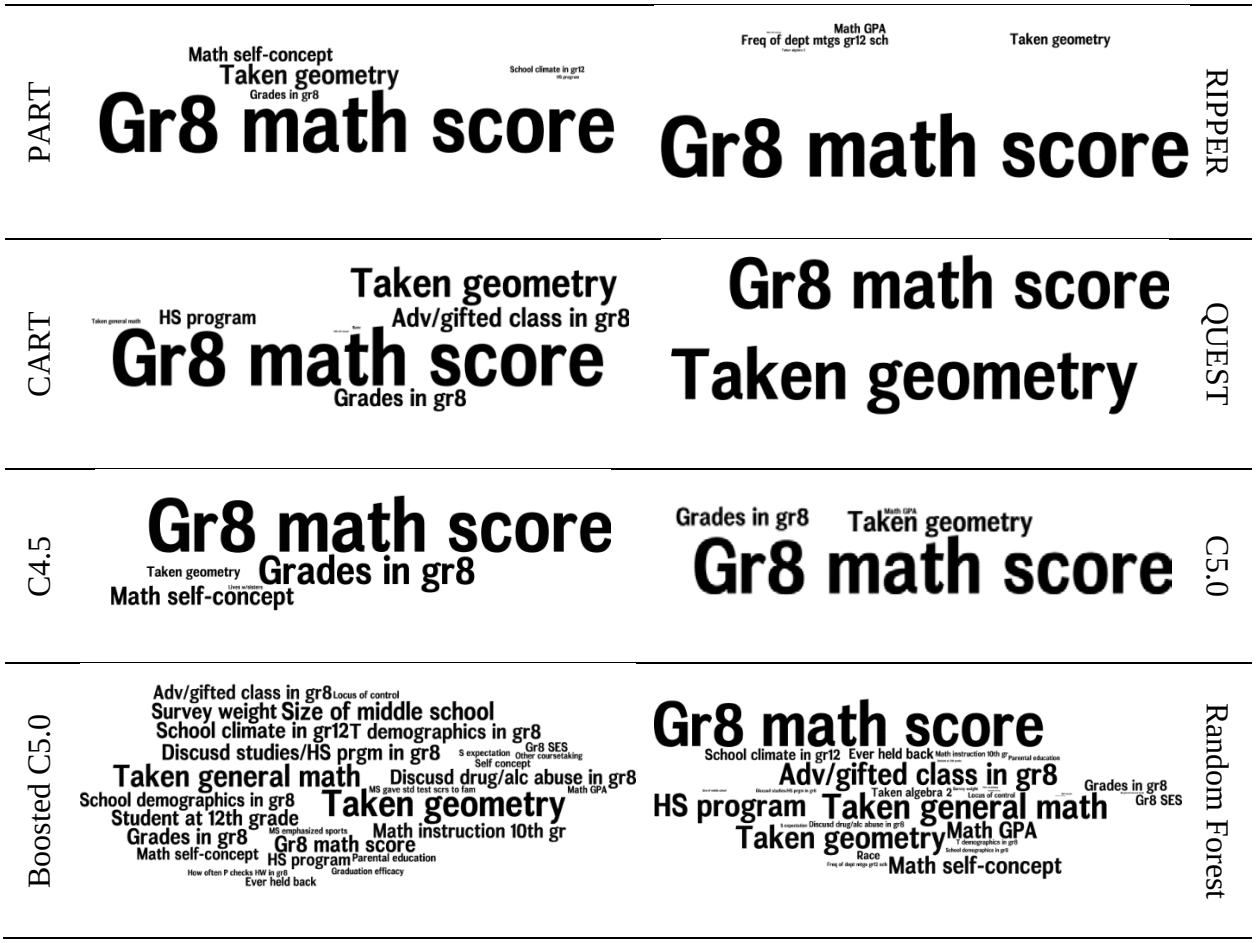


Figure 14. Predictors included in model, sized proportionally to importance (Study 2, 1933 possible predictors)

4.2.2.3 Interesting rules within rulesets and their accuracy

The rulesets and trees that were mined for each algorithm, including their confidence (proportion the rule was correct, given the antecedent applied) and coverage (proportion of the sample to which the rule applied) are presented in Appendix D. For most rules, the accuracy metrics in the training and test sets were identical. This—I realized upon reflection—was likely due to the large sample size, and made me question whether I should have created the training and test sets before expanding the dataset using case weights. Another curious observation was that while I set each algorithm to detect rules that account for at least 2 percent of the sample, RIPPER and C50

produced 1 or 2 rules that applied to less of the sample. These rules are listed in Appendix D, but not considered in the analysis.

The number of rules is indicated in Table 38. CBA generated too many (832) rules for me to examine individually, and anyway had a considerably lower predictive validity than the other algorithms, so I did not analyze these rules further. The rest of the algorithms produced between 4 and 20 rules.

Table 38. Number of rules discovered (Study 2)

	Variables considered	Rules
CBA	19	832
RIPPER	19	6
	1372	6
PART	19	6
	1372	10
C4.5	19	7
	1372	12
CART	19	20
	1372	15
C5.0	19	9
	1372	9
QUEST	19	10
	1372	4
Total		946

Because 8th grade math score was a prominent predictor across most algorithms, I created a series of mosaic plots to represent the rulesets, placing the math score on the y-axis for each. Because the math predictor was represented the same way across all algorithms, the partitioning diagrams allowed me to better identify similarity and differences in the classifications across them.

The entire block represents the students. The segmentation of the block represents subgroups of students that were identified by the algorithm. The attribute-values that created the

subgroups are labeled within the blocks and/or by the y-axis, and the outcome (a numeric value for CART, and "HIGH"/"LOW" for others). The area of the partitions roughly corresponds to the proportion of students to whom the rule applied, or coverage. The coverage is also indicated numerically within each subgroup. For example, in the first diagram (Figure 15), the proportion of students who scored 18 or below in 8th grade math is described both numerically ("31%"), and visually be having the box for that rule cover approximately 31% of the block's area. However, it should be noted that when vertical splits are made to the block (i.e., splits parallel to the y-axis), the relative difference in areas across higher and lower 8th grade math scores are not necessarily accurately represented.

For results where the outcome was dichotomous, I represented predictions of high achievement in green, and low achievement in gold. The confidence values are indicated within each block and is also represented by the darkness; higher confidence is represented by a darker color.

For CART results, where the outcome was continuous, the figures are black-and-white, with the darkness corresponding to the magnitude of the outcome. The higher the predicted score, the darker the color. The accuracy of each CART rule was estimated by the proportion of students who were within 5 and 10 percentage points of the estimate. Those are not indicated in the mosaic diagrams, and only indicted in Appendix D (Table 64 and Table 65).

The partitioning diagrams are presented in Figure 15 through Figure 26. There were clear and consistent patterns across the results. 8th grade math scores and (for the most part) geometry course-taking were positively correlated with the outcome. Math self-concept and general math course-taking were predictive of some subgroups but not others. More specifically, the results indicated that:

- 8th grade math score is very predictive of 12th grade math scores. If it is greater than 30, it is highly likely that the student will be high achieving (when the outcome is dichotomized) in math in 12th grade. If it is less than or equal to 18, then there is a 93% chance of being lower achieving.
- Within each band of 8th grade math score, course-taking is typically most predictive of 12th grade math scores. For example, students are more likely to be higher achieving if they have taken geometry by grade 10, and/or have not taken general math in high school by grade 10. However, course-taking is not particularly relevant to those who had the highest 8th grade math scores (>24, and certainly >30), likely because many of those students probably have taken geometry.
- For those with higher grade 8 math scores (>24), 12th grade math score tends to be positively associated with math self-concept (e.g., PART conducted with 1933 variables, CART). Math self-concept did not appear to strongly relate to the outcome among those whose 8th grade math scores were <=24, except for those who had 8th grade scores of at least 18 and had taken geometry as detected by C4.5.
- While those who have taken Algebra 2 and Geometry tend to score high, those who have taken Algebra 2 but not Geometry tend to score high only if they also had a low 8th grade math score (QUEST with 29 possible predictors).
- Even if students had not taken geometry, if they had high test scores in 8th grade, they tended to be higher achieving in 12th grade math. Conversely, even if students had taken geometry, if they had low math scores in 8th grade, they tended not to do well in 12th grade math.

- The CART models provided the most detail for predicting the outcome, understandably because the outcome was numeric rather than dichotomous. 8th grade math scores were most predictive, and what was next most predictive depended on this math score. For those who scored the lowest (below 24), geometry and general math course-taking were next most predictive. For those whose math scores were higher (24-30), geometry course-taking and math self-concept were next most predictive. For those who scored even higher in 8th grade math (30-35), math self-concept was next most predictive. For those who scored the highest in 8^{th} grade math (above 35), grade 8 GPA was next most predictive according to one model (and nothing added to the prediction according the other model).

The trends above may have implications for policy. For example, the finding that math self-concept is related to the outcome for only a subset of students suggest that math achievement cannot be improved by targeting self-concept by itself. Improving math achievement in middle school, and ensuring students take geometry and Algebra 2 in high school (and are sufficiently prepared for it), seems most important as a focus.

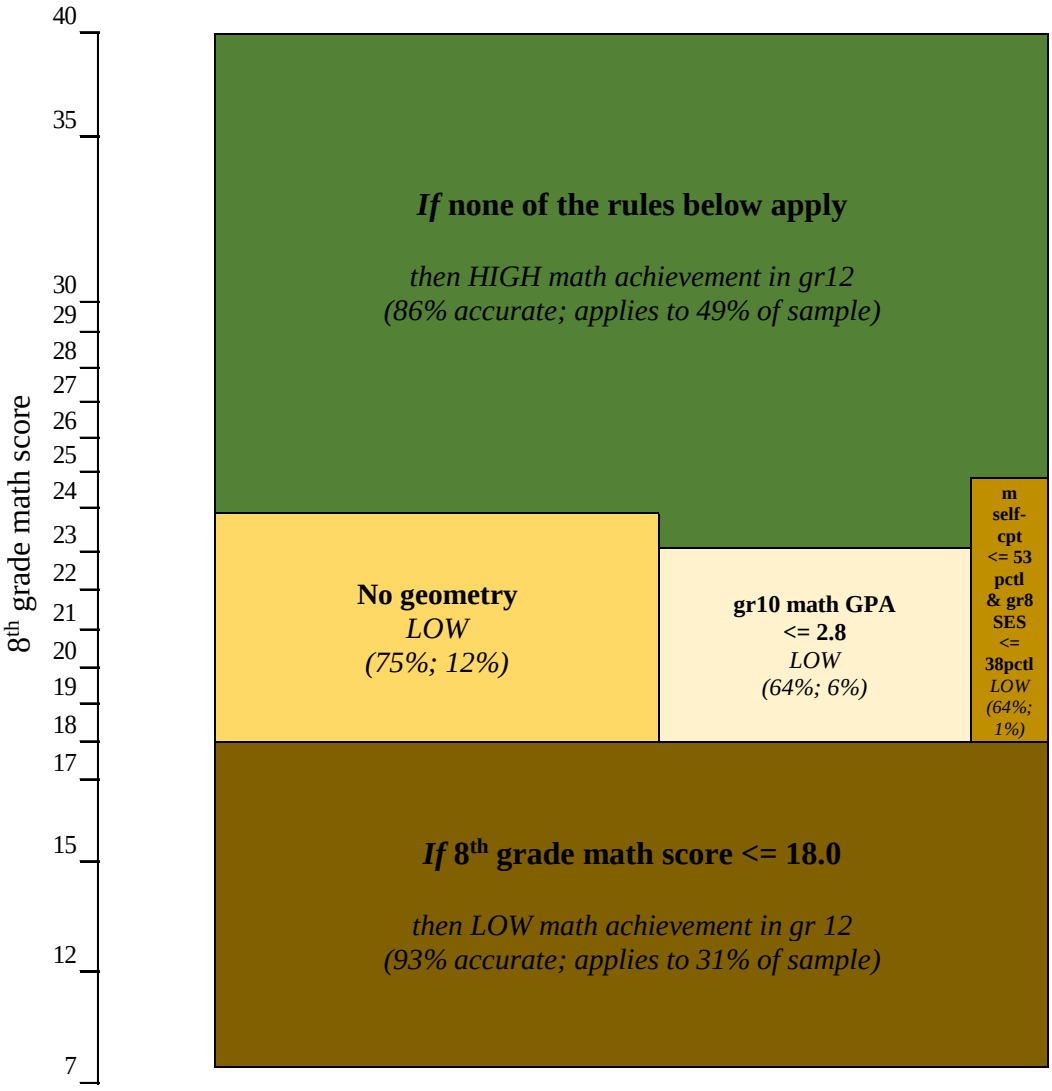


Figure 15. Mosaic plot for RIPPER (Study 2, 29 possible predictors)

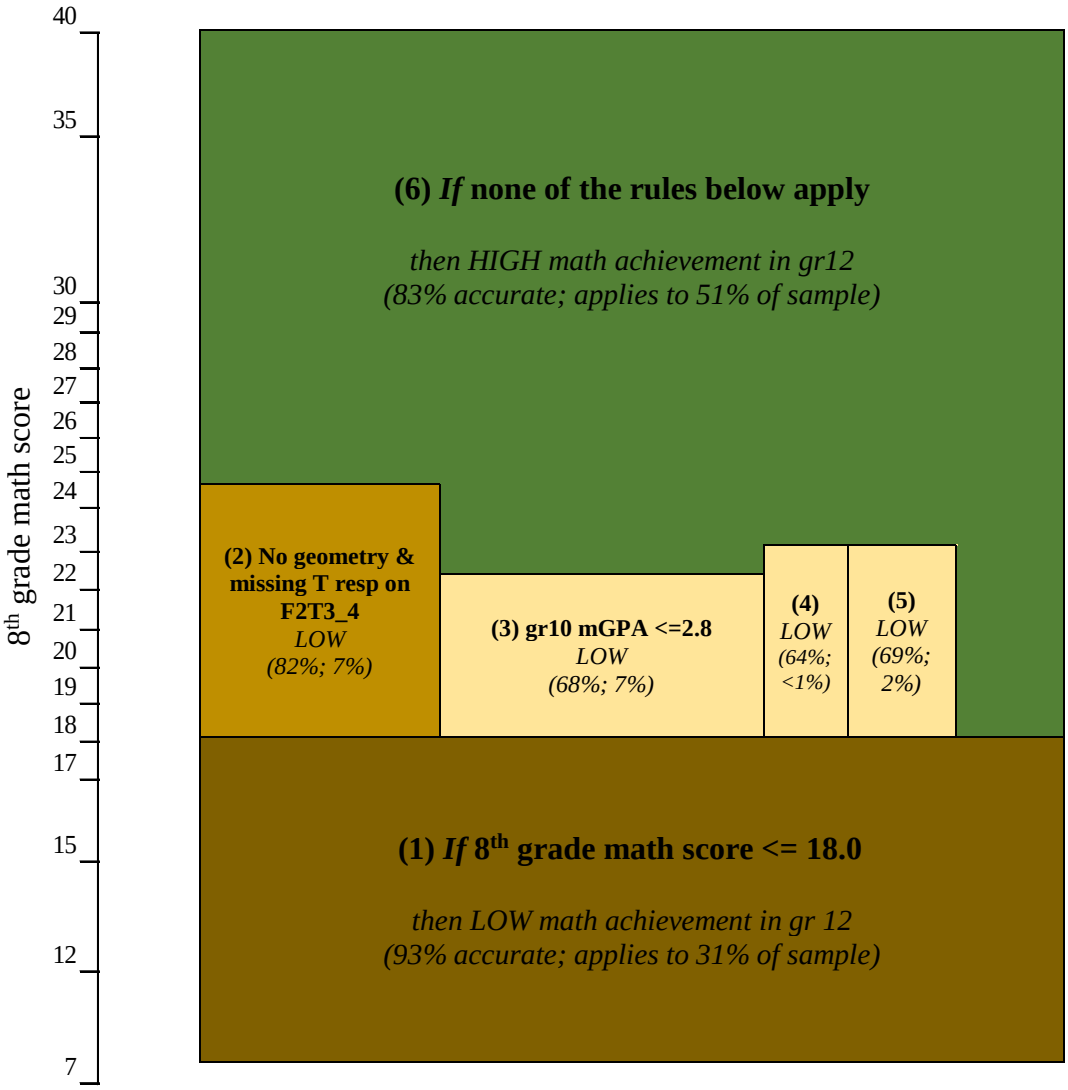


Figure 16. Mosaic plot for RIPPER (Study 2, 1933 possible predictors)

Note: F2T3_4 = frequency of departmental meetings; (4) math self-concept is <= 53rd percentile and teacher response is missing on frequency of departmental meetings; (5) Have taken neither geometry nor algebra 2.

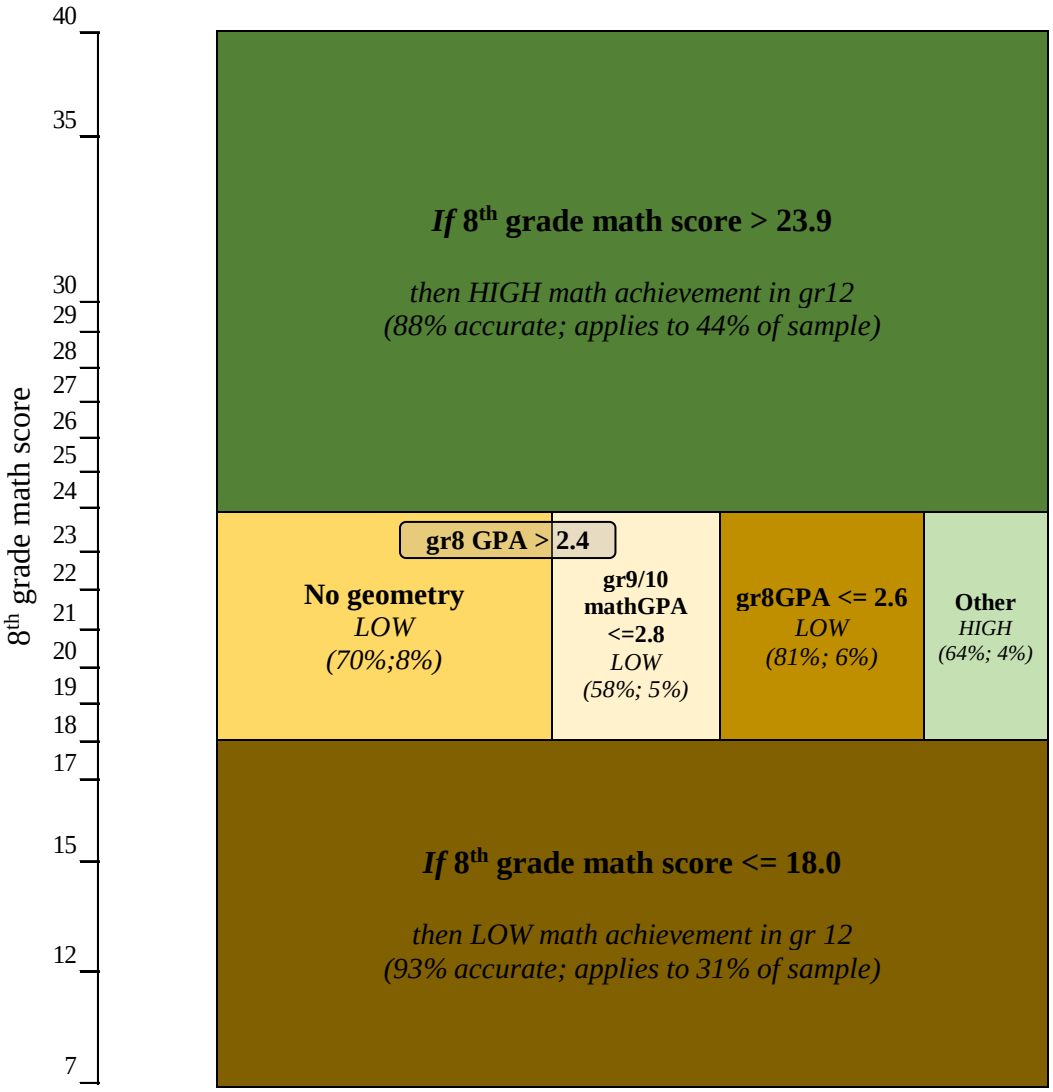


Figure 17. Mosaic plot for PART (Study 2, 29 possible predictors)

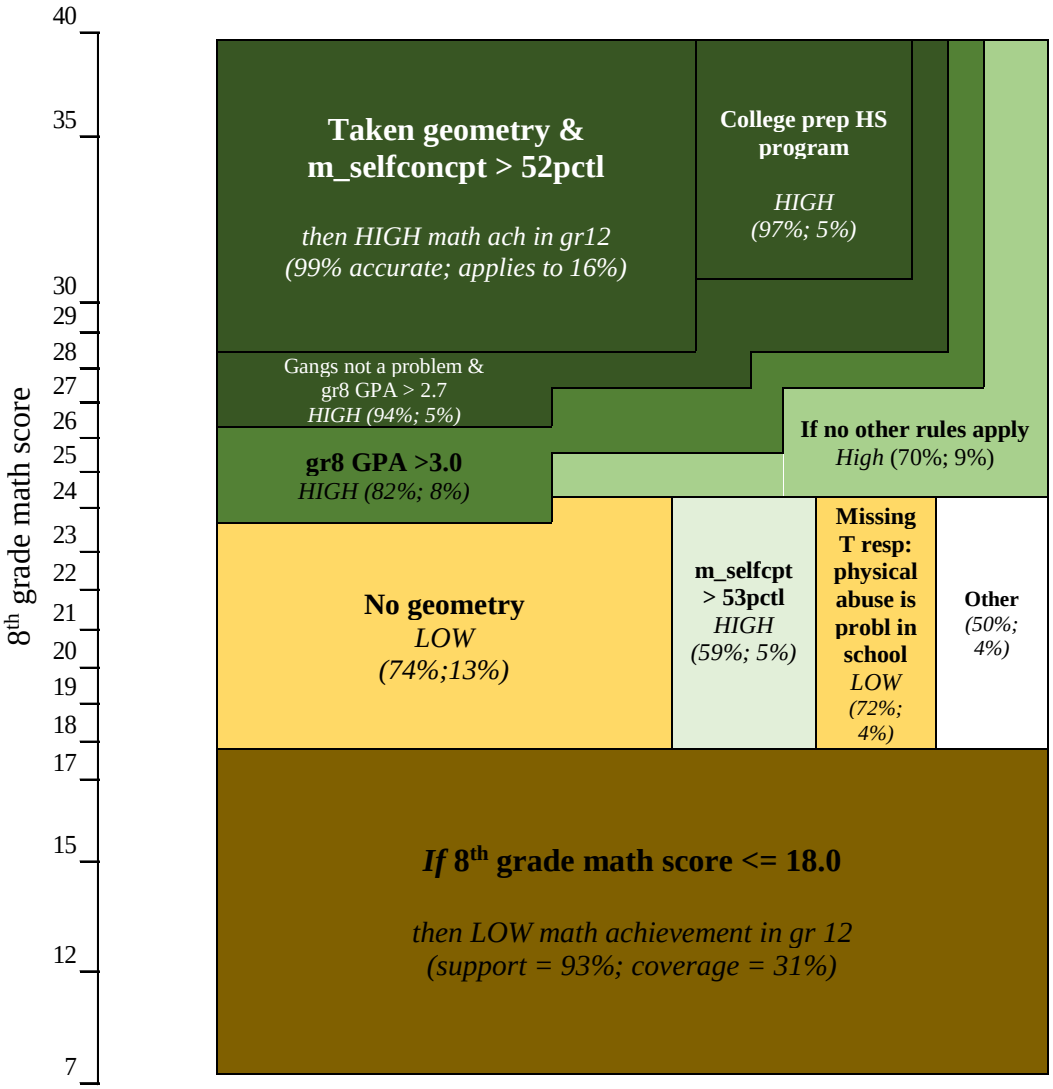


Figure 18. Mosaic plot for PART (Study 2, 1933 possible predictors)

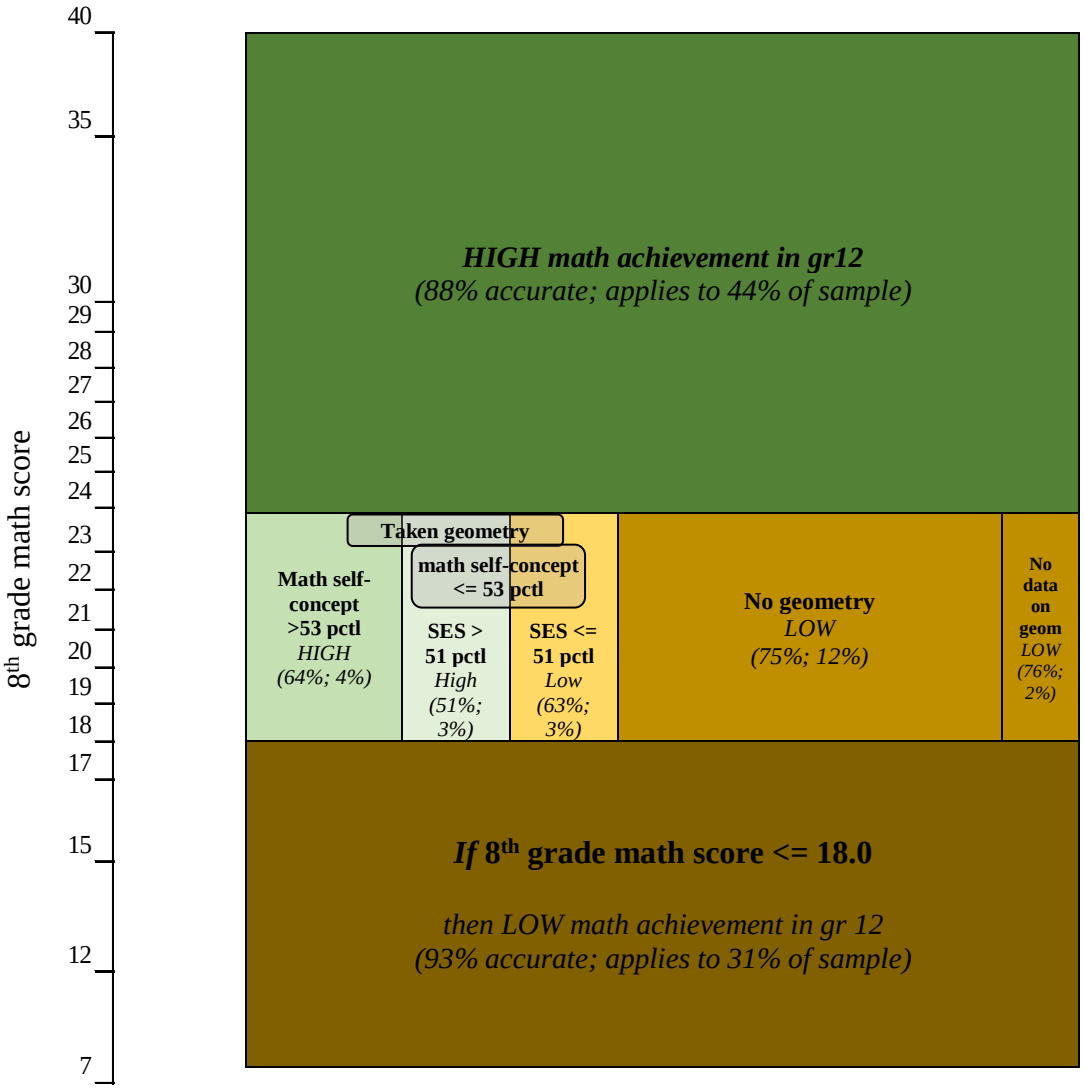


Figure 19. Mosaic plot for C4.5 (Study 2, 29 possible predictors)

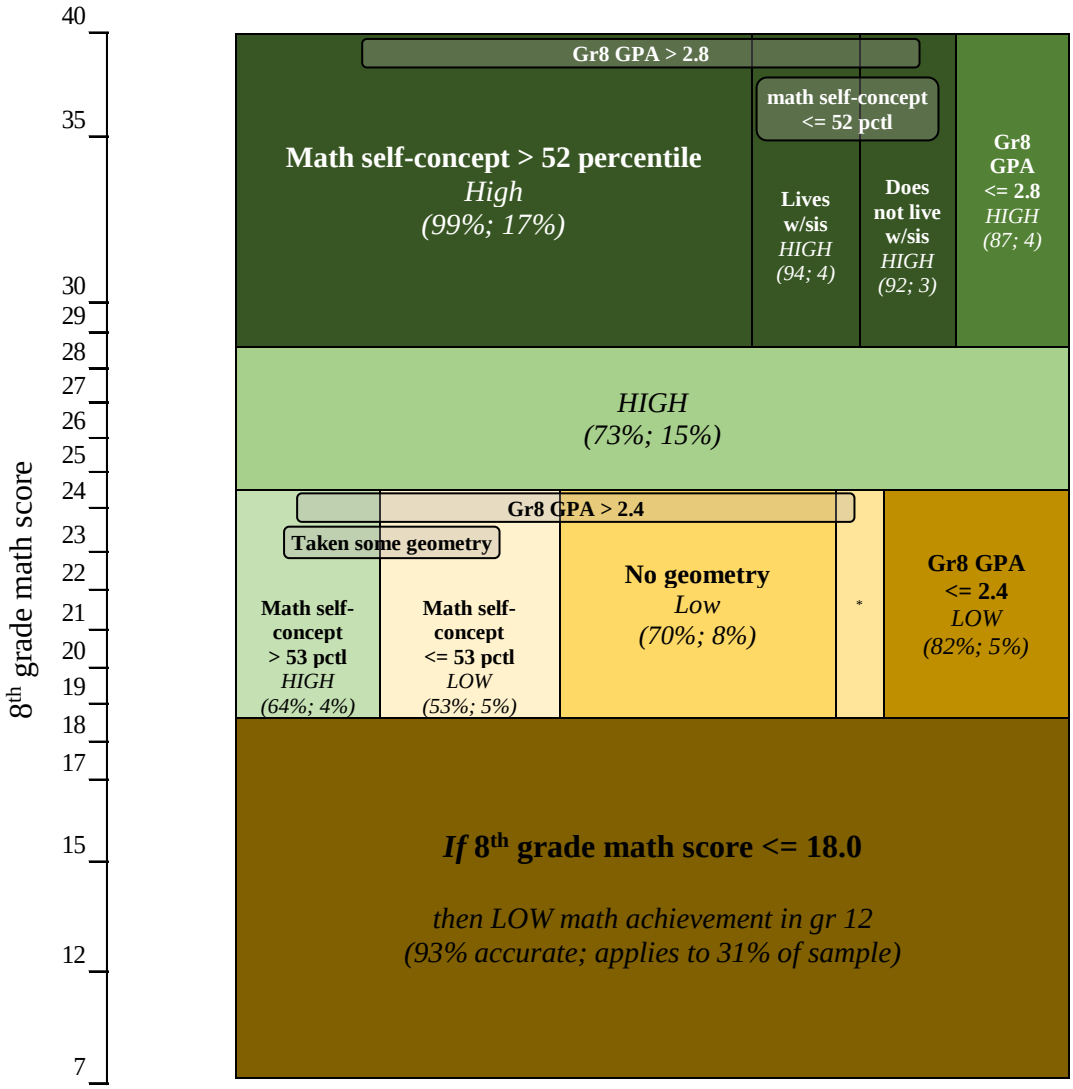


Figure 20. Mosaic plot for C4.5 (Study 2, 1933 possible predictors)

Note: A rule with <.001 coverage is excluded.

* Information about geometry coursework is missing. 68% confidence, 1% coverage.

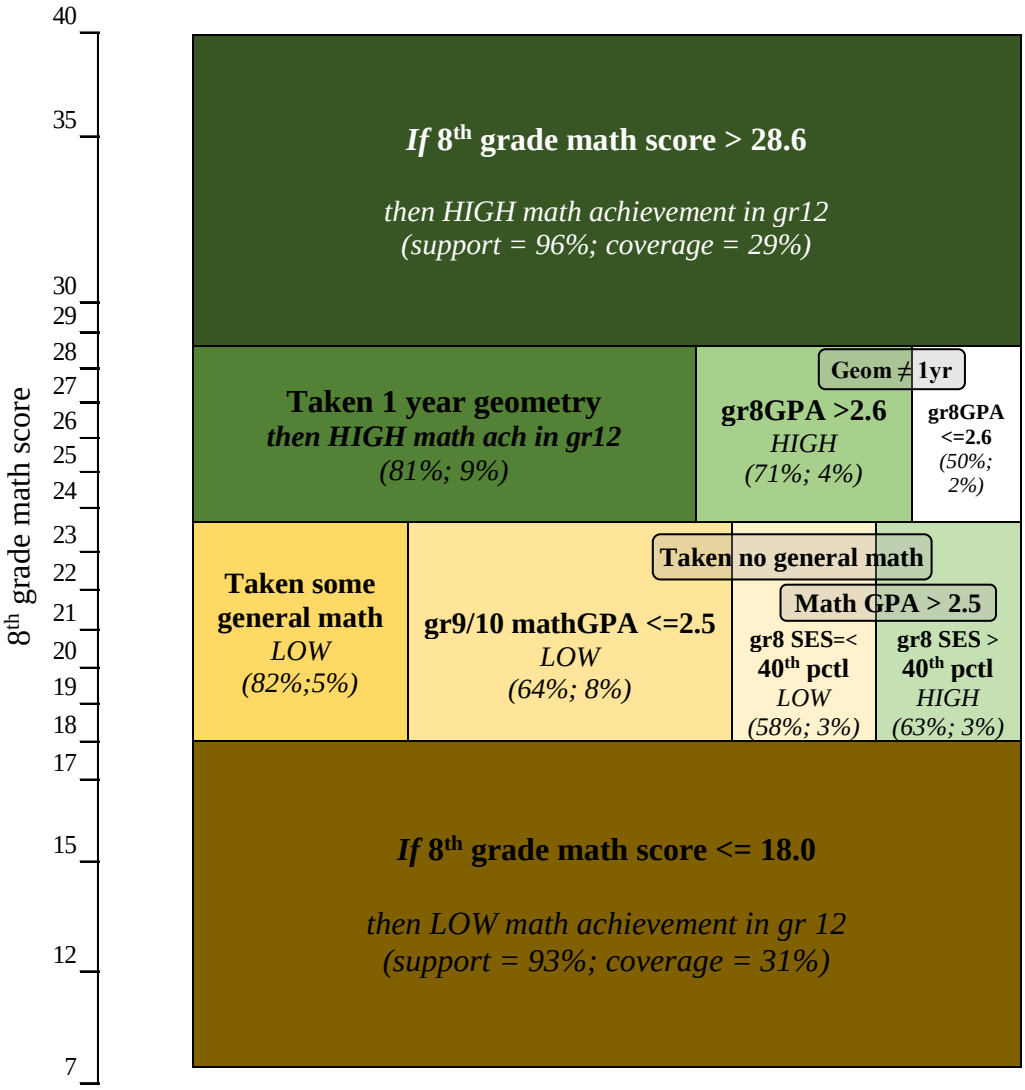


Figure 21. Mosaic plot for C5.0 (Study 2, 29 possible predictors)

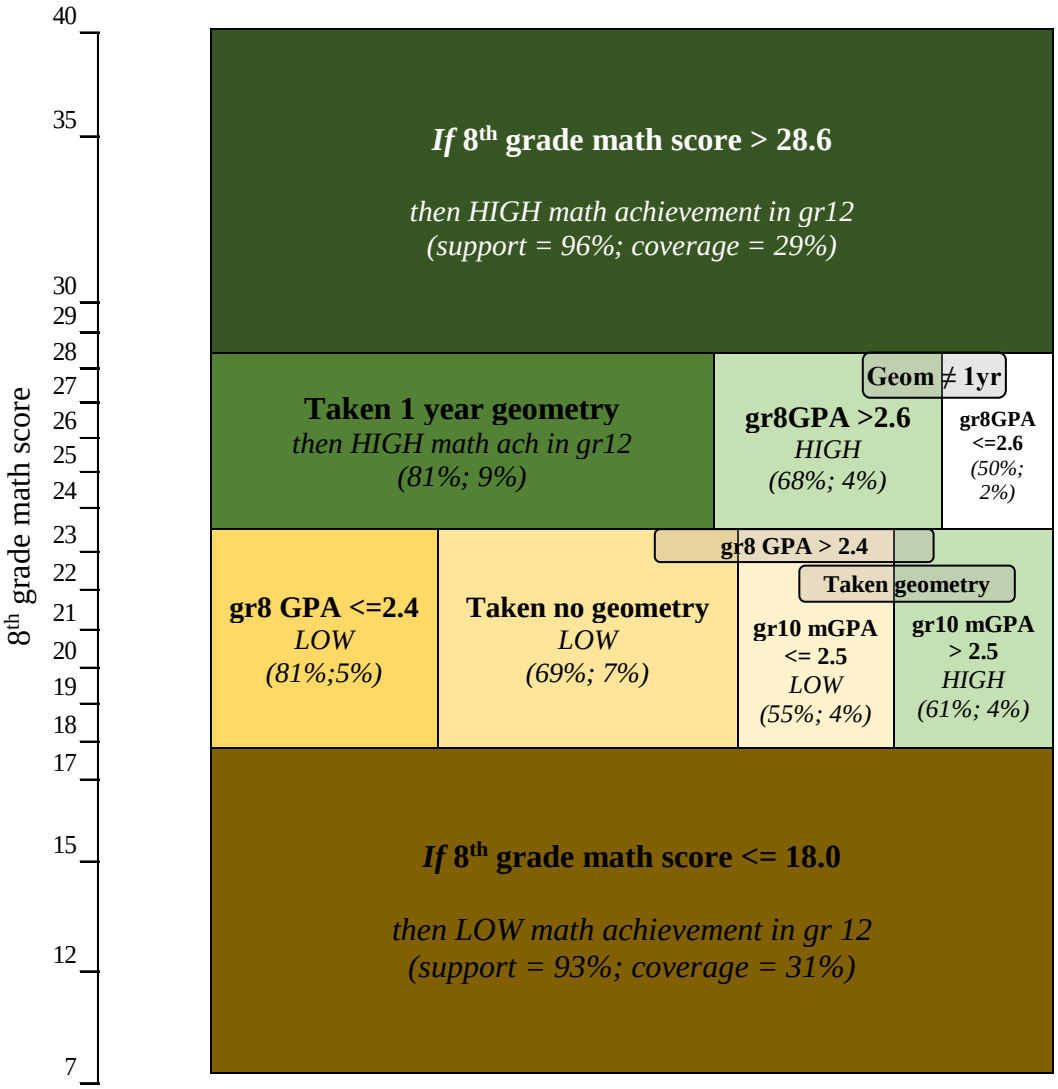


Figure 22. Mosaic plot for C5.0 (Study 2, 1933 possible predictors)

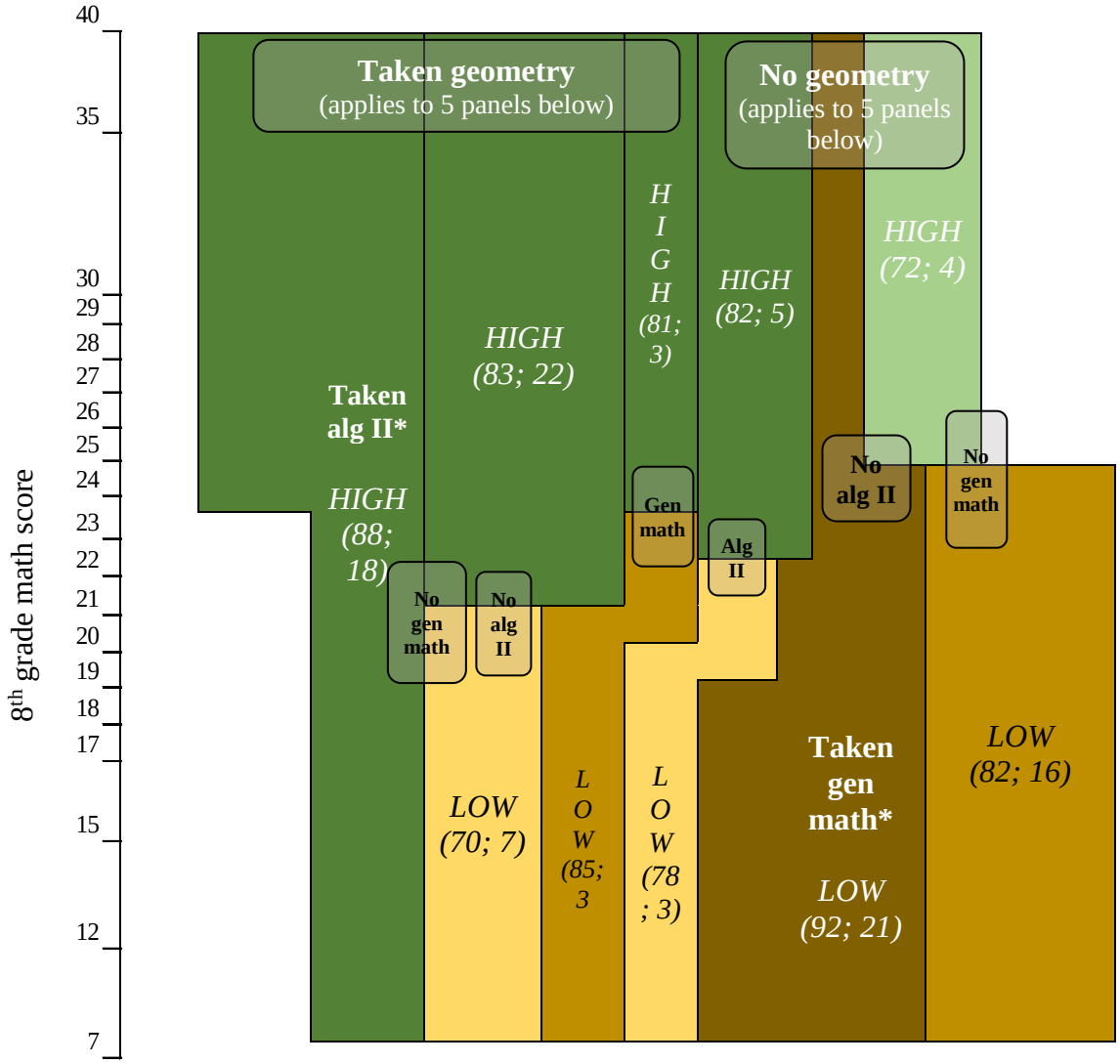


Figure 23. Mosaic plot for QUEST (Study 2, 29 possible predictors)

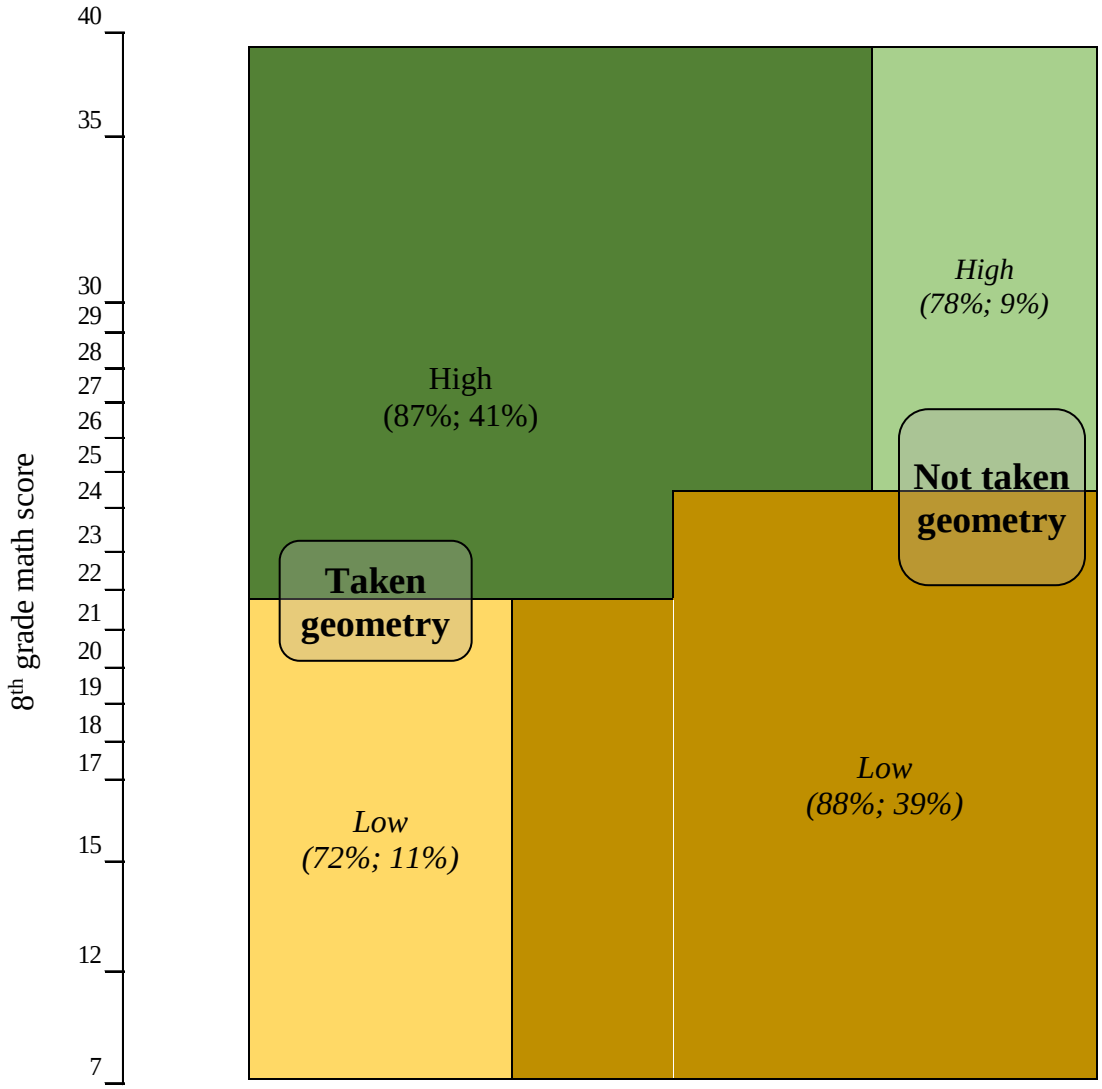


Figure 24. Mosaic plot for QUEST (Study 2, 1933 possible predictors)

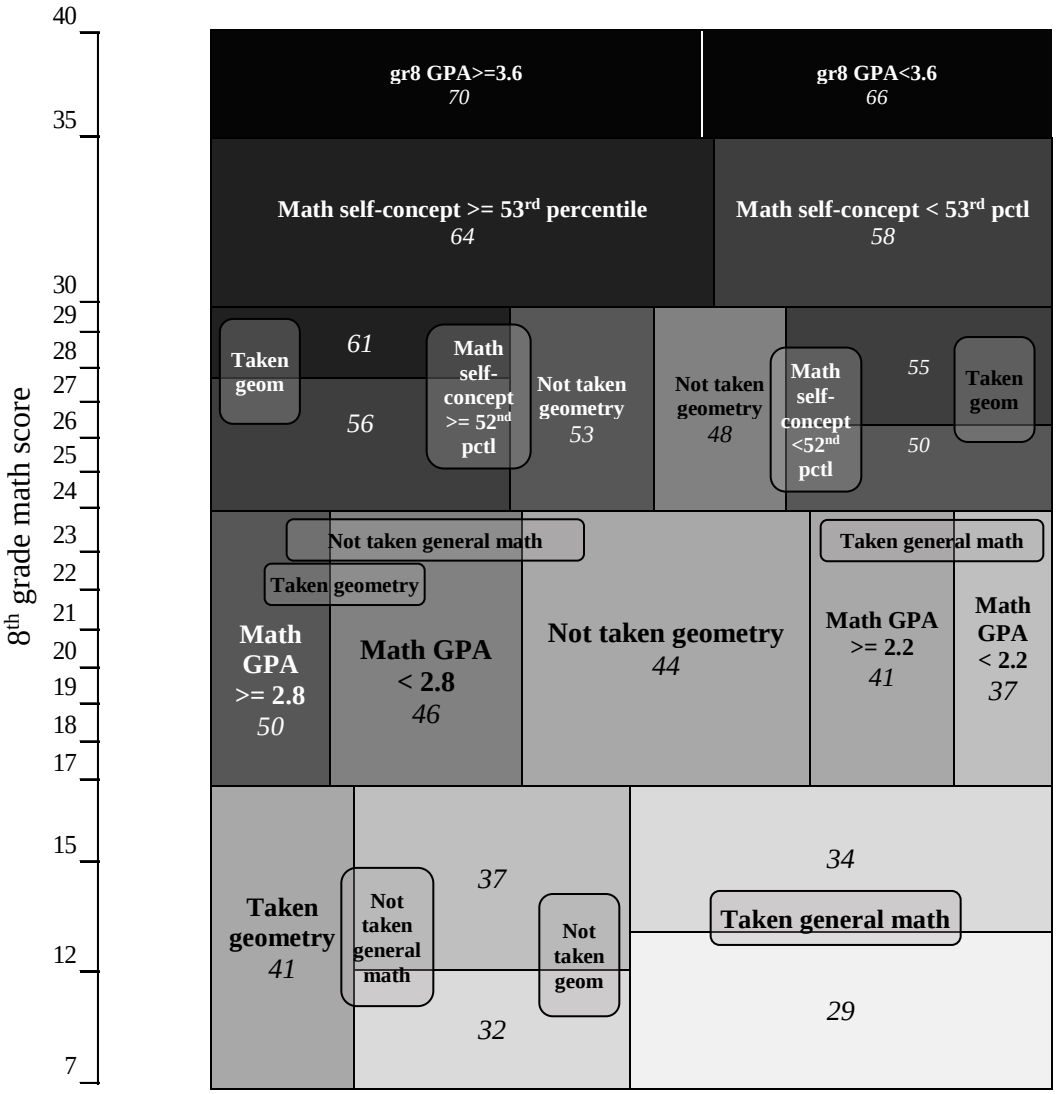


Figure 25. Mosaic plot for CART (Study 2, 29 possible predictors)

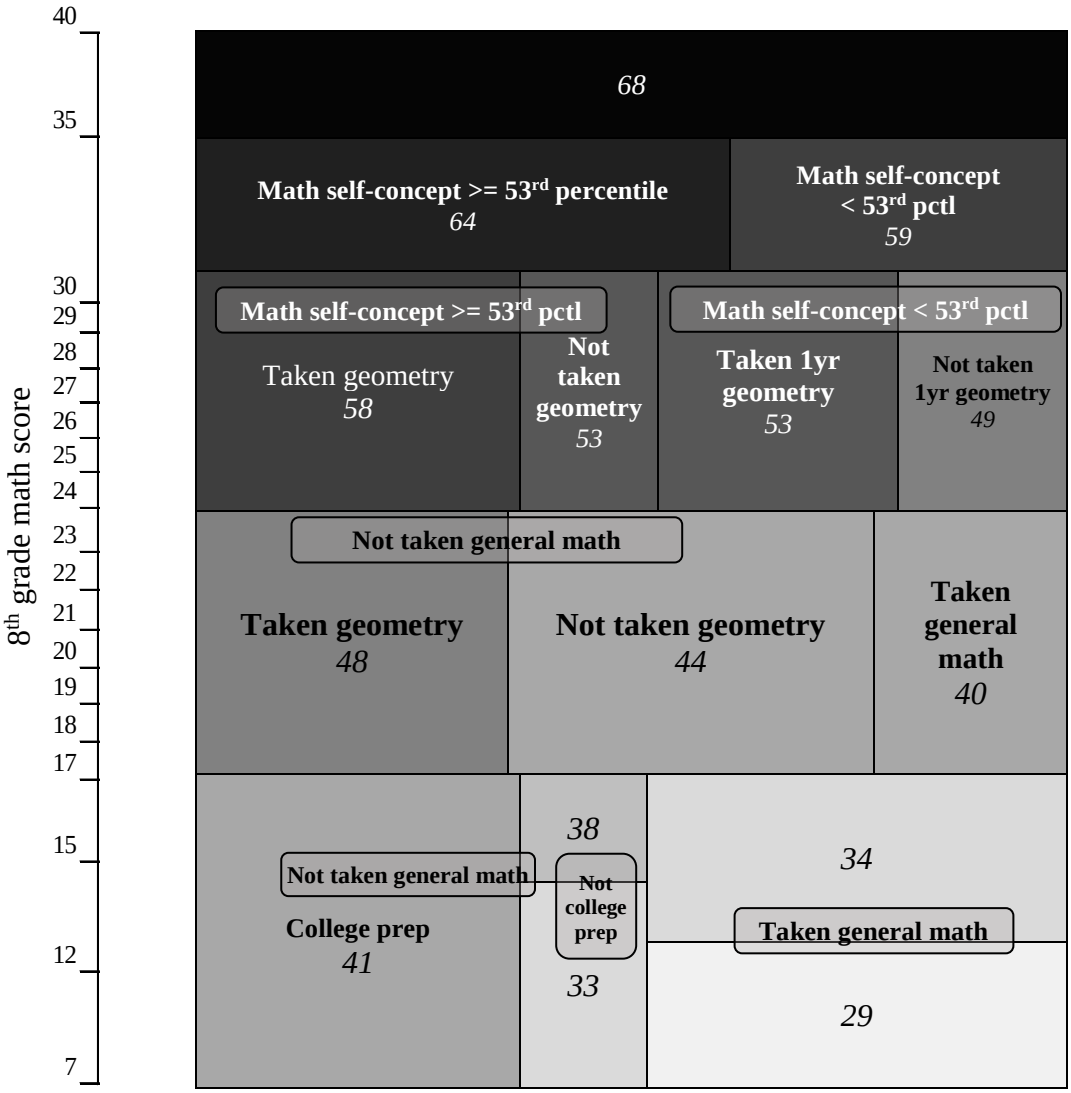


Figure 26. Mosaic plot for CART (Study 2, 1933 possible predictors)

4.2.2.4 Results from association rule mining

The number of rules generated by association rule mining for each of the subgroups is described in Table 39. For each subgroup, between 800 thousand to 1.1 million rules of lengths 2 to 3 described at least 25% of the high achievers in the generation set, and of those, approximately 700 thousand rules described at least 25% of the high achievers in the screening set. Among those, fewer than 100 had a positive likelihood ratio of 2.5 or greater on the training set, and between 2 and 26 met that cut-off in the test set. Six to fourteen rules of length 2 had positive likelihood ratios that were greater than 1.5 in the test set.

Table 39. Number of association rules generated by subgroup (Study 2)

8^{th} grade math score	# rules generated	# rules after screening	# rules of len=3 & PLR >= 2.5 (training/test)	# rules of len=2 & PLR >= 1.5 (training/test)
24-30	1,126,239 (2226 len= 2)	739,075 (1882 len= 2)	50/26	23/6
17-23	801,591 (2037 len= 2)	667,152 (1838 len= 2)	92/4	46/14
0-16	948,606 (2127 len= 2)	662,082 (1857 len= 2)	33/2	30/13

Note. True positive ratio was set to be >= .25 for all groups. Only rules with length 2 or 3 were generated. In the last two columns, the first number represents the number of rules in the training set, while the second represents that in the test set.

Table 40 summarizes the association rules with length 2 that were found among the three subsamples of 8^{th} grade math achievement (individual rules are described in Appendix F, Table 76). Regardless of students' 8^{th} grade math scores, math course-taking and high school's emphasis in academics were associated with 12^{th} grade math achievement. However, as we saw through the ruleset mining, whether students took Algebra 2 mattered more to those whose 8^{th} grade scores were higher (17-30), while geometry was more strongly associated with future achievement for

those with lower 8th grade scores (16 or less). Between 30-50% of high achievers had taken these math courses, while the percentage was about half for the lower achievers. Similarly, there were also differences across the three groups on the specific elements associated with students' math achievement pertaining to "high school's emphasis in academics."

Other factors were associated with achievement among one or two subgroups. Parent's expectations for the 8th grader to take algebra, attend college, and/or qualify for college financial aid were associated with math achievement only for students whose 8th grade test scores were very low. School climate, safety, and teacher perception of their professional environment were associated with achievement only for students whose 8th grade scores were between 17 and 23. Among students whose 8th grade math scores were higher (24-31), the potential difference-makers for achievement were more specific—e.g., enrollment in a foreign language class while in 8th grade, high school science class assigning lab reports once a week, and at least one parent being relatively old when the student was born (34-38 years old). Some factors, such as parental expectation, math course-taking and school climate, seemed more reliably related to the outcome because more than one variable suggested the relationship, the relationship pertained to more than 1 subgroup, and/or because the relationship seemed sound theoretically. Relationships that did not meet these criteria—e.g., relationship between availability of cheerleading and math achievement, among students who scored 24-31 in the 8th grade math assessment—seemed indefensible without further supportive evidence.

Table 41 lists all the length 3 rules predicting math achievement that had a positive likelihood ratio of at least 2.5. For students whose 8th grade math scores were between 24 and 31, all of the length 3 rules included Algebra course-taking—high achievers took at least some Algebra 2 in high school. The second condition generally had to do with school factors such as safety,

worthwhileness of homework, availability of student clubs, and use of textbooks and hall passes. Interestingly, student and family demographics, peer and community variables were not included in any rule antecedents. The rule antecedents characterized between 31 and 39 percent of the high achievers, applying 2.5-2.8 times more to them than to the lower achievers. If the rule antecedents applied, the student's chance of being a high achiever was about 1 in 4 to 1 in 5, while if they did not apply the chances were about 1 in 12.

Only 4 rules of length 3 described factors that differentiate high achievers among those whose 8th grade scores were between 17 and 23. Three included the condition that the high school teacher agrees they are encouraged to experiment with teaching, which by itself was true for 50% of the high achievers and 29% of others. When paired with another condition—either about 8th grader having visited with science/history museum or it being "very accurate" that the high school encourages students to take academic classes—the rule became true for about 30% of high achievers and just 12-15 % of lower achievers. If both conditions applied, the student had about a ¼ chance of being a high achiever, where as if that was not the case, the chances of being a high achiever were about 1 in 11. The fourth rule was about the 10th grader reporting a major emphasis in problem solving in math class, and the 12th grade teacher report about what was not discussed in their professional development that year ("in-depth study of a specialized subject" was not discussed). While the accuracy metrics resembled the other three rules, the second condition seems too specific, one-off and theoretically difficult to justify as a valid rule.

There were only two rules of length 3 that described factors that differentiate high achievers among those whose 8th grade scores were less than 17 with a positive likelihood ratio of over 2.5. Both rules included the condition of being male, which alone was not associated with higher achievement. One rule stated that being male and attending a school that places high emphasis on

academics is predictive of high achievement. The latter condition by itself applied to 28 percent of high achieving students and 17 percent of lower achieving students, and with the male condition added still applied to 28 percent of the higher achieving students but only to 11 percent of lower achieving students. The other rule stated that being male and having a parent who (in 8th grade) does *not* believe that the student's test scores will be too low to qualify for college financial aid are more likely to be high achieving. The condition about parent expectations applied to 49% of the high achievers and 30% of the lower achievers, while after excluding females, it applied to 28 percent of high achievers and 11% of lower achievers.

Table 40. Variables associated with higher than expected math achievement in 12th grade, within 3 different 8th grade math achievement subgroups (Study 2)

Category	**Details** (square brackets refer to subgroups for which the rule was discovered, where 1 through 3 are groups with 8th grade math scores <17, 17-23 and 24-31, respectively)	**Groups to which rules applied** (TPR = true positive rate, PLR=positive likelihood ratio)
[Opportunity] Math course-taking in HS	Taken geometry. [1] Taken algebra 2. [2, 3]	**Gr8 math score <17** (TPR = 0.31/0.38; PLR = 1.8/2.3) **Gr8 math score 17-23** (TPR = 0.27/0.38; PLR = 1.7/2.3) **Gr8 math score 24-31** (TPR = .43-.49/.44-.5; PLR = 2.1-1.9/1.8-1.9)
[Opportunity] HS emphasis in academics	S attends academic HS program; "Very accurate that HS Ss are expected to do HW; 75-100% of HS Ss in academic HS program. [1] S attends academic HS program; T "agrees" that dept is committed to AP and honors programs. [2] Ss write science labs once a week [3]	**Gr8 math score <17** (TPR = .29-.44/.28-.43; PLR = 1.5-1.8/1.6-1.9) **Gr8 math score 17-23** (TPR = .31-.50/.40-.61; PLR = 1.6-1.7/1.7-1.9) **Gr8 math score 24-31** (TPR = 0.48/0.52; PLR = 1.5/1.5)
[Opportunity] School safety and climate	HS teacher considers robbery/theft, illegal drugs, alcohol and possession of weapons to be a "minor" problem at the school.	**Gr8 math score 17-23** (TPR = .32-.43/.35-.47; PLR = 1.6-1.9/1.5-2)
[Opportunity/other] Teachers have necessary materials	HS teacher "agrees" that necessary materials are readily available.	**Gr8 math score <17** (TPR = 0.29/0.43; PLR = 1.5/2) **Gr8 math score 17-23** (TPR = 0.46/0.53; PLR = 1.6/1.7)
[Opportunity/other] Teachers in a positive, learning-oriented culture	HS teacher "agrees" that they are encouraged to experiment with teaching, grading practices are consistent and fair, and/or department chair consults staff before decision. Teacher "disagrees" that routine practices interfere with teaching. Teacher reports that cooperative learning and higher-order thinking skills are discussed.	**Gr8 math score 17-23** (TPR = .31-.51/.25-/53; PLR = 1.6-1.9/1.5-1.8)
[Distal] Parent expects college	Parent expects 8th grader to attend a 4-5 year college program.	**Gr8 math score <17** (TPR = 0.42/0.38; PLR = 1.7/1.5)
[Propensity] P believes 8th grader's academics will not negatively interfere with college financial aid	Parent does not expect 8th grader's test scores and/or grades to be too low to qualify for financial aid.	**Gr8 math score <17** (TPR = .47-.51/.46-.44; PLR = 1.6-1.7/1.6-1.7)
[Propensity] P does not believe that they have not	8th grader's parent does not believe they have not been able to get much	**Gr8 math score <17** (TPR = 0.35/0.31; PLR = 1.6/1.6)

Table 40 continued

been able to get information about how to apply for financial aid.	information on how and where to apply for financial aid.	
[Distal] Parent expectation for algebra	8th grader believes their parents/guardian wanted 8th grader to take Algebra.	**Gr8 math score <17** (TPR = 0.3/0.4; PLR = 1.6/2)
[Opportunity] Enrollment in foreign language class	8th grader enrolled in a foreign language class.	**Gr8 math score 24-31** (TPR = 0.39/0.51; PLR = 1.5/1.7)
[Other] Availability of cheerleading	Cheerleading available to 8th graders at the school.	**Gr8 math score <17** (TPR = 0.35/0.32; PLR = 1.6/1.5)
[Other] 8th grader has been threatened once or twice	"Once or twice," someone has threatened to hurt 8th grader at school.	**Gr8 math score 24-31** (TPR = 0.29/0.28; PLR = 1.6/1.5)
[Distal] Parent/guardian was in mid 30s when 8th grader was born	Parent who responded to base year survey was born in 1940-1944 (48-52 years old in 1988; i.e., 34-38 years old when 8th grader was born).	**Gr8 math score 24-31** (TPR = 0.26/0.26; PLR = 1.6/1.5)
[Opportunity/other] Social studies teacher's teaching was observed several times by supervisor.	Social studies teacher reports that supervisor observed their teaching "several times."	**Gr8 math score <17** (TPR = 0.28/0.27; PLR = 1.7/1.5)

Table 41. Variables pairs associated with higher than expected math achievement in 12th grade, within 3 different 8th grade math achievement subgroups (Study 2)

Gr8 math	Conditions associated with high achievement in 12th grade math	TPR	FPR	PLR	Prec	FOR	RP
24-31	1yr of Alg2 & Parent agreed that 8th grader's homework is worthwhile	.35	.13	2.72	.23	.08	3.04
24-31	1yr of Alg2 & Parent agreed that school is preparing 8th grader well for HS	.31	.12	2.54	.22	.08	2.75
24-31	1yr of Alg2 & Parent agreed that school is a safe place	.32	.11	2.78	.23	.08	3.00
24-31	1yr of Alg2 & 8th grader did not talk to teacher about jobs/careers after HS	.39	.16	2.37	.21	.07	2.77
24-31	1yr of Alg2 & HS calendar is semesterized	.34	.14	2.53	.22	.08	2.82
24-31	1yr of Alg2 & HS had no plans to offer parent workshops on adolescent problems, drug/alcohol use prevention	.36	.14	2.54	.22	.08	2.90
24-31	1yr of Alg2 & HS at some point had offered staff development in adolescent characteristics and teaching strategies for secondary school students	.36	.14	2.51	.22	.08	2.84
24-31	1yr of Alg2 & Not the case that HS never offered staff development in adolescent characteristics and teaching strategies for secondary school students	.37	.15	2.50	.21	.08	2.86
24-31	1yr of Alg2 & Trigonometry is a regular course in HS	.35	.13	2.70	.23	.08	3.03
24-31	1yr of Alg2 & Hall passes needed for HS students to visit the library	.34	.13	2.61	.22	.08	2.90
24-31	1yr of Alg2 & Not the case that HS will transfer student to another school for the 2nd time skipping school 1-2 days.	.38	.15	2.51	.22	.07	2.91
24-31	1yr of Alg2 & Not the case that HS will transfer student to another school for the 2nd time skipping school 3 or more days.	.36	.14	2.52	.22	.08	2.86
24-31	1yr of Alg2 & In most recent/current math class, student has never used book other than textbook.	.36	.15	2.36	.21	.08	2.69
24-31	Has taken some Alg2 & 8th grader did not study religion outside of school	.32	.13	2.53	.22	.08	2.76
24-31	Has taken some Alg2 & Parent agreed that school is a safe place	.35	.13	2.63	.22	.08	2.94
24-31	Has taken some Alg2 & 8th grader did not talk to teacher about HS programs	.38	.15	2.57	.22	.07	2.97
24-31	Has taken some Alg2 & Student newspaper was available to 8th graders	.37	.15	2.58	.22	.07	2.97
24-31	Has taken some Alg2 & 21+ college reps sent to HS during 1989-90	.30	.11	2.58	.22	.08	2.74
24-31	Has taken some Alg2 & 0% of HS students receive program for pregnant girls	.34	.14	2.42	.21	.08	2.69

Table 41 continued

24-31	Has taken some Alg2 & HS requires less than a year of health	.33	.12	2.75	.23	.08	3.00
24-31	Has taken some Alg2 & Science club(s) available in 10th grade	.32	.13	2.46	.21	.08	2.69
24-31	Has taken some Alg2 & Other subject clubs available in 10th grade	.35	.14	2.50	.21	.08	2.81
24-31	Has taken some Alg2 & HS student morale is high according to school administrator	.33	.12	2.73	.23	.08	2.99
24-31	Has taken some Alg2 & HS student disagrees that drug sale/use is a problem	.33	.12	2.69	.23	.08	2.95
24-31	Has taken some Alg2 & HS student disagrees that violence on school grounds is a problem	.38	.15	2.53	.22	.07	2.92
24-31	Has taken some Alg2 & HS student disagrees that lack of discipline in class is a problem	.34	.13	2.58	.22	.08	2.87
17-31	8th grader goes to science museums & HS teacher agrees that they are encouraged to experiment with teaching	.31	.12	2.51	.24	.09	2.67
17-31	8th grader goes to history museums & HS teacher agrees that they are encouraged to experiment with teaching	.30	.12	2.40	.23	.09	2.53
17-31	Very accurate that HS Ss are encouraged to enroll in academic classes & HS teacher agrees that they are encouraged to experiment with teaching	.28	.10	2.86	.27	.09	2.89
17-31	10th grader reports that most recent/current math class places major emphasis on thinking about what a problem means and how it might be solved & In-depth study of a specialized subject was not discussed in teacher enrichment programs HS teacher attended this year	.24	.10	2.42	.23	.10	2.44
< 17	Student is male & Was not the case that parent believed 8th grader's test scores will not be good enough to qualify for college financial aid	.30	.12	2.50	.22	.08	2.69
< 17	Student is male & 75-100% of HS students in academic counseling program	.28	.11	2.51	.22	.08	2.64

TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving.

4.2.3 Summary of Study 2 Results

Key findings from Study 2 are summarized in Table 42. Predictive accuracy of ruleset mining was comparable to regression, regardless of dataset size. In contrast to regression and hierarchical regression models that attributed the outcome variance to many different predictors, ruleset mining accounted for most of the explainable variance with 8th grade math scores, followed by math course-taking. In addition, the ruleset models (but not the regression models) identified a few predictor-outcome relationships that were specific to just some subgroups—for example, the rulesets suggested that additional variables such as math course-taking, improved the prediction for some students, but generally not for those who score highest or lowest in 8th grade math.

The expansion of the dataset allowed for identification of predictors that were not considered by Byrnes and Miller, such as high school's emphasis in academics, and parents' expectations about the students' ability to qualify for college financial aid. Association rule mining using subgroups of students (e.g., only students of a particular SES range or level of parental education) helped identify characteristics related to the outcome that were unique to the subgroup.

Table 42. Key findings from Study 2

Method	Key findings
Hierarchical multiple regression	Model accounted for 76% of variance (with over 90% of that sufficiently explained by math self-concept; and with most variables accounting for <1% of unique variance in the outcome). Distal factors (incl., SES, parent & student expectations and middle school GPA) explained 43% of 12th grade math score variance, opportunity factors (esp., math course-taking) explained 11% more (or 45% by itself), propensity factors (esp. MS math achievement) explained 22% more (or 73% by itself). Demographic factor explained less than 1% of remaining variance. Following factors had high correlations with outcome variable: each of the four distal factors such as 8th grade SES and parent expectations (correlation between .42 and .56), 1 year of general math (-.37), 1 year of geometry (.53), and 1 year of Algebra II (.37), math achievement before the start of 8th grade (.84), GPA in grade 9 and 10 (.44), and math self-concept (.40)
Ruleset mining with 29/1933 possible predictors	Predictive accuracy was comparable to regression. 8th grade math scores most predictive of outcome, followed by math course-taking. 8th grade math scores are the most important predictor across all ruleset models, included in every rule. When outcome is dichotomized, additional factors (primarily math course-taking and sometimes other factors such as middle school GPA, math self-concept, SES) improve the prediction, but generally not for those who score highest or lowest in 8th grade math. The CART mosaic plots—with the numeric outcome—indicated that 8th grade math scores were most predictive, and that what was next most predictive depended on that score. For those who scored the lowest in 8th grade math, geometry and general math course-taking were next most predictive. For those who scored higher, geometry course-taking and math self-concept were next most predictive. For those who scored even higher in 8th grade math, math self-concept was next predictive. For those who scored the highest, grade 8 GPA was next predictive, if anything.

Table 42 continued

Method	Key findings
Association rule mining	Math course-taking and HS emphasis in academics were associated with higher-than-predicted* achievement regardless of 8th grade scores, but there were slight differences on how these predictors were operationalized across different subgroups of 8th grade math scores. Re: Rules of length 2, with PLR >1.5: Parent's academic expectations for the 8th grader to take algebra, attend college, and/or qualify for college financial aid were associated with math achievement only for students whose 8th grade test scores were very low. School climate, safety, and teacher perception of their professional environment were associated with achievement only for students whose 8th grade scores were between 17 and 23. Some rules appeared less reliable (more one-off) than others. Re: Rules of length 3, with PLR >2.5: For students whose 8th grade math scores 24-31, algebra course-taking was included in every rule antecedent, some type of school factor (e.g., safety, worthwhileness of homework, availability of student clubs, and use of textbooks and hall passes) was included in many, while demographics, and peer or community variables were not included in any. Very few rules characterized those whose 8th grade math scores were lower than 24. For students with 8th grade math scores 17-23, the condition that the high school teacher agreed they are encouraged to experiment with teaching was true for 50% of the high achievers and 29% of others. If in addition, the 8th grader had visited with science/history museum or it was "very accurate" that the high school encourages students to take academic classes, the rule became true for about 30% of high achievers and just 12-15 % of lower achievers. For those with 8th grade math scores <17, being male and either attending a school that places emphasis on academics, or whose parents do not believe the students' test scores will be too low to qualify for college financial aid, were likely to be higher achieving than CART predicts. *Higher than predicted by CART, which primarily relied on 8th grade math scores.

5.0 DISCUSSION

This chapter discusses what I learned through this experiment with rule induction methods, including: difference between generating rulesets vs generating rules; stages of mining rulesets and rules and associated considerations; how rule induction and data mining methods do and do not add value beyond regression approaches; and recommendations on practical and principled ways to use rule induction approaches in education research. I also discuss limitations to my study and next steps for research.

5.1 WHAT I LEARNED

My two substudies suggested several ways that rule induction is different from and adds value to regression approaches, and some ways they may not be so helpful. They also gave me several ideas about how to integrate rule induction approaches into education research and evaluation in practical and principled ways. Discussion on these—addressing my main research questions—are presented in section 5.1.3 and 5.1.4. What also became much clearer to me through this project was the difference between generating rules and generating rulesets. Because this difference is key to understanding how and why rule/ruleset induction approaches are and are not helpful, I begin this section with this topic (section 5.1.1). Another somewhat unexpected and hard-earned realization was how the output of rules and rulesets did not automatically inspire insight. In fact, moving from data output to insight required a good amount of labor, including identifying and creating appropriate representations of the output. I discuss this in section 5.1.2, prior to

responding to the main research questions, because the effort required to move from output to insight is an important consideration for assessing the usefulness of rule induction.

5.1.1 Difference between generating rulesets vs rules

This project helped me realized that ruleset induction is better suited for modeling the population rather than identifying specific rules. This is because rules within rulesets are interrelated, and as a newspaper article loses much of its meaning when viewed in isolation from the context in which it was created, rules lose the relational meaning against its context when viewed in isolation from the ruleset within which it was created. For example, a C4.5 rule associating 12th grade math scores with high 8th grade math achievement, geometry, and high math self-concept (Figure 19), becomes more meaningful when examining the entire model in which that rule is embedded. The context indicates the order in which the predictor variables were associated with the outcome, and how frequently specific attribute-values appear in other rules. For this example, the rule by itself suggests that interventions on 8th grade math, geometry and math self-concept may help a small group of individuals, while the ruleset additionally suggests that interventions on 8th grade math is helpful for everyone while interventions on math self-concept may matter to just a smaller portion of students, and so on.

Thus, the ruleset approach is less suited for identifying interesting individual rules because of its unexhaustive nature of the rule search, and built-in dependency among rules (i.e., predictors are necessarily shared in tree models, while for covering models later rules presume earlier rules had been applied). In other words, rules within a ruleset are less reliable because they are not only a function of the predictors' relationship to the outcome, but also a function of the search algorithm and other rules within the ruleset. Association rule mining, on the other hand, conducts an

exhaustive search for rules, and each rule it generates is independent of the search algorithm or of other rules that were identified within the ruleset. Thus, association rule mining would be more helpful than ruleset modeling for identifying interesting rules, and indeed it turned out as such with this project.

What was least obvious about this to me until conducting ruleset induction, was how association rule mining could be more helpful than trees or covering models for identifying attribute-values that are associated to some subgroups and not others. Because this task involves identifying relationships across different rules, I had initially thought that ruleset mining would be the best and perhaps only way to go. However, trying to infer meaning from the rulesets in this project helped me better understand the limits associated with rulesets in accomplishing this task, and helped me think of a way to use association rule mining in a comprehensive and arguably more helpful way.

5.1.2 Stages of mining rulesets and rules

Rule and ruleset induction happened in stages, each associated with unique difficulties and considerations that were sometimes learned the hard way. Figure 27 sketches the stages of ruleset and rule mining process that used for this project. I first generated rulesets from the NELS data, then created representations of the output to make it easier to interpret, then finally gleaned findings and possible implications. For both studies, the ruleset findings helped decide subgroups for which to conduct association rule mining by suggesting variables (including cut-scores of numeric variables) that were strongly related to the outcome. I then conducted similar processes for association rule mining—going from data to output (steps 1 and 5), to representations of output (steps 2 and 6), then to insight (steps 3 and 7). While it not necessary that ruleset mining (steps 1-

3) precede rule mining (steps 5-7), for this project, ruleset results helped identify why and how further rule induction should occur (step 4).

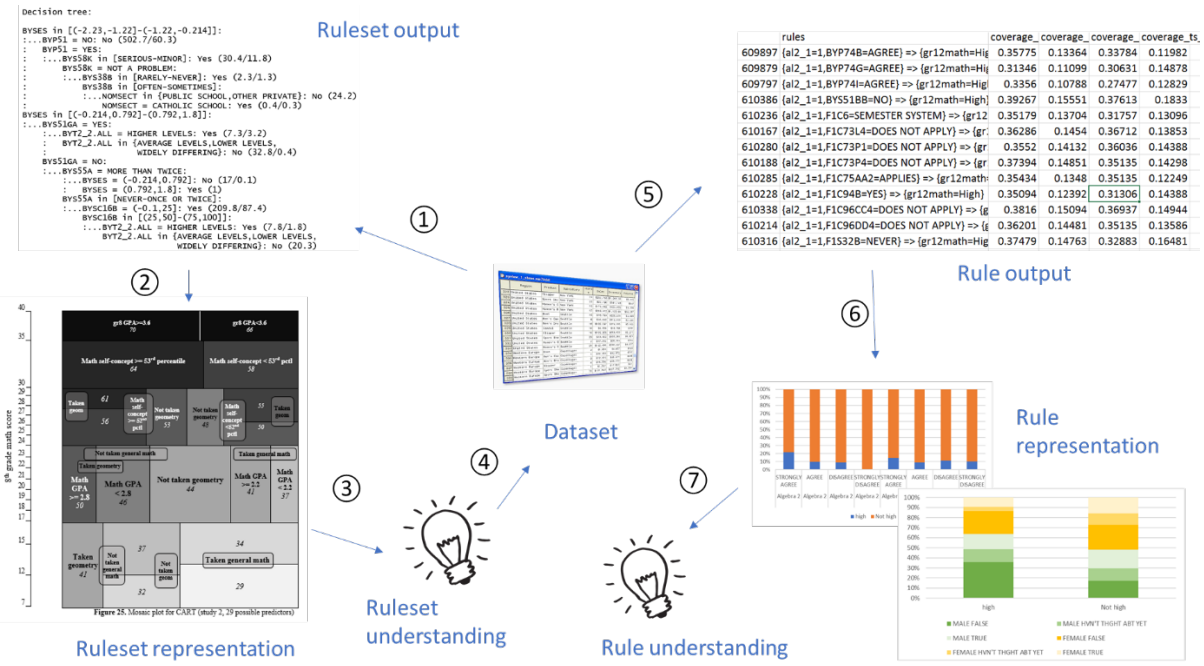


Figure 27. Illustration of rule and ruleset mining process for project

5.1.2.1 From dataset to rule induction output

For both ruleset and rule-mining, the first step was going from a dataset to algorithm output (steps 1 and 5 in Figure 27). This process took many iterative sub-steps, including understanding the problem, data, and available algorithms, learning R programming, and cleaning the data so that it was appropriate to the problem and algorithm. It took particularly long to identify variables that should/should not be included as predictors, and decide what to do with categorical variables with many levels. Several times, I had to re-run analyses due to accidental inclusion of variables that were too related to the outcome. The process of getting to the ruleset also involved understanding and making decisions about parameter settings, and how to make the results comparable as possible across algorithms. The large size of the dataset (especially for the second rounds of each

study) wreaked havoc, particularly when generating lengthy rules using Apriori. My laptop ran out of memory. As a solution, I tried using a super-computer, which involved another learning curve about technical details of remote computing, only to learn that the lengthier rules were not particularly helpful to this project anyway.

5.1.2.2 From rule induction output to output representations

Obtaining insight from the output was not as easy as my literature review suggested. Ruleset and association rule output were generally difficult to interpret due to the lengthiness of the output and the minimally informative variable names. The relative node sizes of trees and rule correctness were also difficult to glean from just the output, since the outputs generally listed the number of people to whom the rule applied correctly/wrongly but not their proportions, and tended to be only indicated in terminal nodes. Interestingness measures of rules were also not automatically generated in the output. In addition, the ordered nature of the sequential covering rules requires one to have to keep in mind the status of every rule that came before it (since new rules apply only to those that were not covered by the previous rules). Each of these factors taxed my working memory as I tried to interpret each result and compare findings across algorithms. Without further processing of these outputs, it was very hard to interpret each ruleset from the raw output, and even more so to compare ten or so rulesets, and hundreds of individual rules, against one another.

Thus, as shown as the second step in Figure 27, I created visualizations of rulesets from the raw output. These included, for ruleset approaches: word clouds that showed the frequency of variable use, bar charts comparing variable importance across algorithms, tables with confusion matrices and validity metrics, scatter plots comparing F-measures and kappa statistic across models, and mosaic diagrams that spatially represented all the rules and relative applicability within each model. Representations for association rule mining included: validity measures for

each rule, correct variable labels, short lists of rules based on relevant validity measures, categories of rules based on conceptual similarity, and final tables of promising rules that allow for ease of understanding. Each of these were time-consuming to create as they required conceptualization, additional calculations, and labeling and interpreting of variables. It was particularly frustrating when I realized, after creating several representations, that there was a variable in the mix that should not have been there, which meant that had to re-run models and re-create possibly all the representations. The process was both quantitative and qualitative: I manipulated large tables of numbers through multiple rounds, with the judgment of what is sensible to do relying heavily on qualitative meaning-making of available results and their relationship with research goals. When comparing qualitative aspects of rulesets or rules, it was helpful to represent common elements (across rules or rulesets) through physical proximity—for example, to streamline the Y-axis across all mosaic plots, and to cluster association rules with common antecedents.

5.1.2.3 From output representations to insight

With the ruleset or rule representations created, the third stage was to interpret the results and obtain findings. How accurately the rulesets predicted the individuals was straight-forwardly understood through the confusion matrices and accuracy measures. Whether any rulesets or rules provided information that was not obtained by regression took a little more effort to ascertain because the relevant resources are spread across multiple pages, and decisions about relevance involved both empirical and theoretical, and qualitative and quantitative considerations. It was often necessary to create higher-level abstractions of representations, to keep from getting lost in the details (so I often went back and forth between meaning making and creating representations). Mental and physical fatigue from the previous stages also sometimes made interpretation difficult, including contribution of first-time jitters and fear that nothing interesting might be found after the

big search. In retrospect, it would have helped me a lot in this stage to not only be clearer about the difference between rules and rulesets as described above (Section 5.1.1), but also have a better understanding about the different questions that can be answered by rule/ruleset-induction as opposed to regression. I explain what I mean by the latter in the following two sections.

5.1.3 How rule induction data mining methods added, or could add value, beyond traditional statistical approaches; how they were *not* more helpful

In both studies, rule induction models had comparable or slightly worse predictive accuracies relative to regression, although the ensemble approaches sometimes performed slightly better. This is unsurprising given that no algorithm is superior across all dataset, and model performance depends largely on the nature of the dataset (i.e., no free lunch theorem). However, with the aid of the visualizations, this study clarifies how rule and ruleset induction could be uniquely helpful relative to traditional regression, and how they cannot.

Rulesets can provide researchers with a unique description of the sample, different from a regression-based picture, that shows at-a-glance, how some of the key predictors were related to the outcome and to each other. This descriptive value was particularly vivid for the models in Study 2, where the mosaic diagrams characterized the sample and their 12^{th} grade math scores first in terms of 8^{th} grade math scores, then mainly in terms of math course-taking. The difference in descriptive power is largely because rulesets describe every respondent, while regression tables describe average contributions by individual variables (an abstraction, so more difficult for the mind to imagine). Ruleset results were also clearer than regression in expressing how just a handful of variables were generally sufficient to explain most of the explainable outcome variance. Byrnes and Miller provided analogous information through presenting partial correlations of each

predictor with the outcome, but in general, relative contributions of predictors to the outcome are somewhat difficult to ascertain in regression unless one knows how to interpret results tables. Thus, the model limitations are perhaps more visible with rulesets, which can keep researchers from overextending their inference.

Rules and rulesets also can help identify relationships between variables that held for some subgroups but not others. For example, rulesets in Study 2 suggested that Algebra 2 and math self-concept were positively related to 12th grade math scores, but only for those who were higher achieving in 8th grade math. Similarly, general math was negatively related to 12th grade math scores but only for those who scored lower on 8th grade math. Association rule mining provided similar kinds of findings. In Study 1, for example, several factors (e.g., participation in honors or gifted and talented programs, school safety) were more strongly associated with 12th grade achievement for lower income students, and students whose parents did not have a college degree. In contrast, regression provides how each predictor contributes, on average, to the population.

In addition, rule induction could identify cut-points of continuous predictors, and groupings of nominal predictors, that could be useful for prediction and further analysis. For example, the 8th grade math subgroups identified by CART in Study 2 motivated the next step of using those scores as cut-points for creating subsamples to conduct association rule mining.

Being data mining approaches, rule and ruleset induction also easily identified variables that are related to the outcome that were not included in the regression model. For example, in Study 1, "ever held back in school" and parental views on 8th grader's academic eligibility for college financial aid were found to be strongly related to the outcome, although they were not considered by the author of the original study. However, it is important to keep in mind that inclusion of unexpected variables in the model is not a unique benefit of ruleset induction *per-se*,

but rather a feature of any kind of data mining that makes its variable choice explicit. It is possible to take a regression-based data mining approach to data mining (e.g., stepwise regression, all-possible subsets regression) that likewise may have identified different variables.

As trite and obvious as this may sound, rule induction approaches were not directly helpful in answering exactly the research questions that regression is specifically designed to answer. Multiple regression asks (1) whether a set of variables together are related to the outcome, and if so (2) whether *on average,* individual variables are related to the outcome after adjusting for the relationships that exist between all other variables and the outcome. Hierarchical regression can additionally find (3) whether *on average,* predictors or sets of predictors predict the outcome over and above another set. It assumes that each predictor is linearly related to the outcome (linearity), that whatever level of predictors we are looking at (e.g., differing levels of parental education, differing levels of math course-taking) the variability of mispredictions are constant (homoscedasticity), errors are normally distributed, there are no outliers, subjects are independent of one another, and that the predictors are not a function of another (no multicollinearity/singularity). Ruleset induction methods can determine the first question. However, they do *not* quite tell us whether individual variables are related to the outcome, *on average, all else being equal*. Instead, it tells us whether a variable could be relevant for some subgroup(s) and suggests a possible type of subgroup in which it would be relevant. Being heavily trained in regression-based approaches, it was easy to forget this, however, and try to look at ruleset results as an alternative method to conducting regression. The next section builds on the structural difference between the two approaches and recommends uses of ruleset and rule induction that do complement rather than conflate it with regression.

5.1.4 Recommendations on practical and principled ways to use the various rule induction approaches in education research

The key for using rule induction as a complement to regression is to keep clear in mind the difference in the research questions it answers. Ruleset induction answers an exploratory and sequential question: *What is a set of characteristics that tend to be commonly associated with each level of the outcome, and to those to whom that set of characteristics do not apply, what (if any) is another set of characteristics that would apply, (and so on)?* Regression answers: *Assuming every independent variable has the same amount of impact on the individual after controlling for other factors, how much unique impact does each independent variable have?* So instead of whether and which variable significantly predicts the outcome, rule induction is better equipped to characterizing *to whom* does a factor matter for predicting a desired (or undesired) outcome, and what are some factors that those with similar levels of outcome have in common? The temptation I fell into while assessing ruleset results was to focus on the relative predictor importance, and wonder about the average effect it had on the population, which regression answers much better.

With that in mind, there appear to be at least three practical and principled ways to incorporate rule induction into education research in the future. (1) Use of ruleset mining to describe the sample, and how some of the key independent and dependent variables relate. This was the way ruleset induction was used in this study, and leads to descriptions of how characteristics associated with different level of subgroups. (2) Use of association rule mining to identify what factors, if any, are different across groups. The groups could reflect differences in outcome (e.g., high achieving vs low achieving), in treatment, or background. This was the way association rule mining was used in this study. (3) Use of decision trees to identify whether a predictor or predictors are related to the outcome, *after controlling for key covariates*. I have not

tried this method in this project, but can see this being answering similar questions to tests for individual variables in regression, and to hierarchical linear regression. The idea would be to first model the outcome with a set key of covariates, then at each of the terminal nodes to investigate the relationship(s) between the outcome and the independent variables(s) of most interest using e.g., regression or a non-parametric approach. There are algorithms that instantiate some versions of this, such as the logistic model tree that conducts logistic regression at each terminal node of a decision tree.

Other factors that could be helpful to keep in mind when conducting ruleset- and rule induction are:

- Sharpen the research question prior to mining—identify precisely the outcome and the sample.
- Make sure in the data cleaning stage to exclude from consideration variables that essentially are duplicates of the outcome, and to exclude or categorize variables with many levels. Accidental inclusion of these variables in an uncategorized form would unnecessarily increase computational demand.
- Thoroughly check whether levels of categorical data should be combined (esp., levels in a Likert scale, and nominal categories with many levels).
- Remain focused on the big picture. Curb enthusiasm and interpretation until accuracy & generality of each rule is calculated. When possible, for ruleset generation, use software and algorithms that produce key accuracy statistics for each rule.
- Maintain a codebook in a spreadsheet format and use it to automate the labeling of variable names in output.
- Maintain clear records of data manipulation process, syntax files, and output.

- It is generally beneficial to limit the number of variables to some of the most predictive and relevant. Irrelevant or weakly relevant variables only makes the data mining and examination lengthier and more difficult.
- Attending to new packages and package updates could help save time.
- Use multiple algorithms for ruleset-induction, and keep in mind that the view that each output provides is one correct view of many.
- Guard against sampling bias by training-test approach, and/or by considering rule validity in other ways (e.g./esp., distrusting one-off rules, considering coherence with theory, considering practical significance of findings). Consider the effects of multiple comparisons increasing the chance of Type I error when generalizing a finding from a sample to a population.
- For association rule mining, generate the rules only once per sample, and generate all metrics needed so that all elements of the rule confusion matrix can be derived. This involves splitting the sample by key subgroups. Splitting the sample also helps reduce computational burden.

5.2 LIMITATIONS AND NEXT STEPS

5.2.1 Limitations to my study

There were several limitations to my study. First, being an exploratory study, there were many decisions that I made about problem understanding, methods and interpretation that were felt rather subjective. My personal experiences, knowledge and capacities impacted each step of the study

such that specific results may not be easily replicated, and other researchers may glean different insights than me. Second, I did not consider statistical approaches to account for the increase in Type I error that comes with multiple comparisons. This is a relatively serious limitation and a very important topic for future research. Third, predictions get worse for CART and bagged CART when the number of variables were increased, but I was unable to identify how that happened (it might have been sampling bias that occurred when splitting the sample to training and test sets).

In addition, there were many small procedural steps in the data preparation process that could have been done differently in retrospect. For example, for Study 2, I applied weights before splitting into training and test sets for ruleset mining, and wonder whether I should have done it differently so that the same subject does not appear in both the training and test sets. In addition, there might have been a better way to handle missing values for rWeka ruleset algorithms than to take the mean of numeric variables and combining the missing values in categorical variables as a new category. It may also have been beneficial to create new categories for Likert scale items for association rule mining, especially categories that combine e.g., "strongly agree" and "agree." The cleaned data also did not account for reporter-report dependencies (e.g., demographic of teacher may affect the judgment of the student).

5.2.2 Next steps for research

Furthering understanding and use of rule induction to education research would require more frequent application in research and evaluation, and more methodological research. Applying the rule induction more frequently in research and evaluation will help improve understanding of the practical significance and barriers associated with the method and its various instantiations (i.e., different software packages and algorithms). Use in different datasets would also refine and

expand the findings of this project, and help optimize rule/ruleset induction use. In tandem, it would be helpful to further explore methodological fronts such as: hierarchical or otherwise staged approaches to rule induction (including confirmatory approaches); ways to best account for increased Type I error due to multiple comparisons; assumptions underlying ruleset models and their comparison to regression-based approaches; approaches to using sample weights in rule induction; additional targeted uses of association rule mining; theoretical/practical comparisons to other group-induction methods such as cluster analysis; approaches to weighing independent variables for ruleset mining (i.e., for whatever reason, one may want to place more value/weight on certain variables over others); and improvements to non-greedy ruleset-induction approaches.

5.3 CONCLUSION

Rule induction identify sets of attribute-values that are commonly associated with each level of the outcome. This approach differs fundamentally from regression-based approaches that identify average associations between outcome and set of predictors. Rule induction approaches therefore expand the set of quantitative research questions education researchers can ask, to include those about the nature and generalizability (size) of the commonly found attribute-value sets, and whether/how the elements in the set differ across subgroups. Particularly when there is reason to believe that relationships between the predictor and outcome are not uniform across the population, rule induction can provide better help than regression in exploring those relationships. In addition, both rule induction and regression can both be used for prediction and identifying variables that relate more strongly to the outcome, but which method yields better and more useful results is difficult if not impossible to determine *a priori*. So, for those purposes, rule induction may not

necessarily add value. However, because rule induction generates a lot of output, and often requires multiple stages of consideration and distillation to arrive at sensible and relevant results, it can consume time and resources, particularly when the dataset is large and the research questions are not clearly articulated. Having clarity about the quantitative research questions, including why exactly rule induction is to be used for those questions, is a minimal requirement for education researchers to incorporate rule induction effectively into their research.

APPENDIX A

VARIABLE DESCRIPTIONS FOR THOMAS 2006 RE-ANALYSIS

Variable	Relevant variable(s) in NELS:88 dataset	Transformation required for precise replication	Modifications for replication
Sample selection			
1. Race	F4RACE	Drop all cases except 3 (Black, not Hispanic), resulting in 1176 cases.	RACE, F1RACE, F2RACE1
Sample identification			
1. Student ID	STU_ID	Not used	
Outcome variable			
1. High achievement (highach)	F22XRC, F22XMC, F22XSC, F22XHC. (1-99. 12th grade reading, math, science, and history centiles, respectively; 998. Missing; 999. Test not complete.)	Recode missing value to mean, by sex (per author's code). Average 4 scores. If top quartile of sample (43 or above) then highach=1; else = 0.	Use IRT theta scores (F22XRTH, F22XMTH, F22XSTH, F22XHTH; SC248V.) from F2 student part 2, and F2 weight (F2QWT; SC201V.) to create centile scores. Then, impute missing data for F2 dataset. Convert this to centile score within
Student characteristics			
1. Sex (female)	F4SEX (1. Male; 2. Female) No values were missing.	If F4SEX is 2, female = 1; else = 0. Total of 649 females and 527 males.	SEX Use F2SEX instead in Student part 2.
2. Parental education	BYPARED (1. Didn't finish HS; 2. HS Grad or GED; 3. >HS & <4yr deg; 4. College graduate; 5. MA/equivalent; 6. PhD/MD/Other; 7. Don't know; 98. Missing; 99. Legitimate skip/not in wave)	"7. Don't know" was not coded as missing, and instead regarded as part of the education scale. Missing were replaced with median.	
3. Number of siblings 8th grader has	BYP3A (0. None; 1. One, …. 6. Six or more; 96. Multiple response; 98. Missing; 99. Legitimate skip/not in wave.)	Missing were replaced with mean, by sex.	
4. Parents' marital status	BYPARMAR (1. Divorced; 2. Widowed; 3. Separated; 4. Never married; 5. Marriage-like relat; 6. Married; 98. Missing; 99. Legitimate skip/not in wave)	This nominal variable was most likely erroneously included into the model. Replication was achieved when missing were replaced with mean, by sex. Mean, standard deviation and model results were presented but not explained in the original paper.	Excluded from analysis.
5. Single parent (sinpar)	F2P7 (Current marital status. 1. Single, never married; 2. Married; 3. Divorced/separated; 4. Widowed; 5. Living like married;	If F2P7 is 2, sinpar = 0; else = 1. Missing were coded into median.	F2 parent (PC8V.)

Variable	Relevant variable(s) in NELS:88 dataset	Transformation required for precise replication	Modifications for replication
	6. Multiple response; 8. Missing; 9. Legitimate skip/not in wave)		
6. Income from all sources 1991	F2P74 (1. None; 2. Less than $1000; 3. $1000-2999; 4. $3000-4999; 5. $5000-7499; 6. $7500-9999; 7. $10000-14999; 8. $15000-19999; 9. $20000-24999; 10. $25000-34999; 11. $35000-49999; 12. $50000-74999; 13. $75000-99999; 14. 100000-199999; 15. $200000 or more; 96. Multiple response; 98. Missing; 99. Legitimate skip/not in wave.	Missing were recoded with mean, by sex.	F2 parent (PC57V.)
7. Hours of homework in school	F2S25F1 (0. None; 1. Less than 1 hour; 2. 1-3 hours; 3. 4-6 hours; 4. 7-9 hours; 5. 10-12 hours; 6. 13-15 hours; 7. 16-20 hours; 8. Over 20 hours; 96. Mult response; 98. Missing; 99. Legitimate skip/not in wave.)	Coding error in original paper: "8. Over 20 hours" seems to have been coded as missing, and all missing were coded into "2. 1-3 hours". Missing were coded into median.	
8. Hours of homework out of school	F2S25F2 (0. None; 1. Less than 1 hour; 2. 1-3 hours; 3. 4-6 hours; 4. 7-9 hours; 5. 10-12 hours; 6. 13-15 hours; 7. 16-20 hours; 8. Over 20 hours; 96. Mult response; 97. Refused; 98. Missing; 99. Legitimate skip/not in wave.)	Coding error in original paper: "8. Over 20 hours" seems to have been coded as missing, and all missing were coded into "2. 1-3 hours" Missing were coded into median.	Student part1
Family variables			
1. Household resources (hhressc)	F2N12A (family has a specific place to study), F2N12B (Family receives daily newspaper at home), F2N12D (Does family have an encyclopedia), F2N12E (Does family have an atlas), F2N12F (Does family have a dictionary in the home), F2N12H (Does family have a computer in the home), F2N12M (Does family have 50+ books in the home), F2N12O (Does family have a calculator in the home) For all, 1. Have; 2. Do not have; 7. Refused; 8. Missing; 9, Legitimate skip/not in wave.	Recode so that 1 (have) = 1, and else=0. Hhressc = sum of the recoded values. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	Student part 2 (SC266V.)
2. Parents pay for tutor (paytutor)	BYP82D (Do you currently have any of the following educational expenses for any of your children? Any educational expenses for tutoring.) 1. Yes; 2. No; 3. Don't know; 6. Multiple response; 7. Refusal; 8. Missing; 9. Legitimate skip/not in wave.	paytutor = 1 if 1, else=0. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	
3. Private school (privsch)	G12CTRL1 (School classification reported by school) 1. Public; 2. Catholic; 3. Priv/oth relig; 4. Priv/non-relig; 5. Priv/not ascrtnd; 98. Missing.	privsch = 1 if ~=1, else 0. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	Student part2 (SC243V.)

Variable	Relevant variable(s) in NELS:88 dataset	Transformation required for precise replication	Modifications for replication
4. Religious school (religsch)	G12CTRL1 (School classification reported by school)	religsch = 1 if =2 or =3, else 0. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	As above
5. Autonomy (autonomy)	F2S98A, F2S98B, F2S98C, F2S98D (Who decides how late R can stay out, Who decides when R can use car, Who decides if R can have job, Who decides how R will spend money, respectively. 1. Parent(s) decide; 2. Prnts dcided w/me; 3. We decide together; 4. Decide w/prnts; 5. Decide myself; 6. Mult response; 7. Refused; 8. Missing; 9. Legitimate skip/not in wave.)	Missing were recoded with mean, by sex. autonomy = mean of 4 variables.	Student part 2 (SC92V. SC93V. SC94V.)
6. Parental involvement in school (pinvolve)	BYP59A, BYP59B, BYP59C, BYP59D, BYP59E (Do you and your spouse/partner do any of the following at your eight grader's school? A. Belong to a parent-teacher organization, B. Attend meetings of a parent-teacher organization, C. Take part in the activities of a parent-teacher organization, D. Act as a volunteer at the school, E. Belong to any other organization with several parents from your eighth grader's school (e.g., neighborhood or religious organizations). 1. Yes; 2. No; 6. Multiple response; 8. Missing; 9. Legitimate skip/not in wave.)	Recode so if BYP59A-E = 0 unless it's =1. pinvolve = sum of 5 recoded variables. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	
7. Parents expect college (pexpcol)	F2S41A, F2S41B (What do the following people think is the most important thing for you to do right after high school? Father and mother, respectively. 1. Does not apply; 2. Go to college; 3. Get ft job; 4. Enter trade school; 5. Enter military; 6. Get married; 7. Do what I want; 8. They don't care; 9. I don't know; 96. Multiple response; 98. Missing; 99. Legitimate skip/not in wave.)	perxpcol = 1 if either is equal to 2, else = 0. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	Student part 1 (SC41V.)
Peer variables			
1. Good peers (goodpeer)	Among Friends how important is it to: Attend classes regularly? (F2S68A); Study? (F2S68B); Get good grades? (F2S68D); Finish high school? (F2S68F); Continue their education past high school? (F2S68H). For each, 1. Not important; 2. Some importance; 3. Very important; 8. Missing; 9. Legitimate skip/not in wave.	Missing were recoded with mean, by sex. Goodpeer = mean of resulting 5 variables.	Student part 1 (SC40V.)

Variable	Relevant variable(s) in NELS:88 dataset	Transformation required for precise replication	Modifications for replication
2. Bad peers (badpeer)	Among Friends how important is it to: Go to parties? (F2S68M); Have sexual relations? (F2S68N); Use drugs? (F2S68O); Drink alcoholic beverages? (F2S68P). Levels same as goodpeer variables, above.	Missing were recoded with mean, by sex. Badpeer = mean of resulting 4 variables.	F2_DS03 (stu2) (SC40V.)
3. Person student admires is intelligent (admintel)	F1S71D (Person student admires most among all people they personally know is intelligent. 1. Applies; 2. Does not apply; 8. Missing; 9. Legitimate skip/not in wave)	Admintel = 1 if F1S71D = 1, else 0. I was able to replicate Thomas' distributional results but not descriptive results.	F1_DS01(stu)
4. Peers expect college (peerexcl)	F2S41C (What do friends think is the most important thing for you to do right after high school? 1. Does not apply; 2. Go to college; 3. Get ft job; 4. Enter trade school; 5. Enter military; 6. Get married; 7. Do what I want; 8. They don't care; 9. I don't know; 96. Multiple response; 98. Missing; 99. Legitimate skip/not in wave.)	peerexcl = 1 if =2; else =0. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	Student part 1 (SC41V.)
Community variables			
1. Activities outside of school (activity)	Has your eighth grader attended classes outside of his or her regular school to study any of the following? Art (BYP60A), Music (BYP60B), Dance (BYP60C), Language (BYP60D), Religion (BYP60E), The history and culture of his/her ancestors (BYP60F), Computer skills (BYP60G), Other (BYP60H). 1. Yes; 2. No; 6 Multiple response; 8. Missing; 9. Legitimate skip/not in wave.	Recode so that if ~=1, then =0. activity = sum of all recoded variables. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	
2. Student's cultural activities (sculture)	Does your eighth grader take part in any of the following activities? Borrow books from the public library (BYP61AB), Attend concerts or other musical events (BYP61BB), go to art museums (BYP61CB), go to science museums (BYP61DB), go to history museums (BYP61EB). 1. Yes; 2. No; 6 Multiple response; 8. Missing; 9. Legitimate skip/not in wave.	Recode so that if ~=1, then =0. sculture = sum of all recoded variables. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	
3. Neighborhood safety	F2P60 (How safe is neighborhood. 1. Very safe; 2. Somewhat safe; 3. Somewhat unsafe; 4. Very unsafe; 8. Missing; 9. Legitimate skip/not in wave.)		F2 parent (PC49V.)
4. Neighborhood diversity	F4JRDVA (What percentage of the people in the neighborhood where you grew up were the same race and ethnicity as you? 0-100; -1. Don't know; -2; refused; -3.	Seems like 0 was also coded as missing Missing were coded into median.	

Variable	Relevant variable(s) in NELS:88 dataset	Transformation required for precise replication	Modifications for replication
	Legitimate skip; -7. Not reached-partial/abbrev interview.)		
School variables			
1. Public school (public)	G8CTRL (School type. 1. Public school; 2. Catholic school; 3. Private school, other religious affiliation; 4. Private school, no religious affiliation. 9. Legitimate skip/not in wave.)	public = 1 if =1, else =0.	
2. Urbanicity of school (urban)	G8URBAN (1. Urban; 2. Suburban; 3. Rural; 9. Legitimate skip/not in wave.)	urban = 1 if =1, else =0. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	
3. Percent minority in school	G8MINOR (Percent minority in school. 0. None; 1. 1-5; 2. 6-10; 3. 11-20; 4. 21-40; 5. 41-60; 6. 61-90; 7. 91-100; 998. Missing; 999. Legitimate skip/not in wave.)	Missing were coded into median.	
4. Percent receiving free lunch in school	G8LUNCH (Percent students receiving free or reduced lunch. 0. None; 1. 1-5; 2. 6-10; 3. 11-20; 4. 21-30; 5. 31-50; 6. 51-75; 7. 76-100; 998. Missing; 999. Legitimate skip/not in wave.)	Missing were coded into median.	
5. School climate (climate)	Indicate the degree to which each of the following matters are a problem in your school. Student tardiness (BYS58A), student absenteeism (BYS58B), students cutting class (BYS58C), physical conflicts among students (BYS58D), Robbery or theft (BYS58E), vandalism of school property (BYS58F), Student use of alcohol (BYS58G), Student use of illegal drugs (BYS58H), Student possession of weapons (BYS58I), Physical abuse of teachers (BYS58J). (1. Serious; 2. Moderate; 3. Minor; 4. Not a problem; 6. Multiple response; 8. Missing; 9, Legitimate skip/not in wave.)	Recode so that if ~=1, then =0. climate = sum of all recoded variables. [Not ideal since missing value regarded as 0; Should have filled missing values first.]	
6. Student assigned for racial/ethnic composition (diversassg) [I added?]	BYSC24C (Pupils are assigned from particular areas to achieve desired racial or ethnic composition in the school. 1. Yes; 2. No; 8. Missing; 9. Legitimate skip/not in wave.)	Diversassg=1 if BYSC24C = 0. Diversassg=1 if BYSC24C = 2. Missing were coded into median. [Could not match original results]	
7. Student feels unsafe in school (unsafe)	BYS59K (I don't feel safe at this school. 1. Strongly agree; 2. Agree; 3. Disagree; 4. Strongly disagree; 6. Multiple response; 8. Missing; 9. Legitimate skip/not in wave.)	unsafe=1 if BYS59K is 1 or 2. unsafe=0 if BYS59K is 3 or 4. Missing were coded into median.	

Variable	Relevant variable(s) in NELS:88 dataset	Transformation required for precise replication	Modifications for replication
8. Disruptions in school prevent learning (disrupt)	BYS59L (Disruptions by other students get in the way of my learning. 1. Strongly agree; 2. Agree; 3. Disagree; 4. Strongly disagree; 6. Multiple response; 8. Missing; 9. Legitimate skip/not in wave.)	disrupt=1 if BYS59L is 1 or 2. Else, disrupt=0. (Not ideal since missing data was counted as 0, even when there were more students who "agreed" or "strongly agreed")	
9. Student-teacher ratio	BYRATIO (Student-teacher ratio. 10. 10 or fewer students per teacher; 11-29. Number of students per teacher; 30. 30 or more students per teacher.; 99=legitimate skip/not in wave)	Missing were recoded with mean, by sex.	
10. Number of Black, non-Hispanic teachers	BYSC20D (Number of Black, non-Hispanic teachers. 0. None; 1. 1; 2. 2; 3. 3-5; 4. 6-10; 5. 11-20; 6. 21 or more; 997. Refusal; 998. Missing; 999. Legitimate skip/not in wave.)	Missing were coded into median.	96. Don't know; 97. Refusal; 98. Missing; 99. Legitimate skip

APPENDIX B

VARIABLE DESCRIPTIONS FOR BYRNES AND MILLER 2007 RE-ANALYSIS

Variable	Relevant variable(s) in NELS:88 dataset	Recoding required for replication
Sample selection and weight		
1. Flags	F2TRP1FL, F2F1QFLG	Select only if F2TRP1FL=1 and F2F1QFLG=1, resulting in 15855 cases.
2. Weight	F2PNLWT	
Sample identification		
1. Student ID	STU_ID	
Outcome variables		
1. 12th grade math (mathach)	F22XMIRR	Set 99.98 and 99.99 to missing. Categorize to high vs low based on median. (For mathach, 6229 high achievers with scores greater than 50.11, and 6231 low achievers.)
2. 12th grade science	F22XSIRR (99.98 = "MISSING" 99.99 = "TEST NOT COMP")	
Distal factors		
1. 8th grade SES	BYSES (99.998 = "MISSING")	Set 99.998 to missing.
2. Parent expectations in 8th grade (Pexp)	BYP76 (01 = "LT H.S. DIPLOMA" 02 = "GED" 03 = "H.S. GRADUATION" 04 = "VOC, ETC. < 1YR" 05 = "VOC,ETC. 1-2 YRS" 06 = "VOC, ETC.2+ YRS" 07 = "LT 2YRS COLLEGE" 08 = "2+ YRS COLLEGE" 09 = "2YR COLLEGE PGM" 10 = "4-5YR COLLEG PGM" 11 = "MASTER'S DEGREE" 12 = "PH.D., M.D.,OTHR" 96 = "MULTIPLE RESPONSE" 97 = "REFUSAL" 98 = "MISSING" 99 = "LEGITIMATE SKIP")	Code 1-3 as Pexp_HS, 4-9 as Pexp_2yr, 10 as Pexp_BA, and 11-12 as Pexp_MADr, else missing.
3. Student expectations in 8th grade (Sexp)	BYS45 (01 = "WON'T FINISH H.S" 02 = "WILL FINISH H.S" 03 = "VOC,TRD,BUS AFTR H.S" 04 = "WILL ATTEND COLLEGE" 05 = "WILL FINISH COLLEGE" 06 = "HIGHER SCH AFTR COLL" 96 = "MULTIPLE RESPONSE" 97 = "REFUSAL" 98 = "MISSING" 99 = "LEGITIMATE SKIP")	Code 1-2 as Sexp_HS, 3-4 as Sexp_2yr, 5 as Sexp_BA, and 6 as Sexp_MADr, else missing.
4. Middle school GPA	BYGRADS (9.8 = "MISSING")	Set 9.8 as missing
Opportunity factors		
1. General math courses (gm_none, gm_half, gm_1, gm_2; reference is none)	F1S22A (0 = "NONE" 1 = "1/2 YEAR" 2 = "1 YEAR" 3 = "1 1/2 YEARS"	Indicator code so that 0 years is the reference category, and less than 1 year, 1 year, and over 1 year are other categories. Set 6-9 as missing.

Variable	Relevant variable(s) in NELS:88 dataset	Recoding required for replication
	4 = "2 YEARS" 6 = "MULTIPLE RESPNSE" 7 = "REFUSAL" 8 = "MISSING" 9 = "LEGITIMATE SKIP")	
2. Geometry courses (geo_none, geo_half, geo_1, geo_2; reference is none)	F1S22D (same categorization as F1S22A)	
3. Algebra II courses (al2_none, al2_half, al2_1, al2_2; reference is none)	F1S22E (same categorization as F1S22A)	
4. General science courses (gs_none, gs_half, gs_1, gs_2; reference is none)	F1S23A (same categorization as F1S22A)	
5. Biology courses (bio_none, bio_half, bio_1, bio_2; reference is none)	F1S23C (same categorization as F1S22A)	
6. Chemistry courses (chm_none, chm_half, chm_1, chm_2; reference is none)	F1S23E (same categorization as F1S22A)	
7. Student perception of math emphasis (emph_m)	F1S31A, B, C, D, E (0 = "NONE" 1 = "MINOR EMPHASIS" 2 = "MODERATE EMPHASIS" 3 = "MAJOR EMPHASIS" 6 = "MULTIPLE RESPNSE" 7 = "REFUSAL" 8 = "MISSING" 9 = "LEGITIMATE SKIP")	Set 6-9 as missing. 1 point each for indicating a major emphasis on B-D, and either a minor or moderate emphasis on A and E, for a maximum of 5 points.
8. Student perception of science emphasis (emph_sci)	F1S30A, B, C, D, E (same categorization as F1S31A)	Same as above.
9. Student perception of teacher responsiveness (t_rspnsv)	F1S7A (S gets along well with teachers), D (Discipline is fair at school), G (The teacher is good at school), H (Teachers are interested in students), I (When R works hard teachers praise effort), J (In class often feel put down by teachers), L (Most teachers listen to R) (1 = "STRONGLY AGREE" 2 = "AGREE" 3 = "DISAGREE" 4 = "STRONGLY DISAGREE" 6 = "MULTIPLE RESPNSE" 7 = "REFUSAL" 8 = "MISSING" 9 = "LEGITIMATE SKIP")	Recode so that 3 represents high teacher responsiveness and 0 is low. t_rspnsv = (.340584*F1S7A) + (.286039*F1S7D) + (.414373*F1S7G) + (.448220*F1S7H) + (.373865*F1S7I) + (-.337648*F1S7J) + (.419283*F1S7L); Coefficients from PCA using covariance matrix.
Propensity factors		
1. Math achievement before start of 9th grade	BYTXMIRR (999.998 = "MISSING" 999.999 = "LEGITIMATE SKIP")	Set 999.998 and 999.999 to missing.
2. Math GPA in 9th and 10th grades (GPA910_m)	F1S39A (01 = "NOT TAKING" 02 = "MOSTLY A'S" 03 = "HALF A & HALF B" 04 = "MOSTLY B'S" 05 = "HALF B & HALF C"	Set 1 and >9 to missing. Code 1 as 4, 2 as 3.5, 3 as 3, ... and 9 as .5.

Variable	Relevant variable(s) in NELS:88 dataset	Recoding required for replication
	06 = "MOSTLY C'S" 07 = "HALF C & HALF D" 08 = "MOSTLY D'S" 09 = "MOSTLY BELOW D" 10 = "CLASS NOT GRADED" 96 = "MULTIPLE RESPNSE" 97 = "REFUSAL" 98 = "MISSING" 99 = "LEGITIMATE SKIP")	
3. Science achievement before start of 9th grade	BYTXSIRR (999.998 = "MISSING" 999.999 = "LEGITIMATE SKIP")	Set 999.998 and 999.999 to missing.
4. Science GPA in 9th and 10th grades (GPA910_s)	F1S39D (same categorization as F1S39A)	Set 1 and >9 to missing. Code 1 as 4, 2 as 3.5, 3 as 3, ... and 9 as .5.
5. Efficacy for graduating high school (How sure student is about graduating) (grad_eff)	F1S18A (1 = "YES, SURE GRAD" 2 = "PROBABLY" 3 = "PROBABLY NOT" 4 = "NO/ SURE I WON'T" 6 = "MULTIPLE RESPNSE" 7 = "REFUSAL" 8 = "MISSING" 9 = "LEGITIMATE SKIP")	Set >4 to missing. Recode so that scale is 0 to 3, where 3 is student is sure of graduating.
6. Plans to take SAT (SATplan)	F1S50B (1 = "HAVEN'T THOUGHT" 2 = "DON'T PLAN" 3 = "YES, THIS YEAR" 4 = "YES, NEXT YEAR" 5 = "YES, 12TH GRADE" 6 = "MULTIPLE RESPNSE" 7 = "REFUSAL" 8 = "MISSING" 9 = "LEGITIMATE SKIP")	SATplan is 1 if 3-5, 0 if 1-2, and missing if >=6.
7. Math self-concept (m_selfcpt)	F1S63D (Math is one of R's best subjects), J (R has always done well in math), Q (R gets good marks in math), S (R does badly in tests in math) (01 = "FALSE" 02 = "MOSTLY FALSE" 03 = "FALSE THAN TRUE" 04 = "TRUE THAN FALSE" 05 = "MOSTLY TRUE" 06 = "TRUE" 96 = "MULTIPLE RESPNSE" 97 = "REFUSAL" 98 = "MISSING" 99 = "LEGITIMATE SKIP")	Recode so that lowest math concept is 0, and 24-31 is 5 for each item. m_selfcpt = (.557034*F1S63D_r) + (.511711*F1S63J_r) + (.517930*F1S63Q_r) + (.399517*F1S63S_r) Coefficients from PCA using covariance matrix.
Demographic factors		
1. Gender (female)	SEX, F1SEX, F2SEX (1 = "MALE" 2 = "FEMALE" 6 = "MULTIPLE RESPNSE" 7 = "REFUSAL" 8 = "MISSING" 9 = "LEGITIMATE SKIP")	Indicator code so that male is reference group.
2. Race/ethnicity (asian, hispanic, black, nativeam, white)	RACE, F1RACE, F2RACE1 (1 = "API" 2 = "HISPANIC" 3 = "BLACK,NON-HISPANIC"	Indicator code with White as reference group.

Variable	Relevant variable(s) in NELS:88 dataset	Recoding required for replication
	4 = “WHITE,NON-HISPANIC” 5 = “AMERICAN INDIAN” 6 = “MULTIPLE RESPONSE” 7 = “REFUSAL” 8 = “MISSING” 9 = “LEGITIMATE SKIP”)	

APPENDIX C

DETAILED RULESET MINING RESULTS—STUDY 1

Table 43. CBA ruleset (Study 1, 19 possible predictors)

Rule order	Antecedent	Consequent
1	par_ed=M.A./EQUIVALENT, activity=(1-4] med	High achieving
2	par_ed=M.A./EQUIVALENT, privsch=No, peerexcl=Yes	High achieving
3	par_ed=M.A./EQUIVALENT, pinvolve=(2-4 med]	High achieving
4	par_ed=M.A./EQUIVALENT, climate=<=1 best, peerexcl=Yes	High achieving
5	par_ed=M.A./EQUIVALENT, hhressc=gr6 hi	High achieving
6	par_ed=M.A./EQUIVALENT, peerexcl=Yes	High achieving
7	par_ed=M.A./EQUIVALENT, climate=<=1 best	High achieving
8	goodpeer=gr2.5 hi/4, BlkTeacher=NONE, disrupt=No	High achieving
9	hw_outsch=7-9 HOURS, sculture=gr3 hi, peerexcl=Yes, disrupt=No	High achieving
10	hw_outsch=7-9 HOURS, sculture=gr3 hi, pexpcol=Yes, disrupt=No	High achieving
11	badpeer=<=1.5 lo, BlkTeacher=NONE, disrupt=No	High achieving
12	hw_outsch=7-9 HOURS, sculture=gr3 hi, privsch=No, pexpcol=Yes, unsafe=No	High achieving
13	hw_outsch=7-9 HOURS, sculture=gr3 hi, disrupt=No	High achieving
14	badpeer=<=1.5 lo, BlkTeacher=NONE, peerexcl=Yes	High achieving
15	hw_outsch=7-9 HOURS, sculture=gr3 hi, pexpcol=Yes, unsafe=No	High achieving
16	income91=$35000-$49999, hhressc=(4-6 med], goodpeer=gr2.5 hi/4, pexpcol=Yes, unsafe=No	High achieving
17	hw_outsch=7-9 HOURS, privsch=No, peerexcl=Yes, disrupt=No	High achieving
18	hw_outsch=7-9 HOURS, peerexcl=Yes, disrupt=No	High achieving
19	goodpeer=gr2.5 hi/4, BlkTeacher=NONE, peerexcl=Yes	High achieving
20	badpeer=<=1.5 lo, sculture=gr3 hi, privsch=No, pexpcol=Yes, peerexcl=Yes, unsafe=No, disrupt=No	High achieving
21	activity=(1-4] med, sculture=gr3 hi, peerexcl=Yes, unsafe=No, disrupt=No	High achieving
22	goodpeer=gr2.5 hi/4, sculture=gr3 hi, religsch=No, pexpcol=Yes, peerexcl=Yes, unsafe=No, disrupt=No	High achieving
23	hhressc=gr6 hi, BlkTeacher=NONE	High achieving
24	income91=$35000-$49999, hhressc=(4-6 med], goodpeer=gr2.5 hi/4, unsafe=No	High achieving
25	pinvolve=(2-4 med], badpeer=<=1.5 lo, sculture=gr3 hi, pexpcol=Yes, unsafe=No, disrupt=No	High achieving
26	activity=(1-4] med, sculture=gr3 hi, religsch=No, pexpcol=Yes, unsafe=No, disrupt=No	High achieving
27	goodpeer=gr2.5 hi/4, sculture=gr3 hi, privsch=No, pexpcol=Yes, unsafe=No, disrupt=No	High achieving
28	income91=$35000-$49999, climate=(2-5] med/5, unsafe=No	High achieving
29	hw_sch=1-3 HOURS, goodpeer=gr2.5 hi/4, sculture=gr3 hi, unsafe=No	High achieving
30	activity=(1-4] med, religsch=No, pexpcol=Yes, peerexcl=Yes, unsafe=No, disrupt=No	High achieving
31	activity=(1-4] med, badpeer=<=1.5 lo, sculture=gr3 hi, religsch=No, pexpcol=Yes, unsafe=No	High achieving

Table 43 continued

32	hhressc=(4-6 med], goodpeer=gr2.5 hi/4, badpeer=<=1.5 lo, sculture=gr3 hi, pexpcol=Yes, unsafe=No	High achieving
33	BlkTeacher=NONE, religsch=No, peerexcl=Yes, unsafe=No	High achieving
34	pinvolve=gr4 hi, privsch=No, peerexcl=Yes, disrupt=No	High achieving
35	BlkTeacher=3 - 5, privsch=No, pexpcol=Yes, peerexcl=Yes, unsafe=No, disrupt=No	High achieving
36	sculture=gr3 hi, religsch=No, pexpcol=Yes, peerexcl=Yes, disrupt=No	High achieving
37	hw_outsch=7-9 HOURS, privsch=No, pexpcol=Yes, unsafe=No, disrupt=No	High achieving
38	badpeer=(1.5-2] med, sculture=gr3 hi, unsafe=No, disrupt=No	High achieving
39	hhressc=(4-6 med], activity=(1-4] med, sculture=gr3 hi, pexpcol=Yes, unsafe=No	High achieving
40	par_ed=COLLEGE GRADUATE, goodpeer=gr2.5 hi/4, peerexcl=Yes	High achieving
41	hw_outsch=7-9 HOURS, badpeer=<=1.5 lo, pexpcol=Yes, unsafe=No, disrupt=No	High achieving
42	hw_sch=1-3 HOURS, sculture=gr3 hi, privsch=No, unsafe=No, disrupt=No	High achieving
43	par_ed=COLLEGE GRADUATE, peerexcl=Yes	High achieving
44	activity=(1-4] med, goodpeer=gr2.5 hi/4, pexpcol=Yes, peerexcl=Yes, unsafe=No, disrupt=No	High achieving
45	activity=(1-4] med, goodpeer=gr2.5 hi/4, pexpcol=Yes, unsafe=No, disrupt=No	High achieving
46	goodpeer=gr2.5 hi/4, G8LUNCH=NONE, peerexcl=Yes, unsafe=No	High achieving
47	hw_sch=4-6 HOURS, climate=(2-5] med/5, religsch=No, peerexcl=Yes	High achieving
48	G8LUNCH=NONE, climate=<=1 best	High achieving
49	hw_outsch=10-12 HOURS, hhressc=(4-6 med], badpeer=<=1.5 lo, pexpcol=Yes	High achieving
50	hw_sch=1-3 HOURS, hhressc=gr6 hi, religsch=No, disrupt=No	High achieving
51	hhressc=(4-6 med], pinvolve=(2-4 med], badpeer=<=1.5 lo, pexpcol=Yes, unsafe=No, disrupt=No	High achieving
52	BlkTeacher=3 - 5, religsch=No, pexpcol=Yes, unsafe=No, disrupt=No	High achieving
53	hhressc=gr6 hi, G8LUNCH=NONE	High achieving
54	hw_outsch=10-12 HOURS, hhressc=(4-6 med], goodpeer=gr2.5 hi/4, pexpcol=Yes	High achieving
55	hhressc=(4-6 med], activity=(1-4] med, badpeer=<=1.5 lo, pexpcol=Yes, unsafe=No	High achieving
56	badpeer=<=1.5 lo, sculture=gr3 hi, privsch=No, pexpcol=Yes, peerexcl=Yes, unsafe=No	High achieving
57	goodpeer=gr2.5 hi/4, badpeer=<=1.5 lo, climate=(1-2] 2nd best/5, pexpcol=Yes, unsafe=No	High achieving
58	hhressc=(4-6 med], goodpeer=gr2.5 hi/4, climate=(1-2] 2nd best/5, pexpcol=Yes, unsafe=No	High achieving
59	pinvolve=(2-4 med], badpeer=<=1.5 lo, pexpcol=Yes, peerexcl=Yes, unsafe=No, disrupt=No	High achieving
60	pinvolve=(2-4 med], badpeer=<=1.5 lo, pexpcol=Yes, unsafe=No, disrupt=No	High achieving
61	hhressc=(4-6 med], activity=(1-4] med, badpeer=<=1.5 lo, pexpcol=Yes	High achieving
62	hhressc=(4-6 med], goodpeer=gr2.5 hi/4, badpeer=<=1.5 lo, peerexcl=Yes, unsafe=No, disrupt=No	High achieving
63	hhressc=(4-6 med], BlkTeacher=3 - 5, pexpcol=Yes, unsafe=No	High achieving
64	hw_outsch=10-12 HOURS, hhressc=(4-6 med], pexpcol=Yes	High achieving
65	goodpeer=gr2.5 hi/4, BlkTeacher=3 - 5, privsch=No, disrupt=No	High achieving
66	hw_outsch=4-6 HOURS, hhressc=(4-6 med], goodpeer=gr2.5 hi/4, badpeer=<=1.5 lo, pexpcol=Yes, unsafe=No	High achieving
67	If none of the rules apply	Not high achieving

Table 44. RIPPER ruleset (Study 1, 19 possible predictors)

Rule order	Antecedent	Consequent	Coverage	Confidence
1	(peerexcl = Yes) and (disrupt = No) and (pinvolve >= 3)	High achieving	.11/.12	.50/.56
2	(sculture >= 4) and (par_ed = M.A./EQUIVALENT) and (activity >= 2)	High achieving	.02	.76/.26
3	(pexpcol = Yes) and (BlkTeacher = NONE)	High achieving	.06/.05	.48/.28
4	(income91 = $35,000-$49,999) and (goodpeer <= 2.4)	High achieving	.03/.04	.40/.14
5	(unsafe = No) and (par_ed = > HS & < 4YR DEG) and (hhressc >= 5)	High achieving	.19	.30/24
6	(hw_sch = 16-20 HOURS) and (activity >= 1.01178)	High achieving	.008	.86/1
7	If none of the rules apply	Not high achieving	.58/.59	.88/.81

Table 45. RIPPER ruleset (Study 1, 1372 possible predictors)

Rule order	Antecedent	Consequent	Coverage	Confidence
1	(BYSES >= -0.331) and (BYT2_2.ALL = HIGHER LEVELS) and (BYS55A = NEVER)	High achieving	.07	.87/.67
2	(BYP85F = FALSE) and (BYPSEPLN = HIGHER SCH AFTR COLL) and (BYSES >= -0.753)	High achieving	.10/.12	.71/.49
3	(BYLOCUS1 >= -0.04) and (autonomy >= 3.5) and (BYS42B = OVER 5 HRS A DAY) and (BYP73 = HOME) and (BYP81 = ONE)	High achieving	.02/.04	.79/.47
4	(BYS74 = NO) and (BYS44M = DISAGREE) and (BYS52 = PRO,BUSINSS,MGRL)	High achieving	.04	.51/.23
5	(BYS58K = MODERATE) and (G8URBAN = SUBURBAN) and (BYS83G = DID NOT PARTICIPATE)	High achieving	.02/.01	.75/.50
6	(goodpeer <= 2.4) and (BYS59G = AGREE) and (BYP67 = OCCASIONALLY) and (BYS62 = YES) and (autonomy >= 3.25)	High achieving	.01/.008	.78/0
7	(BYP7 = missing) and (hhressc >= 6) and (BYSC47K = VERY MUCH ACCURATE)	High achieving	.01/.02	.90/0
8	(BYFAMINC = $25,000-$34,999) and (BYP62B3 = YES) and (BYS44E = AGREE) and (BYS55D = NEVER)	High achieving	.01/.33	.70/.008
9	(BYT310A2.ALL = missing) and (BYS58D = MODERATE) and (BYS50E = ONCE OR TWICE) and (BYS51DB = YES)	High achieving	.01/.02	.60/.25
10	(BYSC41F = YES) and (BYS44C = STRONGLY DISAGREE) and (BYS70A = DISAGREE)	High achieving	.007/.008	.83/.33

Table 45 continued

11	(hw_outsch = NONE) and (BYS60A = HIGH)	High achieving	.008/.01	.57/0
12	(BYS80 = 6 HRS OR MORE PER WK) and (BYP14 = missing)	High achieving	.005/0	1/0
13	(BYT3_30B.MATH = FOUR HOURS) and (BYT2_9C.MATH = RARELY USED)	High achieving	.008/.003	.57/0
14	(BYS47 = PROBABLY WON'T) and (goodpeer <= 2)	High achieving	.006/.003	.60/0
15	(BYT2_24B.ALL = REVIEW TOPIC ONLY) and (BYP74A = STRONGLY AGREE)	High achieving	.01	.33/.02
16	(BYT3_4.MATH = 13 - 15 YEARS) and (BYT3_3Y.ALL = 1951 - 1955)	High achieving	.005/.02	.50/.33
17	(BYP30 = MA+) and (G8LUNCH = 76-100%)	High achieving	.001/0	1/0
18	If none of the rules apply	Not high achieving	.65/.61	1/.85

Table 46. PART ruleset (Study 1, 19 possible predictors)

Rule order	Antecedent	Ach	Coverage	Confidence
1	unsafe = missing AND activity <= 1.030055	Not high	.05	.95/.89
2	pexpcol = No AND par_ed = H.S. GRAD OR GED AND badpeer <= 1.596122	Not high	.06	1/.85
3	par_ed = DIDN'T FINISH HS	Not high	.12	.95/.93
4	unsafe = Yes	Not high	.11	.87/.83
5	peerexcl = Yes AND hw_outsch = 10-12 HOURS AND BlkTeacher = 6 - 10	High	.01/.008	.66/.33
6	peerexcl = Yes AND disrupt = No AND hw_outsch = 7-9 HOURS AND income91 = missing	High	.02	.93/.67
7	peerexcl = missing AND disrupt = No	Not high	.03	.95/1
8	par_ed = COLLEGE GRADUATE	High	.05/.07	.47/.53
9	par_ed = M.A./EQUIVALENT	High	.05	.58/.55
10	religsch = Yes AND income91 = $35,000-$49,999	High	.006/.003	.8/1
11	par_ed = > HS & < 4YR DEG AND G8LUNCH = 11-20%	High	.05/.04	.36/.31
12	BlkTeacher = NONE	High	.05/.04	.51/.33
13	income91 = $5,000-$7,499	Not high	.03/.01	.91/.75
14	income91 = $25,000-$34,999 AND hw_sch = 4-6 HOURS	High	.02/.03	.31/.10
15	income91 = $25,000-$34,999	High	.03/.02	.29/.38
16	income91 = $7,500-$9,999	Not high	.03/.02	.86/.75
17	income91 = $10,000-$14,999 AND peerexcl = No	Not high	.02	.94/1
18	G8LUNCH = NONE	High	.01/.02	.67/.43
19	hw_sch = missing	Not high	.02/.03	.95/.90
20	income91 = $20,000-$24,999	High	.04	.33/.13
21	BlkTeacher = missing	High	.007/.005	.67/0

Table 46 continued

22	income91 = $35,000-$49,999 AND activity <= 0	Not high	.01	1/.75
23	income91 = $35,000-$49,999 AND activity <= 1.030055 AND hw_sch = 1-3 HOURS	High	.003/.008	1/0
24	income91 = $35,000-$49,999	High	.02	.44/.33
25	hhressc <= 6 AND income91 = $15,000-$19,999	High	.02/.04	.40/.38
26	hhressc <= 6 AND G8LUNCH = 51-75%	Not high	.03/.04	.84/.57
27	hhressc > 6	Not high	.02/.04	1/.69
28	climate <= 6 AND income91 = $10,000-$14,999 AND G8LUNCH = 31-50%	High	.006/.01	.40/.25
29	climate <= 6 AND income91 = missing AND G8LUNCH = 31-50%	High	.01/.008	.25/.33
30	climate <= 6 AND income91 = missing	High	.02/.01	.46/.25
31	If none of the rules apply	Not high	.03/.02	.75/.78

Table 47. PART ruleset (Study 1, 1372 possible predictors)

Rule order	Antecedent	Ach	Coverage	Confidence
1	BYSES > -0.208 AND BYS76 [cutting class] = NEVER/ALMOST NEVER AND BYP45A [reason for being held back] = missing AND BYSC38C [holds 8th grader back for failing science test] = NO AND BYP85F [believes test scores won't qualify S for fin aid] = FALSE AND BYS74B [ever repeated gr 1] = missing AND BYS57B [someone offered to sell S drugs at school] = NEVER AND BYS46 [confidence in graduating HS] = VERY SURE WILL AND BYT3_14B.SOC.STUDIES.HISTORY [proficient in German if proficient in non-English language] = missing AND BYS67A [remedial math >=1/wk] = DO NOT ATTEND	High	.12	.76/.53
2	BYP45C [held back because of other reason] = missing AND BYLOCUS1 [locus of control] <= -0.055065 AND BYT3_25B.MATH [satisfaction with content/curric of gifted & talented program if teaching such program] = missing AND religsch = No	Not high	.28/.24	.93/.87
3	BYP45C [held back for other reason, if held back] = NO	Not high	.07/.10	.97
4	BYP45C [held back for other reason, if held back] = YES	Not high	.05/.04	.94/.93
5	BYT3_25E.ENGLISH [satisfaction with selection procedures for gifted and talented program, if teaches such program] = missing AND BYS68B [enrolled in bilingual ed] = NO AND BYT3_16D.ENGLISH [how well T writes in non-English language if proficient in at least one] = missing AND	Not high	.04/.02	.94/1

Table 47 continued

	BYT3_17E.SOC.STUDIES.HISTORY [learned non-English language informally, if proficient in at least one] = missing AND BYP52A [how important child complete sch faster] = missing AND BYP36A [spouse's current work status] = missing AND BYP45A [held back due to parent request, if held back at all] = missing AND BYS35O [R's family has a pocket calculator] = HAVE AND BYS27C [how well R reads English] = missing AND BYT2_15.ALL [# of hrs class meets per week] > 3 AND BYS51GA [talk to counselor about drug/alc abuse] = YES			
6	BYS76 [how often do you cut or skip class] = NEVER/ALMOST NEVER AND BYT3_17E.SOC.STUDIES.HISTORY [learned non-English language informally, if proficient in one] = missing AND BYP52A [how important child complete sch faster] = missing AND BYT3_16D.ENGLISH [how well T writes non-English, if proficient in one] = missing AND BYT310B1.SOC.STUDIES.HISTORY [graduate degree in English] = NO AND BYS39A [parents trust R to do what they expect] = TRUE	Not high	.03	1/.9
7	BYS76 [how often do you cut or skip class] = NEVER/ALMOST NEVER AND BYT3_17E.SOC.STUDIES.HISTORY [learned non-English language informally, if proficient in one] = missing AND G10COHRT [enrolled in sch in 10th grade (erroneously included in dataset)] = SPRING MEMBER AND BYP52A [how important child complete sch faster] = missing AND BYT3_16D.ENGLISH [how well T writes non-English, if proficient in one] = missing AND BYT3_16B.ALL [how well T speaks non-English, if proficient in one] = missing AND autonomy > 4.5 AND BYP38B [did 8th grader attend nursery/pre-school] = missing	High	.01/.008	.73/1
8	BYT3_22.ALL <= 50 AND BYT3_16D.ENGLISH = missing AND BYS68B = NO AND BYT3_17E.SOC.STUDIES.HISTORY = missing AND BYT3_16D.ALL = missing AND G12COHRT = SPRING MEMBER AND BYT3_23B.MATH = missing AND BYP33A = missing AND BYP52C = missing AND BYT3_33.SOC.STUDIES.HISTORY = missing AND BYS82U = DID NOT PARTICIPATE AND	High	.03/.04	.65/.31

Table 47 continued

	BYT3_30D.SOC.STUDIES.HISTORY = missing AND BYP84B = missing AND BYCNCPT2 > -0.18 AND BYT2_16C.ALL = LESS THAN ONE HR AND BYSC48B = YES			
9	BYT3_22.ALL <= 50 AND BYT3_16D.ENGLISH = missing AND BYSES <= -1.243	Not high	.05/.04	.91/.79
10	BYS55B = NEVER AND G10COHRT = SPRING MEMBER AND BYP52A = NOT VERY IMPORTANT	High	.009/.02	.75/.57
11	BYS55B = NEVER AND G10COHRT = NOT A MEMBER	Not high	.02	.84/1
12	BYS55B = NEVER AND BYS51EA = NO AND BYT3_25B.MATH = missing AND BYSC38D = NO AND BYS51HA = NO AND BYT3_17E.ENGLISH = missing AND BYS76 = NEVER/ALMOST NEVER AND BYS83C = missing AND BYSC16F <= 4	High	.005	1/.5
13	BYS55B = NEVER AND BYS8C = YES AND BYT3_16D.ENGLISH = missing AND BYS67CD = DO NOT ATTEND AND BYS67B = missing	High	.02	.71/.29
14	BYS51EA = NO AND BYS55B = NEVER AND BYS8C = YES AND BYT3_16D.ENGLISH = missing AND BYT2_6.SOC.STUDIES.HISTORY <= 0.358065 AND BYS76 = NEVER/ALMOST NEVER AND BYP36A = missing AND BYT2_7M.ENGLISH <= 20	High	.11/.10	.34/.39
15	BYT3_22.SOC.STUDIES.HISTORY <= 33 AND BYSC50BD = EXPULSION	Not high	.13/.005	.85/1
16	If none of the rules apply	High	.04/.20	.47/.26

Table 48. C4.5 ruleset (Study 1, 19 possible predictors)

Rule	Antecedent	Consequent	Coverage	Confidence
1	par_ed = > HS & < 4YR DEG, unsafe = {missing, Yes}	Not high achieving	.09/.08	.95/.79
2	par_ed = > HS & < 4YR DEG, unsafe = No	High achieving	.37/.35	.31/.22
3	par_ed = {COLLEGE GRADUATE, M.A./EQUIVALENT}	High achieving	.14/.15	.54/.49
4	par_ed = {DIDN'T FINISH HS, H.S. GRAD OR GED}	Not high achieving	.36/.38	.89/.83
5	par_ed = PH.D. M.D. OTHER, peerexcl = No	High achieving	.005/.008	1/.67

Table 48 continued

6	par_ed = PH.D. M.D. OTHER, peerexcl = Missing	Not high achieving	.002/0	1/NA
7	par_ed = PH.D. M.D. OTHER, peerexcl = Yes, BlkTeacher = {0, 1, 2, 3-5, 6-10, 11-20, missing}	High achieving	.007/.01	.83/1
8	par_ed = PH.D. M.D. OTHER, peerexcl = Yes, BlkTeacher = {21 or more}	Not high achieving	.002/0	1/NA

Table 49. C4.5 ruleset (Study 1, 1372 possible predictors)

Rule	Antecedent	Consequent	Coverage	Confidence
1	BYP45A = missing, BYSES <= -0.216	Not high achieving	.56/.50	.83/82
2	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = missing	Not high achieving	.007/.008	.83/1
3	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = MORE THAN TWICE	High achieving	0/.005	NA/0
4	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {AT LEAST ONCE A WEEK, DAILY}	High achieving	.002/.003	0
5	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {missing, LESS THAN ONCE A WK}	Not high achieving	.02/.03	.88/.64
6	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = missing	High achieving	.005/.01	.75/.25
7	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = missing, BYS26B = {1/2 THE TIME, ALWAYS/MOST TIME}	High achieving	.005/.003	.75/0
8	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = missing, BYS26B = missing, BYS74B = missing, BYP85F = {FALSE, missing}	High achieving	.16	.69/.53
9	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = missing, BYS26B = missing, BYS74B =	Not high achieving	.009/.02	1/.63

Table 49 continued

	missing, BYP85F = HVN'T THGHT ABT YET			
10	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = missing, BYS26B = missing, BYS74B = missing, BYP85F = TRUE, BYS51AA = {missing, YES}	High achieving	.005	.75/0
11	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = missing, BYS26B = missing, BYS74B = missing, BYP85F = TRUE, BYS51AA = NO	Not high achieving	.007/.005	1/.5
12	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = missing, BYS26B = missing, BYS74B = NO	High achieving	.006/.005	.4/0
13	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = missing, BYS26B = missing, BYS74B = YES	Not high achieving	.002/.003	.5/1
14	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = missing, BYS26B = NEVER	Not high achieving	.002/.005	1/0
15	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = missing, BYS26B = SOMETIMES	High achieving	.008/.002	1/0
16	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = {NOT VERY WELL, VERY WELL}	High achieving	.003	.67/1
17	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST	Not high achieving	.006/.005	.8/0

Table 49 continued

	NEVER}, BYS51GA = NO, BYT3_16B.SOC.STUDIES.HISTORY = {PRETTY WELL, WELL}			
18	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = NEVER, BYS76 = {NEVER/ALMOST NEVER}, BYS51GA = YES	Not high achieving	.02	.84/.71
19	BYP45A = missing, BYSES > -0.216, BYSC15 = 10% OR LESS, BYS55C = ONCE OR TWICE	Not high achieving	.008/.01	.86/.5
20	BYP45A = missing, BYSES > -0.216, BYSC15 = {11 - 20%, 21 - 30%, 41 - 50%}	Not high achieving	.03/.04	.81/.5
21	BYP45A = missing, BYSES > -0.216, BYSC15 = {31 - 40%, 61 - 70%, 81 OR MORE, missing}	High achieving	.007/.003	.5/0
22	BYP45A = NO	Not high achieving	.08/.10	.95/.97
23	BYP45A = YES	Not high achieving	.05	.96/.84

Table 50. CART ruleset (Study 1, 19 possible predictors)

Rule	Antecedent	Consequent	Coverage	Confidence
1	par_ed=COLLEGE GRADUATE, M.A./EQUIVALENT, PH.D., M.D., OTHER	High achieving	.14/.15	.54/.49
2	par_ed=DIDN'T FINISH HS, H.S. GRAD OR GED, HS & < 4YR DEG	Not high achieving	.86/.83	.79/.78

Note: For ease of interpretation, surrogate branches for missing data are excluded from this table.

Table 51. CART ruleset (Study 1, 1372 possible predictors)

Rule	Antecedent	Consequent	Coverage	Confidence
1	par_ed= MASTER'S DEGREE, PH.D., M.D.,OTHR	High achieving	.11/.13	.53/.32
2	par_ed= LT HS GRAD, HS GRAD, VOC,ETC. LT 2YR, VOC,ETC.2+ YRS, LT 2YRS COLLEGE, 2YR+ COLLEGE, 4-5YR COLLEG PGM	Not high achieving	.67/.69	.85/.82

Note: Surrogate branches for missing data are excluded from this table.

Table 52. C5.0 ruleset (Study 1, 19 possible predictors)

Rule	Antecedent	Consequent	Coverage	Confidence
1	par_ed in [COLLEGE GRADUATE-PH.D., M.D., OTHER]	High achieving	.16/.17	.55/.53
2	par_ed in [DIDN'T FINISH HS], activity <= 3	Not high achieving	.12/.13	.96/.91

Table 52 continued

3	par_ed in [DIDN'T FINISH HS], activity > 3	High achieving	0	NA
4	par_ed in [H.S. GRAD OR GED-> HS & < 4YR DEG], pexpcol = No, religsch = No	Not high achieving	.21/.22	.88/.87
5	par_ed in [H.S. GRAD OR GED-> HS & < 4YR DEG], pexpcol = No, religsch = Yes	High achieving	.004/0	1/NA
6	par_ed in [H.S. GRAD OR GED-> HS & < 4YR DEG], pexpcol = Yes, unsafe = Yes, goodpeer > 2.2	Not high achieving	.05/.03	1/.75
7	par_ed in [H.S. GRAD OR GED-> HS & < 4YR DEG], pexpcol = Yes, unsafe = Yes, goodpeer <= 2.2	High achieving	.01/.02	.27/.33
8	par_ed in [H.S. GRAD OR GED-> HS & < 4YR DEG], pexpcol = Yes, unsafe = No, BlkTeacher in [NONE-6 - 10]	High achieving	.22/.21	.34/.35
9	par_ed in [H.S. GRAD OR GED-> HS & < 4YR DEG], pexpcol = Yes, unsafe = No, BlkTeacher in [11 – 20-21 OR MORE], sculture <= 1	Not high achieving	.01/.08	1
10	par_ed in [H.S. GRAD OR GED-> HS & < 4YR DEG], pexpcol = Yes, unsafe = No, BlkTeacher in [11 – 20-21 OR MORE], sculture > 1, activity <= 0	High achieving	.01/.02	.55/.14
11	par_ed in [H.S. GRAD OR GED-> HS & < 4YR DEG], pexpcol = Yes, unsafe = No, BlkTeacher in [11 – 20-21 OR MORE], sculture > 1, activity >0, climate <= 3	Not high achieving	.01/.008	1/1
12	par_ed in [H.S. GRAD OR GED-> HS & < 4YR DEG], pexpcol = Yes, unsafe = No, BlkTeacher in [11 – 20-21 OR MORE], sculture > 1, activity >0, climate > 3	High achieving	.02	.27/.5

Table 53. C5.0 ruleset (Study 1, 1372 possible predictors)

Rule	Antecedent	Consequent	Coverage	Confidence
1	BYSES in [(-2.23,-1.22]-(-1.22,-0.214]], BYLOCUS1 in [(-2.67,-1.69]-(-1.69,-0.715]], G8CTRL in {PUBLIC SCHOOL,PRIVATE,OTH RELIG, PRIVATE, NO RELIG}	Not high achieving	.13/.12	.98/.96
2	BYSES in [(-2.23,-1.22]-(-1.22,-0.214]], BYLOCUS1 in [(-2.67,-1.69]-(-1.69,-0.715]], G8CTRL = CATHOLIC SCHOOL	High achieving	.004/0	.33/NA
3	BYSES in [(-2.23,-1.22]-(-1.22,-0.214]], BYLOCUS1 in [(-0.715,0.263]-(0.263,1.24]], BYP76 in {LT HS GRAD,HS GRAD,VOC, ETC. LT 2YR,VOC, ETC.2+ YRS, LT 2YRS COLLEGE,2YR+ COLLEGE, 4-5YR COLLEG PGM}	Not high achieving	.39/.37	.88/.84
4	BYSES in [(-2.23,-1.22]-(-1.22,-0.214]], BYLOCUS1 in [(-0.715,0.263]-(0.263,1.24]], BYP76 in { MASTER'S DEGREE,PH.D., M.D.,OTHR }	High achieving	.08/.10	.46/.29

Table 53 continued

5	BYSES in [(-0.214,0.792]-(0.792,1.8]], BYS51GA = YES, BYSC39M = ONE-HALF YEAR	Not high achieving	0	NA
6	BYSES in [(-0.214,0.792]-(0.792,1.8]], BYS51GA = YES, BYSC39M = LESS THN 1/2 YR	High achieving	.002/.003	1/0
7	BYSES in [(-0.214,0.792]-(0.792,1.8]], BYS51GA = YES, BYSC39M = {NO SPECIFIC AMT,FULL YEAR}, BYT3_31.ALL in [NONE-5-9 KID'S PARENTS]	High achieving	.004/.008	.67/0
8	BYSES in [(-0.214,0.792]-(0.792,1.8]], BYS51GA = YES, BYSC39M = {NO SPECIFIC AMT,FULL YEAR}, BYT3_31.ALL in [10-19 KID'S PARENTS-60+ KID'S PARENTS]	Not high achieving	.04	1/.71
9	BYSES in [(-0.214,0.792]-(0.792,1.8]], BYS51GA = NO, BYS55A = MORE THAN TWICE, BYSES = (-0.214,0.792]	Not high achieving	.02/.008	1
10	BYSES in [(-0.214,0.792]-(0.792,1.8]], BYS51GA = NO, BYS55A = MORE THAN TWICE, BYSES = (0.792,1.8]	High achieving	.001/0	1/NA
11	BYSES in [(-0.214,0.792]-(0.792,1.8]], BYS51GA = NO, BYS55A in [NEVER-ONCE OR TWICE], BYP85F in {TRUE,FALSE}	High achieving	.21/.22	.59/.50
12	BYSES in [(-0.214,0.792]-(0.792,1.8]], BYS51GA = NO, BYS55A in [NEVER-ONCE OR TWICE], BYP85F = HVN'T THGHT ABT YET, BYSES = (0.792,1.8]	High achieving	.001/0	1/0
13	BYSES in [(-0.214,0.792]-(0.792,1.8]], BYS51GA = NO, BYS55A in [NEVER-ONCE OR TWICE], BYP85F = HVN'T THGHT ABT YET, BYSES = (-0.214,0.792], G12CTRL1 in {PUBLIC,PRIV/OTH RELIG, PRIV/NOT ASCRTND}	Not high achieving	.01/.03	1/.67
14	BYSES in [(-0.214,0.792]-(0.792,1.8]], BYS51GA = NO, BYS55A in [NEVER-ONCE OR TWICE], BYP85F = HVN'T THGHT ABT YET, BYSES = (-0.214,0.792], G12CTRL1 in {CATHOLIC,PRIV/NON-RELIG}	High achieving	0	NA

Table 54. QUEST ruleset (Study 1, 19 possible predictors)

Rule	Antecedent	Consequent	Coverage	Confidence
1	IF (par_ed <= “H.S. GRAD OR GED”)	Not high achieving	.36/.38	.89/.83
2	IF (par_ed > “H.S. GRAD OR GED”) AND (par_ed <= “COLLEGE GRADUATE”) AND (BlkTeacher <= “NONE”)	High achieving	.06/.05	.44/.39
3	IF AND (par_ed > “H.S. GRAD OR GED”) AND (par_ed <= “COLLEGE GRADUATE”) AND (BlkTeacher IS MISSING OR	Not high achieving	.20	.80/.82

Table 54 continued

	(BlkTeacher > "NONE")) AND (income91 <= "$15,000-$19,999")			
4	IF (par_ed > "H.S. GRAD OR GED") AND (par_ed <= "COLLEGE GRADUATE") AND (BlkTeacher IS MISSING OR (BlkTeacher > "NONE") AND (income91 > "$15,000-$19,999")	High achieving	.29/.28	.31/.29
5	IF (par_ed > "COLLEGE GRADUATE")	High achieving	.08/.09	.64/.56

Note: For ease of interpretation, surrogate branches for missing data are excluded from this table.

Table 55. QUEST ruleset (Study 1, 1372 possible predictors)

Rule	Antecedent	Consequent	Coverage	Confidence
1	IF (BYSES > .523)	High achieving	.11/.13	.61/.54
2	IF (BYSES <= .523) AND (BUT2_2.ALL=HIGHER LEVELS)	High achieving	.15/.16	.49/.52
3	IF (BYSES <= .523) AND (BUT2_2.ALL=AVERAGE LEVELS; LOWER LEVELS; WIDELY DIFFERING) AND (BYP76 = LT 2YRS; 4-5YR COLLEG PRG; VOC, ETC, 2+ YRS; 2YR+ COLLEGE; HSGRAD; MASTER'S DEGREE; VOC,ETC. LT 2YR; LT HS GRAD) AND (BYS80 > 5.0)	Not high achieving	.03	.42/.33
4	IF (BYSES <= .523) AND (BUT2_2.ALL=AVERAGE LEVELS; LOWER LEVELS; WIDELY DIFFERING) AND (BYP76 = LT 2YRS; 4-5YR COLLEG PRG; VOC, ETC, 2+ YRS; 2YR+ COLLEGE; HSGRAD; MASTER'S DEGREE; VOC,ETC. LT 2YR; LT HS GRAD) AND (BYS80 <= 5.0) AND (G8CTRL2 = PRIV., RELIG; PRIV, NON-RELIG)	High achieving	.05	.30/.41
5	IF (BYSES <= .523) AND (BUT2_2.ALL=AVERAGE LEVELS; LOWER LEVELS; WIDELY DIFFERING) AND (BYP76 = LT 2YRS; 4-5YR COLLEG PRG; VOC, ETC, 2+ YRS; 2YR+ COLLEGE; HSGRAD; MASTER'S DEGREE; VOC,ETC. LT 2YR; LT HS GRAD) AND (BYS80 <= 5.0) AND (G8CTRL2 = PUBLIC)	Not high achieving	.60/.58	.90/.92
6	IF (BYSES <= .523) AND (BUT2_2.ALL=AVERAGE LEVELS; LOWER LEVELS; WIDELY DIFFERING) AND (BYP76 =PH.D., M.D., OTHR) AND (BYT3_25C.ENGLISH<=2)	High achieving	.006/.01	.40/.60
7	IF (BYSES <= .523) AND (BUT2_2.ALL=AVERAGE LEVELS; LOWER LEVELS; WIDELY DIFFERING) AND (BYP76 =PH.D., M.D., OTHR) AND	Not high achieving	.03/.01	.92/.50

Table 55 continued

	(BYT3_25C.ENGLISH>2) AND (BYSC38B=YES)			
8	IF (BYSES <= .523) AND (BUT2_2.ALL=AVERAGE LEVELS; LOWER LEVELS; WIDELY DIFFERING) AND (BYP76 =PH.D., M.D., OTHR) AND (BYT3_25C.ENGLISH>2) AND (BYSC38B=NO)	High achieving	.04/.03	.65/.18

Note: For ease of interpretation, surrogate branches for missing data are excluded from this table.

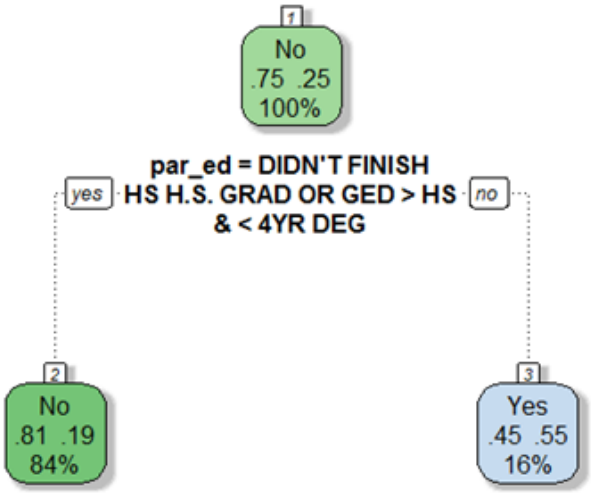


Figure 28. CART tree (Study 1, 19 possible predictors; results shown on training data)

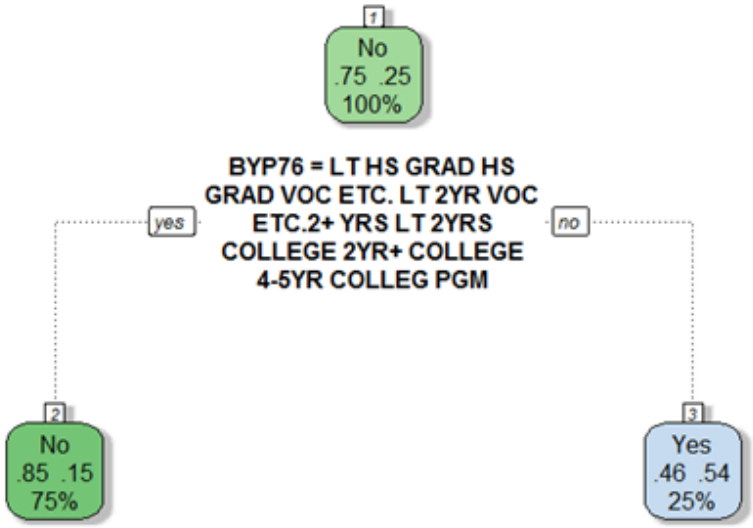


Figure 29. CART tree (Study 1, 1372 possible predictors; results shown on training data)

```
par_ed in [COLLEGE GRADUATE-PH.D., M.D., OTHER]: Yes (56.2/20.8)
par_ed in [DIDN'T FINISH HS-> HS & < 4YR DEG]:
:...par_ed = DIDN'T FINISH HS:
    :...activity <= 3: No (61.2/0.4)
    :   activity > 3: Yes (1.1/0.1)
    par_ed in [H.S. GRAD OR GED-> HS & < 4YR DEG]:
    :...pexpcol = No:
        :...religsch = No: No (92.4/7)
        :   religsch = Yes: Yes (2)
        pexpcol = Yes:
        :...unsafe = missing: Yes (0)
            unsafe = Yes:
            :...goodpeer <= 2.2: Yes (6.8/4.8)
            :   goodpeer > 2.2: No (19.8)
            unsafe = No:
            :...BlkTeacher in [NONE-6 - 10]: Yes (106.5/65.1)
                BlkTeacher in [11 - 20-21 OR MORE]:
                :...sculture <= 1: No (22.1/0.4)
                    sculture > 1:
                    :...activity <= 0: Yes (25.3/15.4)
                        activity > 0:
                        :...climate <= 3: No (19.1)
                            climate > 3: Yes (15.6/11.2)
```

Figure 30. C5.0 tree (Study 1, 19 possible predictors; results shown on training data)

```
BYSES in [(-2.23,-1.22]-(-1.22,-0.214]]:
:...BYLOCUS1 in [(-2.67,-1.69]-(-1.69,-0.715]]:
:   :...G8CTRL in {PUBLIC SCHOOL,PRIVATE,OTH RELIG,
:   :   :          PRIVATE,NO RELIG}: No (113.5/2)
:   :   G8CTRL = CATHOLIC SCHOOL: Yes (3/2)
:   BYLOCUS1 in [(-0.715,0.263]-(0.263,1.24]]:
:   :...BYP76 in {LT HS GRAD,HS GRAD,VOC, ETC. LT 2YR,VOC, ETC.2+ YRS,
:       :         LT 2YRS COLLEGE,2YR+ COLLEGE,
:       :         4-5YR COLLEG PGM}: No (368.9/45.2)
:       BYP76 in {MASTER'S DEGREE,PH.D., M.D.,OTHR}: Yes (74.7/42.8)
BYSES in [(-0.214,0.792]-(0.792,1.8]]:
:...BYS51GA = YES:
    :...BYSC39M = ONE-HALF YEAR: No (0)
    :   BYSC39M = LESS THN 1/2 YR: Yes (2.2/0.2)
    :   BYSC39M in {NO SPECIFIC AMT,FULL YEAR}:
    :   :...BYT3_31.ALL in [NONE-5-9 KID'S PARENTS]: Yes (3.1/1.1)
    :       BYT3_31.ALL in [10-19 KID'S PARENTS-60+ KID'S PARENTS]: No (34.8/
0.5)
    BYS51GA = NO:
    :...BYS55A = MORE THAN TWICE:
        :...BYSES = (-0.214,0.792]: No (17/0.1)
        :   BYSES = (0.792,1.8]: Yes (1)
        BYS55A in [NEVER-ONCE OR TWICE]:
        :...BYP85F in {TRUE,FALSE}: Yes (220.5/94.3)
            BYP85F = HVN'T THGHT ABT YET:
            :...BYSES = (0.792,1.8]: Yes (1.5/0.1)
                BYSES = (-0.214,0.792]:
                :...G12CTRL1 in {PUBLIC,PRIV/OTH RELIG,
                    :            PRIV/NOT ASCRTND}: No (15.6/0.6)
                    G12CTRL1 in {CATHOLIC,PRIV/NON-RELIG}: Yes (0.3/0.2)
```

Figure 31. C5.0 tree (Study 1, 1372 possible predictors; results shown on training data)

```
par_ed = > HS & < 4YR DEG
|   unsafe = missing: No (6.67)
|   unsafe = No: Yes (222.67/98.67)
|   unsafe = Yes: No (32.0/8.0)
par_ed = COLLEGE GRADUATE: Yes (56.67/16.67)
par_ed = DIDN'T FINISH HS: No (51.33/4.0)
par_ed = H.S. GRAD OR GED: No (120.67/44.0)
par_ed = M.A./EQUIVALENT: Yes (52.67/6.67)
par_ed = missing
|   hhressc <= 5.242678: No (6.0)
|   hhressc > 5.242678: Yes (5.33/1.33)
par_ed = PH.D., M.D., OTHER
|   peerexcl = missing: No (0.67)
|   peerexcl = No: Yes (8.0)
|   peerexcl = Yes
|   |   BlkTeacher = 1: Yes (0.0)
|   |   BlkTeacher = 11 - 20: Yes (0.0)
|   |   BlkTeacher = 2: Yes (0.0)
|   |   BlkTeacher = 21 OR MORE: No (0.67)
|   |   BlkTeacher = 3 - 5: Yes (0.0)
|   |   BlkTeacher = 6 - 10: Yes (4.0)
|   |   BlkTeacher = missing: Yes (0.0)
|   |   BlkTeacher = NONE: Yes (4.0)
```

Figure 32. C4.5 tree (Study 1, 19 possible predictors; results shown on training data)

```
BYP45A = missing
|   BYSES <= -0.216: No (278.67/102.0)
|   BYSES > -0.216
|   |   BYSC15 = 10% OR LESS
|   |   |   BYS55C = missing: No (4.67/2.0)
|   |   |   BYS55C = MORE THAN TWICE: Yes (0.0)
|   |   |   BYS55C = NEVER
|   |   |   |   BYS76 = AT LEAST ONCE A WEEK: Yes (0.0)
|   |   |   |   BYS76 = DAILY: Yes (0.0)
|   |   |   |   BYS76 = LESS THAN ONCE A WK: No (6.67/2.0)
|   |   |   |   BYS76 = missing: No (2.0)
|   |   |   |   BYS76 = NEVER/ALMOST NEVER
|   |   |   |   |   BYS51GA = missing: Yes (6.67/0.67)
|   |   |   |   |   BYS51GA = NO
|   |   |   |   |   |   BYT3_16B.SOC.STUDIES.HISTORY = missing
|   |   |   |   |   |   |   BYS26B = 1/2 THE TIME: Yes (2.0)
|   |   |   |   |   |   |   BYS26B = ALWAYS/MOST TIME: Yes (0.0)
|   |   |   |   |   |   |   BYS26B = missing
|   |   |   |   |   |   |   |   BYS74B = missing
|   |   |   |   |   |   |   |   |   BYP85F = FALSE: Yes (146.0/18.0)
|   |   |   |   |   |   |   |   |   BYP85F = HVN'T THGHT ABT YET: No (2.67)
|   |   |   |   |   |   |   |   |   BYP85F = missing: Yes (15.33/3.33)
|   |   |   |   |   |   |   |   |   BYP85F = TRUE
|   |   |   |   |   |   |   |   |   |   BYS51AA = missing: Yes (0.0)
|   |   |   |   |   |   |   |   |   |   BYS51AA = NO: No (3.33)
|   |   |   |   |   |   |   |   |   |   BYS51AA = YES: Yes (6.0)
|   |   |   |   |   |   |   |   |   BYS74B = NO: Yes (4.0)
|   |   |   |   |   |   |   |   |   BYS74B = YES: No (0.67)
|   |   |   |   |   |   |   |   BYS26B = NEVER: No (1.33)
|   |   |   |   |   |   |   |   BYS26B = SOMETIMES: Yes (6.0)
|   |   |   |   |   |   |   BYT3_16B.SOC.STUDIES.HISTORY = NOT VERY WELL: Yes (4.
0)
|   |   |   |   |   |   |   BYT3_16B.SOC.STUDIES.HISTORY = PRETTY WELL: No (0.67)
|   |   |   |   |   |   |   BYT3_16B.SOC.STUDIES.HISTORY = VERY WELL: Yes (0.0)
|   |   |   |   |   |   |   BYT3_16B.SOC.STUDIES.HISTORY = WELL: No (0.67)
|   |   |   |   |   |   BYS51GA = YES: No (10.67/4.0)
|   |   |   BYS55C = ONCE OR TWICE: No (2.67)
|   |   BYSC15 = 11 - 20%: No (6.0)
|   |   BYSC15 = 21 - 30%: No (0.67)
|   |   BYSC15 = 31 - 40%: Yes (2.67/0.67)
|   |   BYSC15 = 41 - 50%: No (0.67)
|   |   BYSC15 = 61 - 70%: Yes (0.0)
|   |   BYSC15 = 81% OR MORE: Yes (0.0)
|   |   BYSC15 = missing: Yes (3.33/1.33)
BYP45A = NO: No (30.67/4.0)
BYP45A = YES: No (22.67)
```

Figure 33. C4.5 tree (Study 1, 1372 possible predictors; results shown on training data)

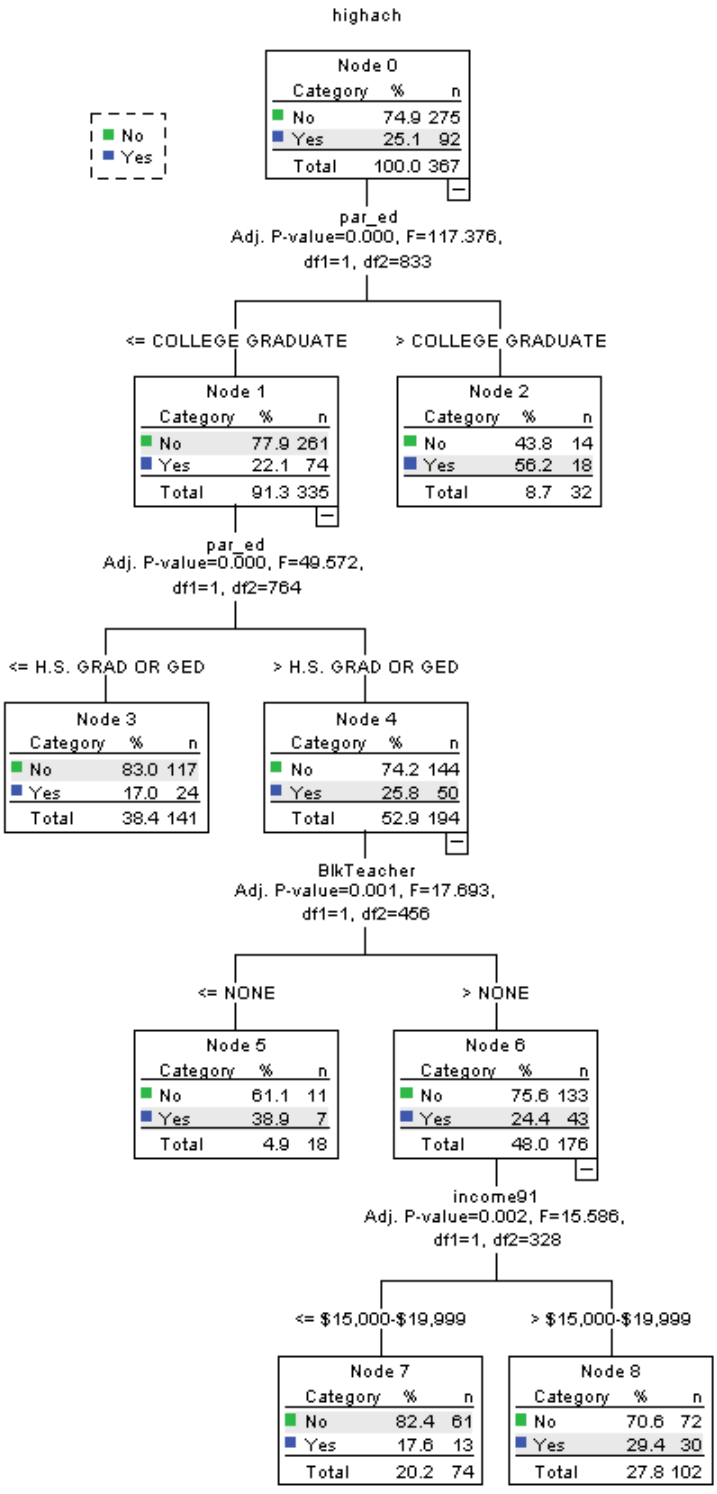


Figure 34. QUEST tree (Study 1, 19 possible predictors; performance on test set)

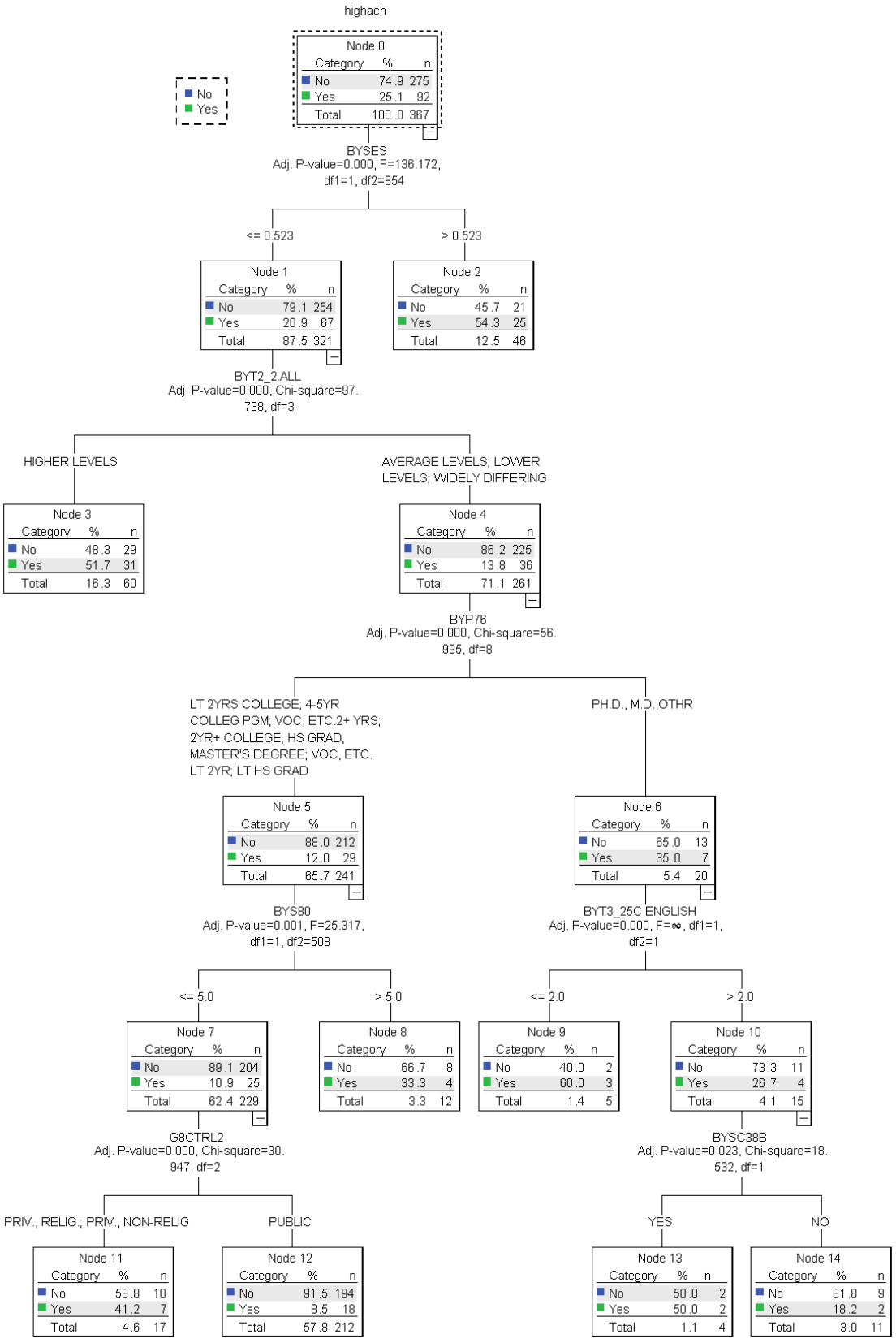


Figure 35. QUEST tree (Study 1, 1372 possible predictors; performance on test set)

Table 56. Categorization of variables included in Study 1 rule induction with 1372 possible predictors

Academic expectation	HOW SURE THAT YOU WILL GRADUATE FROM H.S (BYS46), POST-SECONDARY EDUCATION PLANS (BYPSEPLN)
Academics	R SENT TO OFFICE WITH SCHL WORK PROBLEMS (BYS55B), PARENTS RECEIVED WARNING ABOUT GRADES (BYS55D), ATTEND REMEDIAL MATH AT LEAST ONCE A WK (BYS67A), EVER HELD BACK A GRADE IN SCHOOL (BYS74), EVER REPEAT GRADE 1 (BYS74B), PARTICIPATED IN ACADEMIC HONORS SOCIETY (BYS82O), HELD BACK BECAUSE OF PARENTAL REQUEST (BYP45A), HELD BACK BECAUSE OF OTHER REASON (BYP45C), CHILD ENROLLED IN GIFTED/TALENTED PROG (BYP51), CHLD TEST SCORES NOT GOOD ENOUGH QUALIFY (BYP85F), ACHIEVEMENT LEVEL-THIS CLASS VS AVERAGE (BYT2_2.ENGLISH, BYT2_2.ALL)
Autonomy	Autonomy (autonomy), HOW OFTN PRNTS LIMIT GOING OUT WTH FRNDS (BYS38D)
Behavior	TALK TO COUNSELOR ABT DRUG/ALCOHOL ABUSE (BYS51GA), R SENT TO OFFICE FOR MISBEHAVING (BYS55A), PARENTS RECEIVED WARNING ABT ATTENDANCE (BYS55C), PARENTS RECEIVED WARNING ABOUT BEHAVIOR (BYS55E), HOW OFTEN DO YOU CUT OR SKIP CLASSES (BYS76)
Career expectation	KIND OF WORK R EXPECTS TO DO AT AGE 30 (BYS52)
Good peer	Peers expect college (peerexcl), good peers (goodpeer)
Household income	TOTAL FAMILY INCOME FRM ALL SOURCES 1987 (BYP80), YEARLY FAMILY INCOME (BYFAMINC)
Household resources	Household resources (hhressc), R'S FAMILY HAS A POCKET CALCULATOR (BYS35O)
HW out of school	Homework hours out of school (hw_outsch)
HW per week	HOW MUCH HOMEWORK PER WEEK - MINUTES (BYT2_7M.ENGLISH)
Language at home	HOW OFTEN R'S MOTHER SPEAKS LANG TO R (BYS26B), ENROLLED IN BILINGUAL EDUCATION (BYS68B), LANGUAGE MINORITY COMPOSITE (BYLM)
Lives w/ mother	R LIVES IN HOUSEHOLD WITH MOTHER (BYS8C)
Locus of control	GOOD LUCK MORE IMPORTANT THAN HARD WORK (BYS44C), CHANCE AND LUCK IMPORTANT IN MY LIFE (BYS44M), LOCUS OF CONTROL 1 (BYLOCUS1), TERTILE CODING OF VARIABLE BYLOCUS1 (BYLOCU1T)
Mother birthplace	8TH GRADER'S MOTHER'S BIRTHPLACE (BYP11)

Table 56 continued

Num LEP Ss in class	NUMBER OF LEP STUDENTS IN CLASS (BYT2_6.SOC.STUDIES.HISTORY)
Num of earners	# OF EARNERS CONTRIBUTD TO FAMILY INCOME (BYP81)
Parent employment	MOTHER/FEMALE GUARDIAN'S OCCUPATION (BYS4OCC), FATHER/MALE GUARDIAN EMPLOYMENT STATUS (BYS7A), FATHER/MALE GUARDIAN'S OCCUPATION (BYS7OCC), SPOUSE'S CURRENT WORK STATUS (BYP36A)
Parent expectation	HOW IMPORTANT CHILD COMPLETE SCHL FASTER (BYP52A), HOW FAR IN SCHOOL R EXPECT CHILD TO GO (BYP76)
Parental education	Parent education (par_ed), PARENTS' HIGHEST EDUCATION LEVEL (BYPARED), HIGHEST LEVEL OF EDUCATION R COMPLETED (BYP30)
Parental involvement	DID PRNTS/GRDNS WANT R TO TAKE ALGEBRA (BYS62), R KNOWS PARENT(S) OF CHILD'S 3RD FRIEND (BYP62B3), HOW OFTN TALKS TO CHILD ABOUT H.S. PLANS (BYP67)
Participated in 4-H	PARTICIPATED IN 4-H (BYS83G)
PE requirement	INSTRUCTION REQUIRED FOR PHYSICAL ED (BYSC39J),
Reading on own	HOW MUCH READING DO YOU DO ON YOUR OWN (BYS80),
School climate	SOMEONE OFFERED TO SELL R DRUGS AT SCHL (BYS57B), VERBAL ABUSE OF TEACHERS A PROBLEM (BYS58K), TEACHERS ARE INTERESTED IN STUDENTS (BYS59G)
School demographics	PERCENT MINORITY IN SCHOOL (G8MINOR), PERCENT FREE LUNCH IN SCHOOL (G8LUNCH), % OF WHITE NON-HISPANIC 8TH GRADERS (BYSC13E), % OF 8TH GRADERS LIMITED ENGL PROFICIENT (BYSC15), % OF CLASSROOM TIME TEACHING GIFTED (BYT3_22.ALL)
School structure/policy	ATTEND ENGLISH AT LEAST ONCE A WEEK (BYS67BA), OTHER PRACTICE FOR ASSINGMENT (BYSC24E), COUNSELORS INFLUENCE ASSIGNNG HS COURSES (BYSC36A), 8TH GRADERS RETAINED: FAILED MATH TEST (BYSC38B), 8TH GRADERS RETAINED: FAILED SCIENCE TST (BYSC38C), DISCIPLINE IS EMPHASIZED AT THIS SCHOOL (BYSC47B), DEVIATION FR SCHOOL RULES NOT TOLERATED (BYSC47K), ACTION FOR VERBAL ABUSE OF TCHR: 1ST OCC (BYSC50AI), ACTION FOR INJURY TO OTH STUD:REP OCCUR (BYSC50BB), ACTION FOR DRUG POSS.: REP OCCURRENCES (BYSC50BD)
School type	Religious school (religsch), SCHOOL CONTROL COMPOSITE (G8CTRL), EIGHTH GRADE SCHOOL COMPOSITE 1 (G8CTRL1), EIGHTH GRADE SCHOOL COMPOSITE 2 (G8CTRL2), SCHOOL CLASSIFICATION REPORTED BY SCHOOL (G10CTRL1), SCHOOL CLASSIFICATION REPORTED BY SCHOOL (G12CTRL1)

Table 56 continued

School urbanicity	URBANICITY COMPOSITE (G8URBAN)
Self confidence	I FEEL GOOD ABOUT MYSELF (BYS44A), I AM ABLE TO DO THINGS AS WELL AS OTHERS (BYS44E)
SES	SOCIO-ECONOMIC STATUS COMPOSITE (BYSES), QUARTILE CODING OF BYSES VARIABLE (BYSESQ)
Student at gr 12 (erroneously included)	MEMBER 12TH GRADE IN-SCHOOL CLASS 91-92 (G12COHRT)
T in-service support	NO SUPPORT REC'D FOR IN-SERVICE EDUCATN (BYT3_20A.ALL)
T teaches 8th gr gifted	% OF CLASSROOM TIME TEACHING GIFTED (BYT3_22.SOC.STUDIES.HISTORY, whether <50%), HOW SATISFIED W/ CONTENT OR CURRICULUM (BYT3_25B.MATH, whether responded to "if gifted…"), PROV INSTRUCTION IN 8TH GR GIFTED PROGRM (BYT3_21.ALL)
Talk to counselor about studies in class	TALK TO COUNSELOR ABOUT STUDIES IN CLASS (BYS51EA)
Teacher speaks non-English language	HOW WELL TEACHER WRITES LANGUAGE (BYT3_16D.ENGLISH, whether response provided or not for writing proficiency in non-English language in which teacher is most proficient), PROFICIENT IN GERMAN (BYT3_14B.SOC.STUDIES.HISTORY, whether response provided or not), HOW WELL TEACHER SPEAKS LANGUAGE (BYT3_16B.SOC.STUDIES.HISTORY, whether response provided or not for writing proficiency in non-English language in which teacher is most proficient), LEARNED LANGUAGE INFORMALLY (BYT3_17E.SOC.STUDIES.HISTORY, whether response provided or not for whether T informally learned non-English language in which s/he is most proficient)
Time spent on math HW	TIME SPENT ON MATH HOMEWORK EACH WEEK (BYS79A)
TV hrs on weekend	NO. OF HOURS R WATCHES TV ON WEEKENDS (BYS42B)
Where S goes after school	WHERE DOES CHILD USUALLY GO AFTER SCHL (BYP7), WHERE DOES CHILD USUALLY GO AFTER SCHL (BYP73)

APPENDIX D

DETAILED RULESET MINING RESULTS—STUDY 2

Table 57. CBA ruleset (Study 2, 29 possible predictors)

Rule order	Rule antecedent	Math achievement
1	Sexp=MA or MORE, geo_none=0, BYTXMIRR=(31.840], female=1	high
2	geo_none=0, al2_1=1, emph_m=(3.755], BYTXMIRR=(31.840]	high
3	BYSES=(0.0731.32], Pexp=MA or MORE, BYTXMIRR=(31.840], GPA910_m=(3.124]	high
4	al2_none=0, emph_m=(3.755], BYTXMIRR=(31.840], m_selfcpt=(7.459.94]	high
5	BYSES=(0.0731.32], Pexp=MA or MORE, BYTXMIRR=(31.840], m_selfcpt=(7.459.94]	high
6	BYSES=(-1.170.073], al2_none=0, BYTXMIRR=(31.840], SATplan=1	high
7	BYSES=(-1.170.073], geo_none=0, al2_none=0, BYTXMIRR=(31.840]	high
8	Sexp=MA or MORE, BYTXMIRR=(31.840], m_selfcpt=(7.459.94], female=1	high
9	Pexp=MA or MORE, BYTXMIRR=(31.840], m_selfcpt=(7.459.94], female=1	high
10	BYSES=(0.0731.32], emph_m=(2.53.75], BYTXMIRR=(31.840], GPA910_m=(3.124]	high
11	emph_m=(2.53.75], BYTXMIRR=(31.840], GPA910_m=(3.124], female=1	high
12	BYSES=(0.0731.32], Pexp=MA or MORE, geo_none=0, BYTXMIRR=(31.840]	high
13	Sexp=4YR DEG, geo_none=0, al2_none=0, BYTXMIRR=(31.840]	high
14	Pexp=MA or MORE, al2_1=0, BYTXMIRR=(31.840], white=1	high
15	Pexp=MA or MORE, gm_1=0, al2_none=1, BYTXMIRR=(31.840]	high
16	BYSES=(0.0731.32], t_rspnsv=(5.847.8], BYTXMIRR=(31.840], SATplan=1	high
17	BYSES=(0.0731.32], Sexp=MA or MORE, BYTXMIRR=(31.840], GPA910_m=(3.124]	high
18	BYSES=(0.0731.32], Pexp=MA or MORE, BYGRADS=(3.124], BYTXMIRR=(31.840]	high
19	gm_2=0, geo_none=0, al2_1=1, BYTXMIRR=(31.840]	high
20	Sexp=MA or MORE, emph_m=(3.755], BYTXMIRR=(31.840], m_selfcpt=(7.459.94]	high
21	BYSES=(0.0731.32], Pexp=MA or MORE, BYTXMIRR=(31.840], SATplan=1	high
22	Sexp=MA or MORE, gm_1=0, BYTXMIRR=(31.840], female=1	high
23	Sexp=MA or MORE, BYGRADS=(3.124], BYTXMIRR=(31.840], female=1	high
24	BYSES=(0.0731.32], geo_none=0, al2_none=0, BYTXMIRR=(31.840]	high
25	Sexp=MA or MORE, BYGRADS=(3.124], BYTXMIRR=(31.840], GPA910_m=(3.124]	high
26	al2_1=1, emph_m=(3.755], BYTXMIRR=(31.840], SATplan=1	high

Table 57 continued

27	BYSES=(0.0731.32], Sexp=MA or MORE, al2_none=0, BYTXMIRR=(31.840]	high
28	BYSES=(0.0731.32], al2_none=0, BYTXMIRR=(31.840], GPA910_m=(3.124]	high
29	Pexp=MA or MORE, geo_none=0, al2_half=0, BYTXMIRR=(31.840]	high
30	Pexp=MA or MORE, geo_none=0, BYTXMIRR=(31.840], GPA910_m=(3.124]	high
31	Pexp=MA or MORE, BYTXMIRR=(31.840], m_selfcpt=(7.459.94]	high
32	Sexp=MA or MORE, al2_1=1, BYTXMIRR=(31.840], hispanic=0	high
33	Sexp=MA or MORE, al2_none=0, BYTXMIRR=(31.840], white=1	high
34	BYSES=(0.0731.32], Pexp=MA or MORE, BYTXMIRR=(31.840], hispanic=0	high
35	al2_none=0, BYTXMIRR=(31.840], GPA910_m=(3.124], grad_eff=Yes Sure will grad	high
36	gm_2=0, geo_none=0, al2_none=0, BYTXMIRR=(31.840]	high
37	emph_m=(2.53.75], BYTXMIRR=(31.840], GPA910_m=(3.124]	high
38	al2_none=0, BYTXMIRR=(31.840], m_selfcpt=(7.459.94], female=1	high
39	al2_2=0, emph_m=(3.755], BYTXMIRR=(31.840], m_selfcpt=(7.459.94]	high
40	gm_half=0, emph_m=(3.755], BYTXMIRR=(31.840], m_selfcpt=(7.459.94]	high
41	al2_none=0, BYTXMIRR=(31.840], SATplan=1, m_selfcpt=(7.459.94]	high
42	Pexp=MA or MORE, geo_none=0, BYTXMIRR=(31.840], female=1	high
43	BYSES=(0.0731.32], gm_none=1, BYTXMIRR=(31.840], GPA910_m=(3.124]	high
44	Sexp=4YR DEG, al2_none=0, BYTXMIRR=(31.840], SATplan=1	high
45	geo_none=0, emph_m=(2.53.75], BYTXMIRR=(31.840], m_selfcpt=(7.459.94]	high
46	emph_m=(2.53.75], BYTXMIRR=(31.840], SATplan=1, m_selfcpt=(7.459.94]	high
47	gm_2=0, geo_none=0, BYTXMIRR=(31.840], GPA910_m=(3.124]	high
48	gm_none=1, emph_m=(1.252.5], BYTXMIRR=(31.840], GPA910_m=(3.124]	high
49	BYSES=(-1.170.073], al2_none=0, BYTXMIRR=(31.840], grad_eff=Yes Sure will grad	high
50	Pexp=MA or MORE, geo_none=0, BYTXMIRR=(31.840], black=0	high
51	gm_2=0, BYTXMIRR=(31.840], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
52	BYSES=(-1.170.073], geo_none=0, BYTXMIRR=(31.840], GPA910_m=(3.124]	high
53	BYSES=(-1.170.073], BYTXMIRR=(31.840], m_selfcpt=(7.459.94], white=1	high
54	BYGRADS=(3.124], t_rspnsv=(5.847.8], BYTXMIRR=(31.840], SATplan=1	high
55	BYSES=(-1.170.073], geo_none=0, BYTXMIRR=(31.840], m_selfcpt=(7.459.94]	high
56	gm_2=0, BYTXMIRR=(31.840], m_selfcpt=(7.459.94], female=1	high
57	gm_2=0, geo_none=0, BYTXMIRR=(31.840], m_selfcpt=(7.459.94]	high
58	BYSES=(-1.170.073], geo_none=0, BYTXMIRR=(31.840], SATplan=1	high
59	BYGRADS=(3.124], emph_m=(2.53.75], BYTXMIRR=(31.840], SATplan=1	high
60	BYSES=(-1.170.073], al2_none=0, BYTXMIRR=(31.840], hispanic=0	high
61	BYSES=(0.0731.32], geo_1=1, emph_m=(2.53.75], BYTXMIRR=(31.840]	high

Table 57 continued

62	al2_2=0, BYTXMIRR=(31.840], GPA910_m=(3.124], female=0	high
63	geo_none=0, al2_2=0, BYTXMIRR=(31.840], GPA910_m=(3.124]	high
64	Pexp=MA or MORE, Sexp=MA or MORE, BYTXMIRR=(31.840], white=1	high
65	geo_none=0, BYTXMIRR=(31.840], m_selfcpt=(7.459.94], female=1	high
66	gm_none=1, BYTXMIRR=(31.840], GPA910_m=(3.124], SATplan=1	high
67	BYSES=(0.0731.32], geo_none=0, BYTXMIRR=(31.840], female=1	high
68	Sexp=MA or MORE, BYGRADS=(3.124], emph_m=(2.53.75], BYTXMIRR=(31.840]	high
69	emph_m=(2.53.75], BYTXMIRR=(31.840], grad_eff=Yes Sure will grad, SATplan=1	high
70	BYSES=(0.0731.32], gm_2=0, geo_1=1, BYTXMIRR=(31.840]	high
71	BYSES=(0.0731.32], Sexp=MA or MORE, BYGRADS=(3.124], BYTXMIRR=(31.840]	high
72	gm_2=0, BYTXMIRR=(31.840], m_selfcpt=(7.459.94], black=0	high
73	BYSES=(0.0731.32], Sexp=MA or MORE, BYTXMIRR=(31.840], SATplan=1	high
74	emph_m=(2.53.75], BYTXMIRR=(31.840], grad_eff=Yes Sure will grad, female=1	high
75	t_rspnsv=(5.847.8], BYTXMIRR=(31.840], grad_eff=Yes Sure will grad	high
76	BYGRADS=(3.124], emph_m=(2.53.75], BYTXMIRR=(31.840], female=1	high
77	geo_1=1, al2_half=0, emph_m=(3.755], BYTXMIRR=(31.840]	high
78	geo_none=0, al2_half=0, BYTXMIRR=(31.840], SATplan=1	high
79	BYSES=(-1.170.073], BYGRADS=(3.124], BYTXMIRR=(31.840], GPA910_m=(3.124]	high
80	BYSES=(-1.170.073], BYGRADS=(3.124], BYTXMIRR=(31.840], SATplan=1	high
81	geo_none=0, BYTXMIRR=(31.840], GPA910_m=(3.124]	high
82	emph_m=(-0.0051.25], BYTXMIRR=(31.840], black=0	high
83	BYSES=(0.0731.32], BYGRADS=(3.124], BYTXMIRR=(31.840], female=1	high
84	BYGRADS=(3.124], BYTXMIRR=(31.840], GPA910_m=(3.124], grad_eff=Yes Sure will grad	high
85	gm_2=0, geo_1=1, BYTXMIRR=(31.840], white=1	high
86	Sexp=MA or MORE, al2_half=0, BYTXMIRR=(31.840], SATplan=1	high
87	Sexp=MA or MORE, emph_m=(2.53.75], BYTXMIRR=(31.840]	high
88	BYSES=(0.0731.32], BYTXMIRR=(31.840], grad_eff=Yes Sure will grad, female=1	high
89	Sexp=MA or MORE, geo_2=0, BYTXMIRR=(31.840], white=1	high
90	al2_half=0, emph_m=(3.755], BYTXMIRR=(31.840], SATplan=1	high
91	BYSES=(0.0731.32], geo_1=1, BYTXMIRR=(31.840], m_selfcpt=(4.977.45]	high
92	gm_2=0, geo_none=0, BYTXMIRR=(31.840], white=1	high
93	BYGRADS=(3.124], BYTXMIRR=(31.840], grad_eff=Yes Sure will grad, female=1	high
94	al2_2=0, BYTXMIRR=(31.840], SATplan=1, white=1	high
95	Sexp=4YR DEG, geo_1=1, emph_m=(2.53.75], BYTXMIRR=(31.840]	high
96	BYSES=(0.0731.32], geo_1=1, BYTXMIRR=(31.840], GPA910_m=(2.253.12]	high

Table 57 continued

97	BYSES=(0.0731.32], geo_none=0, geo_half=0, BYTXMIRR=(31.840]	high
98	al2_2=0, BYTXMIRR=(31.840], grad_eff=Yes Sure will grad, white=1	high
99	BYSES=(0.0731.32], gm_none=1, t_rspnsv=(3.895.84], BYTXMIRR=(31.840]	high
100	Pexp=MA or MORE, al2_1=1, GPA910_m=(3.124], female=0	high
101	Pexp=MA or MORE, al2_none=0, GPA910_m=(3.124], female=0	high
102	Pexp=MA or MORE, al2_1=1, m_selfcpt=(7.459.94], female=0	high
103	Pexp=MA or MORE, geo_1=1, al2_1=1, m_selfcpt=(7.459.94]	high
104	BYSES=(0.0731.32], BYGRADS=(3.124], al2_1=1, GPA910_m=(3.124]	high
105	Pexp=MA or MORE, geo_none=0, al2_1=1, m_selfcpt=(7.459.94]	high
106	BYSES=(0.0731.32], geo_1=1, al2_none=0, m_selfcpt=(7.459.94]	high
107	BYSES=(0.0731.32], Pexp=MA or MORE, al2_1=1, m_selfcpt=(7.459.94]	high
108	BYSES=(0.0731.32], geo_none=0, al2_1=1, m_selfcpt=(7.459.94]	high
109	BYSES=(0.0731.32], geo_1=1, al2_1=1, GPA910_m=(3.124]	high
110	BYSES=(0.0731.32], Pexp=MA or MORE, al2_1=1, GPA910_m=(3.124]	high
111	BYSES=(0.0731.32], Pexp=MA or MORE, al2_none=0, m_selfcpt=(7.459.94]	high
112	Pexp=MA or MORE, geo_1=1, al2_1=1, GPA910_m=(3.124]	high
113	BYSES=(0.0731.32], BYGRADS=(3.124], al2_none=0, GPA910_m=(3.124]	high
114	BYSES=(0.0731.32], BYGRADS=(3.124], BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94]	high
115	geo_1=1, al2_none=0, emph_m=(3.755], GPA910_m=(3.124]	high
116	geo_none=0, al2_1=1, emph_m=(3.755], GPA910_m=(3.124]	high
117	BYSES=(0.0731.32], Pexp=MA or MORE, geo_1=1, al2_1=1	high
118	Pexp=MA or MORE, geo_1=1, al2_none=0, m_selfcpt=(7.459.94]	high
119	BYGRADS=(3.124], al2_1=1, GPA910_m=(3.124], female=0	high
120	Pexp=MA or MORE, al2_none=0, m_selfcpt=(7.459.94], female=0	high
121	BYSES=(0.0731.32], geo_none=0, al2_none=0, m_selfcpt=(7.459.94]	high
122	geo_none=0, al2_1=1, emph_m=(3.755], m_selfcpt=(7.459.94]	high
123	BYGRADS=(3.124], geo_1=1, al2_1=1, GPA910_m=(3.124]	high
124	BYSES=(0.0731.32], Sexp=MA or MORE, al2_1=1, m_selfcpt=(7.459.94]	high
125	BYSES=(0.0731.32], geo_1=1, al2_none=0, GPA910_m=(3.124]	high
126	BYGRADS=(3.124], geo_none=0, al2_1=1, GPA910_m=(3.124]	high
127	geo_1=1, al2_none=0, emph_m=(3.755], m_selfcpt=(7.459.94]	high
128	Pexp=MA or MORE, BYGRADS=(3.124], geo_1=1, al2_1=1	high
129	geo_none=0, al2_1=1, BYTXMIRR=(23.631.8], GPA910_m=(3.124]	high
130	geo_1=1, al2_none=0, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
131	BYSES=(0.0731.32], Pexp=MA or MORE, geo_none=0, al2_1=1	high
132	Pexp=MA or MORE, geo_1=1, al2_1=1, white=1	high
133	geo_none=0, al2_none=0, emph_m=(3.755], GPA910_m=(3.124]	high
134	BYSES=(0.0731.32], Pexp=MA or MORE, BYGRADS=(3.124], m_selfcpt=(7.459.94]	high
135	Pexp=MA or MORE, gm_none=1, al2_1=1, female=0	high
136	BYSES=(0.0731.32], BYGRADS=(3.124], al2_none=0, m_selfcpt=(7.459.94]	high

Table 57 continued

137	BYSES=(0.0731.32], al2_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
138	BYGRADS=(3.124], geo_1=1, al2_1=1, m_selfcpt=(7.459.94]	high
139	BYSES=(0.0731.32], Pexp=MA or MORE, al2_1=1, white=1	high
140	BYSES=(0.0731.32], Pexp=MA or MORE, m_selfcpt=(7.459.94], white=1	high
141	BYSES=(0.0731.32], geo_none=0, al2_none=0, GPA910_m=(3.124]	high
142	BYSES=(0.0731.32], al2_1=1, GPA910_m=(3.124], female=0	high
143	BYSES=(0.0731.32], Sexp=MA or MORE, al2_none=0, m_selfcpt=(7.459.94]	high
144	BYSES=(0.0731.32], Pexp=MA or MORE, geo_1=1, GPA910_m=(3.124]	high
145	BYGRADS=(3.124], al2_none=0, GPA910_m=(3.124], female=0	high
146	BYSES=(0.0731.32], BYGRADS=(3.124], geo_1=1, al2_1=1	high
147	BYSES=(0.0731.32], Pexp=MA or MORE, geo_1=1, al2_none=0	high
148	Pexp=MA or MORE, Sexp=MA or MORE, al2_1=1, female=0	high
149	BYGRADS=(3.124], geo_1=1, al2_none=0, m_selfcpt=(7.459.94]	high
150	BYSES=(0.0731.32], Pexp=MA or MORE, geo_1=1, m_selfcpt=(7.459.94]	high
151	Pexp=MA or MORE, BYGRADS=(3.124], geo_none=0, al2_1=1	high
152	Pexp=MA or MORE, BYGRADS=(3.124], al2_1=1, m_selfcpt=(7.459.94]	high
153	Sexp=MA or MORE, geo_1=1, al2_1=1, m_selfcpt=(7.459.94]	high
154	BYGRADS=(3.124], al2_none=0, t_rspnsv=(5.847.8], m_selfcpt=(7.459.94]	high
155	Pexp=MA or MORE, al2_1=1, SATplan=1, m_selfcpt=(7.459.94]	high
156	Pexp=MA or MORE, geo_1=1, al2_1=1, SATplan=1	high
157	BYSES=(0.0731.32], Pexp=MA or MORE, geo_none=0, al2_none=0	high
158	Sexp=MA or MORE, al2_none=0, emph_m=(3.755], GPA910_m=(3.124]	high
159	BYSES=(0.0731.32], geo_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
160	Pexp=MA or MORE, geo_1=1, m_selfcpt=(7.459.94], white=1	high
161	gm_none=1, geo_1=1, al2_1=1, m_selfcpt=(7.459.94]	high
162	BYGRADS=(3.124], geo_1=1, al2_none=0, t_rspnsv=(5.847.8]	high
163	geo_1=1, al2_1=1, m_selfcpt=(7.459.94], white=1	high
164	geo_none=0, al2_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
165	geo_none=0, al2_none=0, emph_m=(3.755], m_selfcpt=(7.459.94]	high
166	Pexp=MA or MORE, geo_1=1, al2_1=1, emph_m=(3.755]	high
167	BYSES=(0.0731.32], BYGRADS=(3.124], geo_none=0, al2_1=1	high
168	BYSES=(0.0731.32], BYGRADS=(3.124], geo_1=1, m_selfcpt=(7.459.94]	high
169	BYSES=(0.0731.32], Pexp=MA or MORE, m_selfcpt=(7.459.94], female=0	high
170	geo_none=0, al2_1=1, m_selfcpt=(7.459.94], white=1	high
171	BYGRADS=(3.124], geo_none=0, al2_none=0, GPA910_m=(3.124]	high
172	Pexp=4YR DEG, geo_none=0, al2_1=1, m_selfcpt=(7.459.94]	high
173	Pexp=MA or MORE, BYGRADS=(3.124], geo_1=1, al2_none=0	high
174	BYSES=(0.0731.32], Pexp=MA or MORE, al2_none=0, white=1	high
175	BYSES=(0.0731.32], Sexp=MA or MORE, geo_1=1, m_selfcpt=(7.459.94]	high
176	BYSES=(0.0731.32], Pexp=MA or MORE, geo_none=0, GPA910_m=(3.124]	high
177	Pexp=MA or MORE, Sexp=4YR DEG, geo_1=1, GPA910_m=(3.124]	high

Table 57 continued

178	BYGRADS=(3.124], al2_none=0, emph_m=(3.755], GPA910_m=(3.124]	high
179	Sexp=MA or MORE, BYGRADS=(3.124], al2_1=1, GPA910_m=(3.124]	high
180	Pexp=MA or MORE, geo_1=1, GPA910_m=(3.124], white=1	high
181	BYSES=(0.0731.32], Pexp=MA or MORE, al2_none=0, female=0	high
182	BYGRADS=(3.124], geo_1=1, al2_1=1, female=0	high
183	BYGRADS=(3.124], al2_1=1, m_selfcpt=(7.459.94], white=1	high
184	Pexp=4YR DEG, BYGRADS=(3.124], al2_1=1, GPA910_m=(3.124]	high
185	BYSES=(0.0731.32], BYGRADS=(3.124], geo_1=1, al2_none=0	high
186	BYSES=(0.0731.32], Sexp=MA or MORE, al2_1=1, GPA910_m=(3.124]	high
187	Sexp=MA or MORE, al2_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
188	Pexp=MA or MORE, geo_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
189	Pexp=MA or MORE, geo_1=1, GPA910_m=(3.124], female=0	high
190	Sexp=MA or MORE, al2_1=1, m_selfcpt=(7.459.94], female=0	high
191	BYGRADS=(3.124], al2_1=1, GPA910_m=(3.124], white=1	high
192	Pexp=MA or MORE, geo_none=0, t_rspnsv=(3.895.84], GPA910_m=(3.124]	high
193	BYGRADS=(3.124], geo_1=1, al2_1=1, white=1	high
194	BYSES=(0.0731.32], BYGRADS=(3.124], geo_none=0, m_selfcpt=(7.459.94]	high
195	BYSES=(0.0731.32], al2_1=1, m_selfcpt=(7.459.94], female=0	high
196	BYGRADS=(3.124], al2_1=1, m_selfcpt=(7.459.94], female=0	high
197	BYSES=(0.0731.32], Pexp=MA or MORE, BYGRADS=(3.124], GPA910_m=(3.124]	high
198	BYSES=(0.0731.32], geo_1=1, emph_m=(3.755], GPA910_m=(3.124]	high
199	BYGRADS=(3.124], al2_1=1, SATplan=1, m_selfcpt=(7.459.94]	high
200	Pexp=MA or MORE, Sexp=4YR DEG, geo_1=1, m_selfcpt=(7.459.94]	high
201	gm_none=1, geo_1=1, al2_none=0, m_selfcpt=(7.459.94]	high
202	Sexp=MA or MORE, BYGRADS=(3.124], geo_1=1, al2_1=1	high
203	BYSES=(0.0731.32], geo_none=0, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
204	Pexp=MA or MORE, geo_1=1, al2_none=0, SATplan=1	high
205	Pexp=MA or MORE, Sexp=MA or MORE, al2_1=1, t_rspnsv=(3.895.84]	high
206	Sexp=MA or MORE, geo_1=1, emph_m=(3.755], m_selfcpt=(7.459.94]	high
207	Pexp=MA or MORE, BYGRADS=(3.124], geo_1=1, m_selfcpt=(7.459.94]	high
208	BYSES=(0.0731.32], Pexp=MA or MORE, GPA910_m=(3.124], white=1	high
209	BYSES=(0.0731.32], al2_1=1, SATplan=1, m_selfcpt=(7.459.94]	high
210	BYGRADS=(3.124], al2_none=0, m_selfcpt=(7.459.94], white=1	high
211	BYSES=(0.0731.32], geo_1=1, BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94]	high
212	geo_none=0, al2_none=0, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
213	Sexp=MA or MORE, geo_1=1, emph_m=(3.755], GPA910_m=(3.124]	high
214	Sexp=MA or MORE, al2_1=1, GPA910_m=(3.124], female=0	high
215	BYSES=(0.0731.32], Sexp=MA or MORE, BYGRADS=(3.124], m_selfcpt=(7.459.94]	high
216	Pexp=4YR DEG, BYGRADS=(3.124], BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94]	high

Table 57 continued

217	geo_1=1, emph_m=(3.755], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
218	geo_1=1, emph_m=(3.755], GPA910_m=(3.124], female=0	high
219	Pexp=MA or MORE, m_selfcpt=(7.459.94], female=0, white=1	high
220	Pexp=MA or MORE, geo_1=1, GPA910_m=(3.124], black=0	high
221	BYSES=(0.0731.32], BYGRADS=(3.124], geo_none=0, al2_none=0	high
222	Sexp=MA or MORE, al2_1=1, t_rspnsv=(3.895.84], GPA910_m=(3.124]	high
223	BYSES=(0.0731.32], BYGRADS=(3.124], BYTXMIRR=(23.631.8], GPA910_m=(3.124]	high
224	Pexp=MA or MORE, geo_none=0, al2_none=0, white=1	high
225	Pexp=MA or MORE, geo_none=0, al2_none=0, emph_m=(3.755]	high
226	Sexp=MA or MORE, geo_1=1, m_selfcpt=(7.459.94], white=1	high
227	BYSES=(0.0731.32], BYGRADS=(3.124], geo_1=1, GPA910_m=(3.124]	high
228	Sexp=MA or MORE, BYGRADS=(3.124], geo_none=0, al2_1=1	high
229	BYSES=(0.0731.32], geo_1=1, m_selfcpt=(7.459.94], white=1	high
230	Pexp=MA or MORE, BYGRADS=(3.124], GPA910_m=(3.124], female=0	high
231	BYSES=(0.0731.32], Sexp=MA or MORE, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
232	BYSES=(0.0731.32], geo_none=0, BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94]	high
233	Pexp=MA or MORE, BYGRADS=(3.124], al2_1=1, t_rspnsv=(3.895.84]	high
234	BYSES=(0.0731.32], BYGRADS=(3.124], emph_m=(3.755], m_selfcpt=(7.459.94]	high
235	BYSES=(0.0731.32], BYGRADS=(3.124], geo_none=0, GPA910_m=(3.124]	high
236	BYSES=(0.0731.32], Sexp=MA or MORE, geo_none=0, m_selfcpt=(7.459.94]	high
237	BYGRADS=(3.124], al2_none=0, SATplan=1, m_selfcpt=(7.459.94]	high
238	geo_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94], female=0	high
239	Pexp=MA or MORE, BYGRADS=(3.124], al2_1=1, SATplan=1	high
240	geo_1=1, BYTXMIRR=(23.631.8], GPA910_m=(3.124], female=0	high
241	Pexp=MA or MORE, BYGRADS=(3.124], emph_m=(3.755], m_selfcpt=(7.459.94]	high
242	BYSES=(0.0731.32], geo_1=1, emph_m=(3.755], m_selfcpt=(7.459.94]	high
243	Pexp=MA or MORE, geo_none=0, al2_none=0, female=0	high
244	geo_none=0, emph_m=(3.755], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
245	Pexp=MA or MORE, BYGRADS=(3.124], m_selfcpt=(7.459.94], female=0	high
246	Pexp=MA or MORE, geo_1=1, m_selfcpt=(7.459.94], black=0	high
247	BYGRADS=(3.124], geo_1=1, al2_1=1, hispanic=0	high
248	Pexp=MA or MORE, t_rspnsv=(3.895.84], m_selfcpt=(7.459.94], white=1	high
249	BYSES=(0.0731.32], Pexp=MA or MORE, m_selfcpt=(7.459.94], black=0	high
250	BYSES=(0.0731.32], BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94], female=0	high
251	BYSES=(0.0731.32], geo_1=1, al2_1=1, emph_m=(3.755]	high
252	Pexp=MA or MORE, al2_1=1, emph_m=(3.755], SATplan=1	high
253	Sexp=MA or MORE, geo_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high

Table 57 continued

254	BYGRADS=(3.124], geo_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
255	BYSES=(0.0731.32], Sexp=MA or MORE, geo_1=1, al2_1=1	high
256	BYSES=(0.0731.32], geo_1=1, GPA910_m=(3.124], white=1	high
257	BYSES=(0.0731.32], BYGRADS=(3.124], m_selfcpt=(7.459.94], white=1	high
258	BYSES=(0.0731.32], Sexp=MA or MORE, m_selfcpt=(7.459.94], white=1	high
259	BYGRADS=(3.124], geo_none=0, al2_none=0, white=1	high
260	BYGRADS=(3.124], geo_1=1, al2_1=1, m_selfcpt=(4.977.45]	high
261	Pexp=MA or MORE, GPA910_m=(3.124], m_selfcpt=(7.459.94], female=0	high
262	Pexp=MA or MORE, al2_1=1, t_rspnsv=(3.895.84], white=1	high
263	BYGRADS=(3.124], geo_1=1, emph_m=(3.755], m_selfcpt=(7.459.94]	high
264	BYSES=(0.0731.32], Sexp=MA or MORE, geo_none=0, al2_1=1	high
265	Pexp=MA or MORE, al2_1=1, SATplan=1, white=1	high
266	BYSES=(0.0731.32], BYGRADS=(3.124], geo_1=1, t_rspnsv=(5.847.8]	high
267	BYGRADS=(3.124], al2_1=1, SATplan=1, female=0	high
268	BYSES=(0.0731.32], Sexp=MA or MORE, BYGRADS=(3.124], GPA910_m=(3.124]	high
269	BYSES=(0.0731.32], BYGRADS=(3.124], gm_none=1, m_selfcpt=(7.459.94]	high
270	BYGRADS=(3.124], geo_1=1, emph_m=(3.755], GPA910_m=(3.124]	high
271	geo_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
272	Sexp=MA or MORE, BYGRADS=(3.124], emph_m=(3.755], m_selfcpt=(7.459.94]	high
273	Pexp=MA or MORE, emph_m=(3.755], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
274	BYSES=(0.0731.32], geo_1=1, al2_none=0, emph_m=(3.755]	high
275	gm_none=1, geo_none=0, al2_none=0, GPA910_m=(3.124]	high
276	al2_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94], female=0	high
277	Sexp=4YR DEG, BYGRADS=(3.124], BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94]	high
278	geo_1=1, BYTXMIRR=(23.631.8], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
279	Sexp=MA or MORE, geo_none=0, m_selfcpt=(7.459.94], white=1	high
280	BYGRADS=(3.124], al2_1=1, t_rspnsv=(5.847.8], SATplan=1	high
281	Sexp=MA or MORE, gm_none=1, geo_1=1, al2_1=1	high
282	BYGRADS=(3.124], geo_none=0, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
283	BYSES=(0.0731.32], BYGRADS=(3.124], emph_m=(3.755], GPA910_m=(3.124]	high
284	Pexp=MA or MORE, GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
285	BYSES=(0.0731.32], Pexp=MA or MORE, BYGRADS=(3.124], geo_1=1	high
286	al2_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
287	Sexp=MA or MORE, geo_1=1, GPA910_m=(3.124], white=1	high
288	BYSES=(0.0731.32], gm_none=1, geo_1=1, GPA910_m=(3.124]	high
289	BYGRADS=(3.124], geo_1=1, m_selfcpt=(7.459.94], white=1	high
290	BYSES=(0.0731.32], BYGRADS=(3.124], geo_none=0, t_rspnsv=(5.847.8]	high
291	BYGRADS=(3.124], geo_1=1, GPA910_m=(3.124], female=0	high

Table 57 continued

292	al2_none=0, emph_m=(3.755], GPA910_m=(3.124], SATplan=1	high
293	geo_1=1, emph_m=(3.755], GPA910_m=(3.124], white=1	high
294	BYGRADS=(3.124], geo_none=0, al2_none=0, female=0	high
295	BYGRADS=(3.124], geo_1=1, al2_none=0, hispanic=0	high
296	geo_1=1, t_rspnsv=(5.847.8], GPA910_m=(3.124], white=1	high
297	geo_1=1, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], GPA910_m=(3.124]	high
298	Pexp=MA or MORE, emph_m=(3.755], SATplan=1, m_selfcpt=(7.459.94]	high
299	BYSES=(0.0731.32], BYGRADS=(3.124], gm_none=1, GPA910_m=(3.124]	high
300	Sexp=MA or MORE, gm_1=0, geo_1=1, al2_1=1	high
301	BYSES=(0.0731.32], BYGRADS=(3.124], al2_1=1, SATplan=1	high
302	BYGRADS=(3.124], BYTXMIRR=(23.631.8], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
303	BYSES=(0.0731.32], Pexp=MA or MORE, geo_1=1, female=0	high
304	BYSES=(0.0731.32], geo_none=0, BYTXMIRR=(23.631.8], GPA910_m=(3.124]	high
305	BYSES=(0.0731.32], Sexp=MA or MORE, geo_1=1, emph_m=(3.755]	high
306	BYSES=(0.0731.32], Pexp=MA or MORE, BYGRADS=(3.124], geo_none=0	high
307	BYGRADS=(3.124], geo_1=1, al2_1=1, black=0	high
308	geo_1=1, GPA910_m=(3.124], m_selfcpt=(7.459.94], black=0	high
309	al2_none=0, GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
310	BYSES=(0.0731.32], Sexp=MA or MORE, geo_none=0, al2_none=0	high
311	BYSES=(0.0731.32], Pexp=MA or MORE, geo_1=1, emph_m=(3.755]	high
312	geo_none=0, GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
313	al2_none=0, GPA910_m=(3.124], m_selfcpt=(7.459.94], female=0	high
314	geo_1=1, emph_m=(3.755], GPA910_m=(3.124], SATplan=1	high
315	geo_none=0, GPA910_m=(3.124], m_selfcpt=(7.459.94], female=0	high
316	BYGRADS=(3.124], geo_none=0, GPA910_m=(3.124], female=0	high
317	geo_none=0, BYTXMIRR=(23.631.8], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
318	Sexp=MA or MORE, geo_1=1, GPA910_m=(3.124], female=0	high
319	Pexp=MA or MORE, GPA910_m=(3.124], female=0, black=0	high
320	Pexp=MA or MORE, BYGRADS=(3.124], al2_none=0, SATplan=1	high
321	Pexp=MA or MORE, geo_none=0, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8]	high
322	BYGRADS=(3.124], geo_1=1, GPA910_m=(3.124], white=1	high
323	al2_1=1, SATplan=1, m_selfcpt=(7.459.94], white=1	high
324	al2_1=1, emph_m=(3.755], SATplan=1, m_selfcpt=(7.459.94]	high
325	geo_none=0, emph_m=(3.755], m_selfcpt=(7.459.94], female=0	high
326	BYGRADS=(3.124], emph_m=(3.755], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
327	BYSES=(0.0731.32], BYGRADS=(3.124], al2_none=0, SATplan=1	high
328	geo_1=1, emph_m=(3.755], m_selfcpt=(7.459.94], white=1	high
329	Pexp=MA or MORE, BYGRADS=(3.124], emph_m=(3.755], GPA910_m=(3.124]	high

Table 57 continued

330	Pexp=MA or MORE, geo_1=1, BYTXMIRR=(23.631.8], female=0	high
331	Pexp=MA or MORE, BYGRADS=(3.124], geo_1=1, white=1	high
332	Pexp=MA or MORE, BYGRADS=(3.124], SATplan=1, m_selfcpt=(7.459.94]	high
333	geo_1=1, al2_none=0, emph_m=(3.755], white=1	high
334	geo_1=1, BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94], female=0	high
335	BYGRADS=(3.124], al2_1=1, BYTXMIRR=(23.631.8], SATplan=1	high
336	al2_none=0, emph_m=(3.755], SATplan=1, m_selfcpt=(7.459.94]	high
337	BYSES=(0.0731.32], geo_1=1, al2_none=0, emph_m=(2.53.75]	high
338	BYSES=(0.0731.32], Pexp=MA or MORE, geo_1=1, white=1	high
339	Pexp=4YR DEG, BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94], female=0	high
340	Pexp=MA or MORE, BYTXMIRR=(23.631.8], SATplan=1, female=0	high
341	Pexp=MA or MORE, gm_none=1, m_selfcpt=(7.459.94], female=0	high
342	BYSES=(0.0731.32], geo_1=1, al2_1=1, white=1	high
343	BYGRADS=(3.124], emph_m=(3.755], t_rspnsv=(5.847.8], GPA910_m=(3.124]	high
344	BYGRADS=(3.124], al2_none=0, emph_m=(3.755], white=1	high
345	BYSES=(0.0731.32], BYGRADS=(3.124], al2_none=0, BYTXMIRR=(23.631.8]	high
346	Sexp=MA or MORE, gm_none=1, geo_1=1, al2_none=0	high
347	geo_1=1, SATplan=1, m_selfcpt=(7.459.94], white=1	high
348	BYSES=(0.0731.32], Pexp=MA or MORE, geo_none=0, female=0	high
349	Pexp=MA or MORE, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], white=1	high
350	geo_1=1, al2_half=0, BYTXMIRR=(23.631.8], GPA910_m=(3.124]	high
351	BYSES=(0.0731.32], BYTXMIRR=(23.631.8], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
352	BYSES=(0.0731.32], Pexp=MA or MORE, BYGRADS=(3.124], female=0	high
353	geo_1=1, al2_1=1, emph_m=(3.755], hispanic=0	high
354	Sexp=MA or MORE, emph_m=(3.755], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
355	gm_none=1, geo_1=1, GPA910_m=(3.124], female=0	high
356	BYGRADS=(3.124], BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94], white=1	high
357	BYGRADS=(3.124], geo_1=1, GPA910_m=(3.124], black=0	high
358	Pexp=MA or MORE, BYGRADS=(3.124], t_rspnsv=(3.895.84], GPA910_m=(3.124]	high
359	BYSES=(0.0731.32], BYTXMIRR=(23.631.8], GPA910_m=(3.124], female=0	high
360	BYSES=(0.0731.32], gm_none=1, geo_1=1, al2_1=1	high
361	BYSES=(0.0731.32], Sexp=MA or MORE, GPA910_m=(3.124], white=1	high
362	BYGRADS=(3.124], t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], GPA910_m=(3.124]	high
363	BYSES=(0.0731.32], BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94], black=0	high
364	Sexp=MA or MORE, geo_1=1, emph_m=(3.755], female=0	high
365	Pexp=MA or MORE, BYGRADS=(3.124], geo_1=1, emph_m=(3.755]	high
366	geo_1=1, emph_m=(3.755], m_selfcpt=(7.459.94], black=0	high

Table 57 continued

367	BYTXMIRR=(23.631.8], SATplan=1, m_selfcpt=(7.459.94], female=0	high
368	geo_none=0, GPA910_m=(3.124], m_selfcpt=(7.459.94], black=0	high
369	BYGRADS=(3.124], BYTXMIRR=(23.631.8], GPA910_m=(3.124], female=0	high
370	Pexp=MA or MORE, BYGRADS=(3.124], geo_none=0, white=1	high
371	Pexp=MA or MORE, geo_1=1, emph_m=(3.755], white=1	high
372	BYSES=(0.0731.32], Pexp=MA or MORE, BYGRADS=(3.124], emph_m=(2.53.75]	high
373	BYSES=(0.0731.32], geo_1=1, al2_none=0, white=1	high
374	Pexp=MA or MORE, emph_m=(3.755], female=0, white=1	high
375	BYSES=(0.0731.32], Sexp=MA or MORE, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8]	high
376	al2_1=1, emph_m=(2.53.75], GPA910_m=(3.124], white=1	high
377	geo_none=0, emph_m=(3.755], m_selfcpt=(7.459.94], white=1	high
378	gm_none=1, geo_none=0, al2_1=1, t_rspnsv=(5.847.8]	high
379	Sexp=MA or MORE, BYGRADS=(3.124], GPA910_m=(3.124], female=0	high
380	BYSES=(0.0731.32], emph_m=(3.755], m_selfcpt=(7.459.94], female=0	high
381	Sexp=MA or MORE, geo_none=0, emph_m=(3.755], female=0	high
382	BYSES=(0.0731.32], Pexp=MA or MORE, BYGRADS=(3.124], white=1	high
383	BYSES=(0.0731.32], geo_none=0, al2_none=0, emph_m=(3.755]	high
384	geo_1=1, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94]	high
385	gm_none=1, al2_1=1, SATplan=1, m_selfcpt=(7.459.94]	high
386	BYSES=(0.0731.32], gm_none=1, geo_none=0, al2_1=1	high
387	BYSES=(0.0731.32], geo_none=0, al2_1=1, female=1	high
388	BYGRADS=(3.124], emph_m=(3.755], GPA910_m=(3.124], white=1	high
389	geo_none=0, SATplan=1, m_selfcpt=(7.459.94], white=1	high
390	Sexp=MA or MORE, geo_none=0, al2_none=0, white=1	high
391	Pexp=MA or MORE, BYGRADS=(3.124], geo_1=1, black=0	high
392	BYSES=(0.0731.32], geo_none=0, al2_none=0, emph_m=(2.53.75]	high
393	geo_1=1, GPA910_m=(3.124], SATplan=1, white=1	high
394	Pexp=MA or MORE, BYTXMIRR=(23.631.8], female=0, hispanic=0	high
395	BYSES=(0.0731.32], Sexp=MA or MORE, al2_none=0, emph_m=(3.755]	high
396	BYGRADS=(3.124], al2_1=1, SATplan=1, white=1	high
397	Pexp=MA or MORE, t_rspnsv=(3.895.84], GPA910_m=(3.124], black=0	high
398	BYSES=(0.0731.32], Pexp=MA or MORE, geo_none=0, white=1	high
399	Sexp=MA or MORE, gm_1=0, al2_1=1, emph_m=(3.755]	high
400	gm_1=0, al2_1=1, m_selfcpt=(7.459.94], white=1	high
401	Sexp=MA or MORE, BYGRADS=(3.124], geo_1=1, t_rspnsv=(5.847.8]	high
402	Pexp=MA or MORE, gm_2=0, BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94]	high
403	al2_none=0, GPA910_m=(3.124], female=0, white=1	high
404	geo_none=0, emph_m=(3.755], m_selfcpt=(7.459.94], black=0	high
405	Pexp=MA or MORE, geo_1=1, emph_m=(3.755], SATplan=1	high
406	geo_1=1, t_rspnsv=(3.895.84], m_selfcpt=(7.459.94], white=1	high

Table 57 continued

407	gm_none=1, geo_1=1, GPA910_m=(3.124], white=1	high
408	Pexp=MA or MORE, geo_1=1, female=0, white=1	high
409	BYSES=(0.0731.32], gm_none=1, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
410	BYSES=(0.0731.32], Sexp=MA or MORE, al2_1=1, SATplan=1	high
411	BYSES=(0.0731.32], Pexp=MA or MORE, BYGRADS=(3.124], emph_m=(3.755]	high
412	BYGRADS=(3.124], emph_m=(3.755], GPA910_m=(3.124], female=0	high
413	Sexp=MA or MORE, emph_m=(3.755], GPA910_m=(3.124], white=1	high
414	gm_none=1, geo_1=1, al2_1=1, white=1	high
415	Pexp=MA or MORE, geo_1=1, SATplan=1, white=1	high
416	gm_1=0, al2_1=1, m_selfcpt=(7.459.94], female=0	high
417	BYTXMIRR=(23.631.8], GPA910_m=(3.124], m_selfcpt=(7.459.94], female=0	high
418	BYGRADS=(3.124], GPA910_m=(3.124], m_selfcpt=(7.459.94], female=0	high
419	gm_none=1, geo_1=1, m_selfcpt=(7.459.94], white=1	high
420	Pexp=MA or MORE, BYGRADS=(3.124], geo_1=1, SATplan=1	high
421	Sexp=MA or MORE, BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94], white=1	high
422	BYGRADS=(3.124], al2_1=1, BYTXMIRR=(23.631.8], white=1	high
423	geo_1=1, SATplan=1, m_selfcpt=(7.459.94], black=0	high
424	al2_1=1, GPA910_m=(3.124], female=0, hispanic=0	high
425	Sexp=MA or MORE, al2_2=0, BYTXMIRR=(23.631.8], female=0	high
426	Sexp=MA or MORE, BYGRADS=(3.124], SATplan=1, m_selfcpt=(7.459.94]	high
427	BYSES=(0.0731.32], Pexp=MA or MORE, geo_1=1, hispanic=0	high
428	BYSES=(0.0731.32], BYGRADS=(3.124], BYTXMIRR=(23.631.8], female=0	high
429	BYSES=(0.0731.32], gm_none=1, geo_none=0, al2_none=0	high
430	BYSES=(0.0731.32], geo_none=0, al2_none=0, BYTXMIRR=(23.631.8]	high
431	BYGRADS=(3.124], geo_1=1, emph_m=(3.755], female=0	high
432	Sexp=MA or MORE, al2_1=1, t_rspnsv=(3.895.84], white=1	high
433	BYGRADS=(3.124], t_rspnsv=(5.847.8], m_selfcpt=(7.459.94], white=1	high
434	Pexp=MA or MORE, BYGRADS=(3.124], female=0, white=1	high
435	Sexp=MA or MORE, geo_1=1, emph_m=(3.755], white=1	high
436	Pexp=MA or MORE, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], black=0	high
437	BYSES=(0.0731.32], Sexp=MA or MORE, BYGRADS=(3.124], geo_1=1	high
438	BYSES=(0.0731.32], BYGRADS=(3.124], geo_1=1, emph_m=(3.755]	high
439	Pexp=MA or MORE, Sexp=MA or MORE, geo_1=1, white=1	high
440	geo_1=1, al2_1=1, SATplan=1, white=1	high
441	Pexp=4YR DEG, emph_m=(3.755], m_selfcpt=(7.459.94], female=0	high
442	BYSES=(0.0731.32], BYGRADS=(3.124], BYTXMIRR=(23.631.8], SATplan=1	high
443	BYGRADS=(3.124], geo_1=1, t_rspnsv=(5.847.8], white=1	high
444	geo_1=1, SATplan=1, m_selfcpt=(7.459.94], female=0	high

Table 57 continued

445	BYGRADS=(3.124], t_rspnsv=(5.847.8], GPA910_m=(3.124], white=1	high
446	gm_none=1, geo_1=1, al2_none=0, white=1	high
447	BYGRADS=(3.124], geo_1=1, BYTXMIRR=(23.631.8], female=0	high
448	BYGRADS=(3.124], al2_none=0, female=0, hispanic=0	high
449	BYGRADS=(3.124], emph_m=(3.755], GPA910_m=(3.124], hispanic=0	high
450	BYSES=(0.0731.32], emph_m=(3.755], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
451	gm_none=1, al2_none=0, m_selfcpt=(7.459.94], white=1	high
452	Sexp=MA or MORE, al2_1=1, t_rspnsv=(3.895.84], black=0	high
453	BYSES=(0.0731.32], BYGRADS=(3.124], geo_none=0, BYTXMIRR=(23.631.8]	high
454	emph_m=(3.755], t_rspnsv=(5.847.8], GPA910_m=(3.124], black=0	high
455	Sexp=4YR DEG, geo_1=1, al2_none=0, emph_m=(3.755]	high
456	gm_none=1, geo_none=0, al2_1=1, white=1	high
457	BYSES=(0.0731.32], gm_none=1, SATplan=1, m_selfcpt=(7.459.94]	high
458	emph_m=(3.755], t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], GPA910_m=(3.124]	high
459	Sexp=MA or MORE, BYGRADS=(3.124], emph_m=(3.755], female=0	high
460	BYGRADS=(3.124], SATplan=1, m_selfcpt=(7.459.94], white=1	high
461	BYGRADS=(3.124], GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
462	Pexp=MA or MORE, BYGRADS=(3.124], geo_none=0, t_rspnsv=(3.895.84]	high
463	Sexp=MA or MORE, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], white=1	high
464	al2_1=1, GPA910_m=(3.124], hispanic=0, black=0	high
465	Sexp=MA or MORE, geo_1=1, emph_m=(3.755], SATplan=1	high
466	BYGRADS=(3.124], SATplan=1, m_selfcpt=(7.459.94], female=0	high
467	BYSES=(0.0731.32], BYGRADS=(3.124], gm_none=1, BYTXMIRR=(23.631.8]	high
468	BYSES=(0.0731.32], BYGRADS=(3.124], BYTXMIRR=(23.631.8], white=1	high
469	geo_none=0, al2_none=0, t_rspnsv=(5.847.8], white=1	high
470	BYSES=(0.0731.32], Sexp=MA or MORE, BYGRADS=(3.124], geo_none=0	high
471	BYGRADS=(3.124], geo_none=0, emph_m=(3.755], female=0	high
472	BYSES=(0.0731.32], BYGRADS=(3.124], emph_m=(3.755], BYTXMIRR=(23.631.8]	high
473	Sexp=MA or MORE, SATplan=1, m_selfcpt=(7.459.94], white=1	high
474	al2_1=1, emph_m=(3.755], SATplan=1, female=0	high
475	BYGRADS=(3.124], GPA910_m=(3.124], SATplan=1, white=1	high
476	BYTXMIRR=(23.631.8], GPA910_m=(3.124], SATplan=1, m_selfcpt=(7.459.94]	high
477	gm_none=1, geo_1=1, m_selfcpt=(7.459.94], female=0	high
478	Pexp=4YR DEG, emph_m=(3.755], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
479	Pexp=MA or MORE, Sexp=MA or MORE, BYGRADS=(3.124], white=1	high
480	Sexp=MA or MORE, BYTXMIRR=(23.631.8], GPA910_m=(3.124], black=0	high
481	BYGRADS=(3.124], geo_none=0, BYTXMIRR=(23.631.8], female=0	high

Table 57 continued

482	BYSES=(0.0731.32], Pexp=MA or MORE, Sexp=MA or MORE, female=0	high
483	BYTXMIRR=(23.631.8], GPA910_m=(3.124], SATplan=1, female=0	high
484	emph_m=(3.755], GPA910_m=(3.124], SATplan=1, m_selfcpt=(7.459.94]	high
485	BYSES=(0.0731.32], BYTXMIRR=(23.631.8], GPA910_m=(3.124], white=1	high
486	Sexp=MA or MORE, GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
487	BYGRADS=(3.124], gm_none=1, al2_1=1, white=1	high
488	BYGRADS=(3.124], gm_none=1, GPA910_m=(3.124], female=0	high
489	BYGRADS=(3.124], gm_none=1, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
490	Sexp=MA or MORE, geo_1=1, t_rspnsv=(5.847.8], black=0	high
491	Pexp=MA or MORE, BYGRADS=(3.124], SATplan=1, white=1	high
492	BYGRADS=(3.124], gm_none=1, m_selfcpt=(7.459.94], white=1	high
493	Sexp=MA or MORE, geo_none=0, emph_m=(3.755], white=1	high
494	BYSES=(0.0731.32], Sexp=MA or MORE, geo_1=1, BYTXMIRR=(23.631.8]	high
495	BYSES=(1.322.56], geo_none=0, grad_eff=Yes Sure will grad, SATplan=1	high
496	t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
497	Sexp=MA or MORE, GPA910_m=(3.124], SATplan=1, white=1	high
498	BYSES=(0.0731.32], SATplan=1, m_selfcpt=(7.459.94], female=0	high
499	Sexp=MA or MORE, BYGRADS=(3.124], geo_1=1, female=0	high
500	BYGRADS=(3.124], geo_1=1, emph_m=(1.252.5], BYTXMIRR=(23.631.8]	high
501	Sexp=MA or MORE, geo_none=0, t_rspnsv=(5.847.8], white=1	high
502	Sexp=MA or MORE, BYGRADS=(3.124], geo_1=1, white=1	high
503	geo_none=0, geo_2=0, GPA910_m=(3.124], white=1	high
504	BYSES=(1.322.56], geo_none=0, grad_eff=Yes Sure will grad, black=0	high
505	Sexp=MA or MORE, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], black=0	high
506	gm_none=1, geo_none=0, al2_none=0, white=1	high
507	Pexp=MA or MORE, Sexp=MA or MORE, geo_1=1, black=0	high
508	BYSES=(0.0731.32], geo_1=1, t_rspnsv=(5.847.8], white=1	high
509	Sexp=4YR DEG, geo_1=1, al2_none=0, white=1	high
510	Sexp=MA or MORE, geo_none=0, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8]	high
511	BYGRADS=(3.124], t_rspnsv=(3.895.84], m_selfcpt=(7.459.94], white=1	high
512	BYSES=(0.0731.32], Pexp=MA or MORE, BYTXMIRR=(23.631.8], white=1	high
513	al2_none=0, emph_m=(3.755], SATplan=1, female=0	high
514	BYSES=(0.0731.32], BYGRADS=(3.124], geo_none=0, emph_m=(3.755]	high
515	Pexp=4YR DEG, BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94], white=1	high
516	BYSES=(0.0731.32], geo_1=1, t_rspnsv=(5.847.8], hispanic=0	high
517	Sexp=MA or MORE, geo_1=1, t_rspnsv=(5.847.8], SATplan=1	high
518	BYSES=(0.0731.32], SATplan=1, m_selfcpt=(7.459.94], white=1	high
519	BYSES=(0.0731.32], BYGRADS=(3.124], geo_1=1, white=1	high
520	BYSES=(0.0731.32], gm_none=1, t_rspnsv=(5.847.8], GPA910_m=(3.124]	high
521	Pexp=MA or MORE, SATplan=1, female=0, white=1	high

Table 57 continued

522	geo_1=1, emph_m=(3.755], t_rspnsv=(5.847.8], white=1	high
523	BYSES=(0.0731.32], Pexp=MA or MORE, SATplan=1, female=0	high
524	BYSES=(0.0731.32], t_rspnsv=(3.895.84], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
525	BYSES=(0.0731.32], BYGRADS=(3.124], geo_1=1, female=0	high
526	Pexp=MA or MORE, Sexp=MA or MORE, t_rspnsv=(3.895.84], white=1	high
527	BYGRADS=(3.124], emph_m=(2.53.75], GPA910_m=(3.124], female=0	high
528	Pexp=MA or MORE, emph_m=(3.755], SATplan=1, white=1	high
529	BYGRADS=(3.124], BYTXMIRR=(23.631.8], SATplan=1, female=0	high
530	gm_none=1, geo_1=1, al2_none=0, SATplan=1	high
531	Pexp=MA or MORE, Sexp=MA or MORE, emph_m=(3.755], white=1	high
532	geo_none=0, al2_none=0, SATplan=1, white=1	high
533	Sexp=MA or MORE, geo_none=0, t_rspnsv=(5.847.8], black=0	high
534	BYSES=(0.0731.32], geo_1=1, al2_half=0, t_rspnsv=(5.847.8]	high
535	BYSES=(0.0731.32], geo_none=0, t_rspnsv=(5.847.8], white=1	high
536	t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], SATplan=1, m_selfcpt=(7.459.94]	high
537	BYSES=(0.0731.32], Sexp=MA or MORE, geo_1=1, female=0	high
538	al2_2=0, BYTXMIRR=(23.631.8], GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
539	BYSES=(0.0731.32], al2_1=1, BYTXMIRR=(23.631.8], SATplan=1	high
540	gm_none=1, BYTXMIRR=(23.631.8], SATplan=1, m_selfcpt=(7.459.94]	high
541	BYGRADS=(3.124], geo_none=0, SATplan=1, female=0	high
542	BYSES=(0.0731.32], BYGRADS=(3.124], geo_none=0, female=0	high
543	Sexp=4YR DEG, BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94], female=0	high
544	BYSES=(0.0731.32], BYGRADS=(3.124], geo_none=0, SATplan=1	high
545	BYGRADS=(3.124], geo_1=1, emph_m=(2.53.75], BYTXMIRR=(23.631.8]	high
546	BYSES=(0.0731.32], Sexp=MA or MORE, BYGRADS=(3.124], female=0	high
547	BYSES=(0.0731.32], Sexp=MA or MORE, geo_1=1, white=1	high
548	Sexp=MA or MORE, BYGRADS=(3.124], geo_1=1, SATplan=1	high
549	BYGRADS=(3.124], al2_2=0, emph_m=(2.53.75], BYTXMIRR=(23.631.8]	high
550	BYSES=(0.0731.32], Sexp=MA or MORE, geo_1=1, SATplan=1	high
551	BYGRADS=(3.124], geo_none=0, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8]	high
552	BYGRADS=(3.124], geo_none=0, t_rspnsv=(5.847.8], female=0	high
553	Sexp=MA or MORE, gm_none=1, m_selfcpt=(7.459.94], black=0	high
554	BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94], female=0, black=0	high
555	Sexp=4YR DEG, gm_none=1, BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94]	high
556	BYSES=(0.0731.32], Sexp=MA or MORE, BYTXMIRR=(23.631.8], white=1	high
557	gm_none=1, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], GPA910_m=(3.124]	high
558	BYGRADS=(3.124], geo_1=1, t_rspnsv=(5.847.8], black=0	high
559	t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], GPA910_m=(3.124], SATplan=1	high

Table 57 continued

560	Pexp=MA or MORE, BYGRADS=(3.124], SATplan=1, black=0	high
561	BYSES=(0.0731.32], geo_1=1, BYTXMIRR=(23.631.8], female=0	high
562	BYSES=(0.0731.32], gm_1=0, al2_1=1, emph_m=(3.755]	high
563	Sexp=4YR DEG, BYTXMIRR=(23.631.8], GPA910_m=(3.124], female=0	high
564	BYSES=(0.0731.32], gm_none=1, t_rspnsv=(3.895.84], m_selfcpt=(7.459.94]	high
565	Sexp=MA or MORE, BYGRADS=(3.124], emph_m=(3.755], white=1	high
566	BYSES=(0.0731.32], Pexp=MA or MORE, Sexp=MA or MORE, gm_none=1	high
567	geo_none=0, al2_2=0, t_rspnsv=(5.847.8], BYTXMIRR=(23.631.8]	high
568	BYTXMIRR=(23.631.8], GPA910_m=(3.124], SATplan=1, white=1	high
569	BYGRADS=(3.124], emph_m=(2.53.75], BYTXMIRR=(23.631.8], white=1	high
570	BYSES=(0.0731.32], Sexp=MA or MORE, BYTXMIRR=(23.631.8], black=0	high
571	BYGRADS=(3.124], t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], SATplan=1	high
572	BYSES=(0.0731.32], gm_none=1, GPA910_m=(3.124], SATplan=1	high
573	Pexp=MA or MORE, Sexp=4YR DEG, BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad	high
574	BYGRADS=(3.124], BYTXMIRR=(23.631.8], female=0, black=0	high
575	BYSES=(0.0731.32], Sexp=MA or MORE, BYGRADS=(3.124], black=0	high
576	BYTXMIRR=(23.631.8], GPA910_m=(3.124], SATplan=1, black=0	high
577	gm_none=1, GPA910_m=(3.124], SATplan=1, m_selfcpt=(7.459.94]	high
578	Sexp=4YR DEG, BYGRADS=(3.124], geo_none=0, BYTXMIRR=(23.631.8]	high
579	Pexp=MA or MORE, Sexp=MA or MORE, emph_m=(3.755], black=0	high
580	Sexp=MA or MORE, geo_1=1, female=0, white=1	high
581	Sexp=MA or MORE, gm_none=1, emph_m=(3.755], female=0	high
582	BYGRADS=(3.124], geo_1=1, BYTXMIRR=(23.631.8], white=1	high
583	BYSES=(0.0731.32], geo_2=0, emph_m=(3.755], GPA910_m=(3.124]	high
584	Sexp=MA or MORE, BYGRADS=(3.124], geo_none=0, black=0	high
585	geo_none=0, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], female=0	high
586	al2_none=0, al2_2=0, emph_m=(2.53.75], BYTXMIRR=(23.631.8]	high
587	BYGRADS=(3.124], geo_1=1, SATplan=1, white=1	high
588	BYTXMIRR=(23.631.8], GPA910_m=(3.124], female=0, white=1	high
589	Pexp=MA or MORE, emph_m=(3.755], female=0, hispanic=0	high
590	BYSES=(0.0731.32], geo_1=1, emph_m=(3.755], female=0	high
591	Pexp=MA or MORE, Sexp=MA or MORE, SATplan=1, white=1	high
592	emph_m=(3.755], SATplan=1, m_selfcpt=(7.459.94], female=0	high
593	al2_none=0, emph_m=(3.755], female=0, white=1	high
594	Pexp=MA or MORE, gm_none=1, female=0, white=1	high
595	geo_1=1, emph_m=(3.755], SATplan=1, female=0	high
596	Pexp=MA or MORE, BYGRADS=(3.124], t_rspnsv=(5.847.8], hispanic=0	high
597	al2_1=1, emph_m=(3.755], SATplan=1, hispanic=0	high
598	t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], m_selfcpt=(7.459.94], white=1	high
599	Sexp=4YR DEG, BYGRADS=(3.124], gm_none=1, BYTXMIRR=(23.631.8]	high

Table 57 continued

600	Pexp=MA or MORE, geo_none=0, SATplan=1, black=0	high
601	BYSES=(0.0731.32], BYGRADS=(3.124], SATplan=1, female=0	high
602	emph_m=(3.755], SATplan=1, m_selfcpt=(7.459.94], white=1	high
603	al2_none=0, emph_m=(3.755], SATplan=1, white=1	high
604	t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], GPA910_m=(3.124], white=1	high
605	GPA910_m=(3.124], SATplan=1, m_selfcpt=(7.459.94], female=0	high
606	BYSES=(0.0731.32], Sexp=MA or MORE, gm_none=1, emph_m=(3.755]	high
607	GPA910_m=(3.124], SATplan=1, m_selfcpt=(7.459.94], white=1	high
608	BYSES=(0.0731.32], geo_1=1, emph_m=(3.755], BYTXMIRR=(23.631.8]	high
609	Sexp=MA or MORE, BYGRADS=(3.124], geo_1=1, al2_half=0	high
610	Sexp=MA or MORE, geo_none=0, female=0, white=1	high
611	BYSES=(0.0731.32], al2_none=0, BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad	high
612	BYGRADS=(3.124], geo_1=1, emph_m=(3.755], white=1	high
613	al2_1=1, BYTXMIRR=(23.631.8], SATplan=1, female=0	high
614	BYGRADS=(3.124], geo_none=0, SATplan=1, white=1	high
615	Pexp=4YR DEG, emph_m=(3.755], GPA910_m=(3.124], grad_eff=Yes Sure will grad	high
616	geo_1=1, BYTXMIRR=(23.631.8], SATplan=1, female=0	high
617	BYSES=(0.0731.32], geo_none=0, BYTXMIRR=(23.631.8], female=0	high
618	geo_none=0, emph_m=(2.53.75], BYTXMIRR=(23.631.8], female=0	high
619	BYGRADS=(3.124], t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], black=0	high
620	Sexp=MA or MORE, geo_1=1, BYTXMIRR=(23.631.8], black=0	high
621	geo_1=1, t_rspnsv=(5.847.8], SATplan=1, white=1	high
622	gm_none=1, SATplan=1, m_selfcpt=(7.459.94], white=1	high
623	BYSES=(0.0731.32], BYGRADS=(3.124], gm_none=1, female=0	high
624	Sexp=MA or MORE, BYGRADS=(3.124], SATplan=1, white=1	high
625	BYSES=(0.0731.32], al2_1=1, SATplan=1, female=0	high
626	BYSES=(0.0731.32], al2_none=0, emph_m=(2.53.75], white=1	high
627	emph_m=(3.755], GPA910_m=(3.124], grad_eff=Yes Sure will grad, SATplan=1	high
628	BYSES=(0.0731.32], geo_none=0, emph_m=(3.755], BYTXMIRR=(23.631.8]	high
629	BYSES=(0.0731.32], gm_none=1, al2_1=1, SATplan=1	high
630	BYSES=(0.0731.32], BYGRADS=(3.124], geo_half=0, emph_m=(3.755]	high
631	Pexp=MA or MORE, BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad, hispanic=0	high
632	gm_none=1, al2_1=1, emph_m=(3.755], SATplan=1	high
633	geo_1=1, emph_m=(3.755], female=0, white=1	high
634	BYSES=(0.0731.32], geo_1=1, SATplan=1, female=0	high
635	Pexp=4YR DEG, GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
636	gm_none=1, al2_half=0, t_rspnsv=(5.847.8], BYTXMIRR=(23.631.8]	high
637	BYSES=(0.0731.32], gm_1=0, t_rspnsv=(3.895.84], m_selfcpt=(7.459.94]	high
638	BYSES=(0.0731.32], gm_none=1, geo_1=1, emph_m=(3.755]	high

Table 57 continued

639	BYSES=(0.0731.32], geo_none=0, emph_m=(2.53.75], BYTXMIRR=(23.631.8]	high
640	BYGRADS=(3.124], gm_none=1, geo_none=0, female=0	high
641	BYSES=(0.0731.32], BYGRADS=(3.124], emph_m=(3.755], SATplan=1	high
642	BYSES=(0.0731.32], gm_none=1, GPA910_m=(3.124], grad_eff=Yes Sure will grad	high
643	BYSES=(0.0731.32], Sexp=MA or MORE, emph_m=(3.755], t_rspnsv=(3.895.84]	high
644	Sexp=MA or MORE, gm_none=1, t_rspnsv=(3.895.84], m_selfcpt=(7.459.94]	high
645	t_rspnsv=(5.847.8], BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad, SATplan=1	high
646	BYSES=(0.0731.32], BYGRADS=(3.124], emph_m=(3.755], white=1	high
647	BYGRADS=(3.124], geo_1=1, SATplan=1, black=0	high
648	geo_1=1, BYTXMIRR=(23.631.8], female=0, black=0	high
649	BYGRADS=(3.124], gm_none=1, geo_1=1, white=1	high
650	al2_1=1, BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad, SATplan=1	high
651	gm_none=1, geo_1=1, emph_m=(3.755], female=0	high
652	Sexp=MA or MORE, gm_none=1, geo_1=1, SATplan=1	high
653	BYSES=(0.0731.32], t_rspnsv=(3.895.84], m_selfcpt=(7.459.94], white=1	high
654	geo_none=0, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], SATplan=1	high
655	BYTXMIRR=(23.631.8], GPA910_m=(3.124], grad_eff=Yes Sure will grad, black=0	high
656	BYGRADS=(3.124], geo_1=1, al2_2=0, white=1	high
657	BYSES=(0.0731.32], geo_1=1, BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad	high
658	gm_none=1, GPA910_m=(3.124], SATplan=1, white=1	high
659	gm_none=1, GPA910_m=(3.124], m_selfcpt=(7.459.94], female=0	high
660	BYSES=(0.0731.32], BYTXMIRR=(23.631.8], SATplan=1, female=0	high
661	BYSES=(0.0731.32], geo_1=1, al2_2=0, BYTXMIRR=(23.631.8]	high
662	gm_none=1, emph_m=(3.755], m_selfcpt=(7.459.94], female=0	high
663	Pexp=MA or MORE, SATplan=1, female=0, black=0	high
664	gm_none=1, GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
665	gm_1=0, geo_1=1, emph_m=(3.755], female=0	high
666	emph_m=(2.53.75], BYTXMIRR=(23.631.8], SATplan=1, female=0	high
667	BYSES=(0.0731.32], geo_1=1, SATplan=1, white=1	high
668	Sexp=MA or MORE, gm_none=1, emph_m=(3.755], white=1	high
669	emph_m=(3.755], BYTXMIRR=(23.631.8], SATplan=1, female=0	high
670	gm_none=1, SATplan=1, m_selfcpt=(7.459.94], black=0	high
671	BYSES=(0.0731.32], BYGRADS=(3.124], gm_none=1, white=1	high
672	geo_1=1, emph_m=(2.53.75], BYTXMIRR=(23.631.8], white=1	high
673	BYGRADS=(3.124], emph_m=(3.755], female=0, white=1	high
674	emph_m=(2.53.75], t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], female=0	high
675	gm_none=1, GPA910_m=(3.124], SATplan=1, female=0	high
676	gm_none=1, SATplan=1, m_selfcpt=(7.459.94], female=0	high

Table 57 continued

677	Pexp=MA or MORE, gm_none=1, female=0, black=0	high
678	geo_none=0, emph_m=(2.53.75], BYTXMIRR=(23.631.8], white=1	high
679	Pexp=4YR DEG, al2_none=0, BYTXMIRR=(23.631.8], white=1	high
680	al2_1=1, emph_m=(3.755], hispanic=0, black=0	high
681	BYSES=(0.0731.32], geo_1=1, emph_m=(2.53.75], white=1	high
682	BYSES=(0.0731.32], geo_none=0, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8]	high
683	geo_none=0, emph_m=(3.755], t_rspnsv=(5.847.8], black=0	high
684	geo_1=1, BYTXMIRR=(23.631.8], SATplan=1, black=0	high
685	geo_1=1, emph_m=(3.755], SATplan=1, white=1	high
686	geo_1=1, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], white=1	high
687	BYSES=(0.0731.32], BYGRADS=(3.124], al2_2=0, female=0	high
688	geo_none=0, al2_2=0, emph_m=(2.53.75], BYTXMIRR=(23.631.8]	high
689	BYSES=(0.0731.32], gm_none=1, geo_1=1, female=0	high
690	geo_none=0, emph_m=(1.252.5], t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8]	high
691	SATplan=1, m_selfcpt=(7.459.94], female=0, white=1	high
692	gm_none=1, al2_none=0, emph_m=(3.755], white=1	high
693	t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], SATplan=1, female=0	high
694	Sexp=4YR DEG, gm_none=1, GPA910_m=(3.124], m_selfcpt=(7.459.94]	high
695	geo_none=0, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], white=1	high
696	BYTXMIRR=(23.631.8], SATplan=1, female=0, white=1	high
697	BYSES=(0.0731.32], geo_1=1, SATplan=1, black=0	high
698	BYSES=(0.0731.32], al2_1=1, grad_eff=Yes Sure will grad, white=1	high
699	Pexp=MA or MORE, gm_none=1, grad_eff=Yes Sure will grad, white=1	high
700	BYSES=(0.0731.32], gm_none=1, al2_1=1, white=1	high
701	BYSES=(0.0731.32], Sexp=MA or MORE, gm_none=1, SATplan=1	high
702	BYSES=(0.0731.32], Sexp=MA or MORE, SATplan=1, female=0	high
703	gm_none=1, emph_m=(2.53.75], BYTXMIRR=(23.631.8], SATplan=1	high
704	BYSES=(0.0731.32], gm_none=1, geo_1=1, white=1	high
705	Pexp=MA or MORE, gm_none=1, SATplan=1, female=0	high
706	geo_1=1, SATplan=1, female=0, white=1	high
707	Sexp=MA or MORE, BYGRADS=(3.124], SATplan=1, hispanic=0	high
708	geo_1=1, BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad, white=1	high
709	al2_1=1, BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad, black=0	high
710	gm_none=1, GPA910_m=(3.124], grad_eff=Yes Sure will grad, SATplan=1	high
711	Pexp=4YR DEG, gm_none=1, m_selfcpt=(7.459.94], white=1	high
712	BYGRADS=(3.124], emph_m=(3.755], SATplan=1, white=1	high
713	emph_m=(2.53.75], BYTXMIRR=(23.631.8], SATplan=1, white=1	high
714	gm_1=0, GPA910_m=(3.124], SATplan=1, white=1	high
715	gm_none=1, geo_1=1, SATplan=1, female=0	high
716	GPA910_m=(3.124], SATplan=1, female=0, white=1	high

Table 57 continued

717	BYSES=(0.0731.32], geo_2=0, GPA910_m=(3.124], white=1	high
718	BYSES=(0.0731.32], GPA910_m=(3.124], grad_eff=Yes Sure will grad, white=1	high
719	Sexp=MA or MORE, SATplan=1, female=0, white=1	high
720	BYSES=(0.0731.32], Sexp=MA or MORE, emph_m=(2.53.75], white=1	high
721	gm_none=1, al2_1=1, SATplan=1, female=0	high
722	BYSES=(0.0731.32], geo_none=0, emph_m=(3.755], grad_eff=Yes Sure will grad	high
723	geo_1=1, al2_half=0, BYTXMIRR=(23.631.8], white=1	high
724	BYSES=(0.0731.32], geo_1=1, al2_2=0, female=0	high
725	BYSES=(-1.170.073], emph_m=(3.755], BYTXMIRR=(23.631.8], SATplan=1	high
726	BYSES=(0.0731.32], Sexp=MA or MORE, gm_1=0, SATplan=1	high
727	BYSES=(0.0731.32], geo_1=1, grad_eff=Yes Sure will grad, white=1	high
728	BYGRADS=(3.124], SATplan=1, female=0, white=1	high
729	gm_none=1, al2_1=1, t_rspnsv=(5.847.8], grad_eff=Yes Sure will grad	high
730	BYSES=(0.0731.32], geo_none=0, SATplan=1, female=0	high
731	BYGRADS=(3.124], gm_none=1, SATplan=1, female=0	high
732	gm_none=1, emph_m=(2.53.75], t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8]	high
733	BYSES=(0.0731.32], gm_none=1, BYTXMIRR=(23.631.8], female=0	high
734	t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], SATplan=1, white=1	high
735	gm_none=1, geo_1=1, SATplan=1, white=1	high
736	BYSES=(0.0731.32], BYTXMIRR=(23.631.8], SATplan=1, white=1	high
737	BYSES=(0.0731.32], geo_none=0, emph_m=(2.53.75], white=1	high
738	Pexp=4YR DEG, gm_1=0, m_selfcpt=(7.459.94], white=1	high
739	geo_1=1, al2_half=0, emph_m=(1.252.5], BYTXMIRR=(23.631.8]	high
740	BYSES=(0.0731.32], geo_none=0, SATplan=1, white=1	high
741	emph_m=(2.53.75], t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], white=1	high
742	gm_none=1, emph_m=(3.755], t_rspnsv=(5.847.8], white=1	high
743	GPA910_m=(3.124], SATplan=1, female=0, black=0	high
744	Sexp=MA or MORE, gm_none=1, t_rspnsv=(5.847.8], white=1	high
745	BYGRADS=(3.124], gm_none=1, SATplan=1, white=1	high
746	emph_m=(2.53.75], BYTXMIRR=(23.631.8], female=0, white=1	high
747	Pexp=4YR DEG, grad_eff=Yes Sure will grad, m_selfcpt=(7.459.94], white=1	high
748	BYSES=(0.0731.32], t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], female=0	high
749	geo_1=1, emph_m=(3.755], grad_eff=Yes Sure will grad, white=1	high
750	gm_1=0, geo_1=1, emph_m=(3.755], white=1	high
751	BYSES=(0.0731.32], BYTXMIRR=(23.631.8], female=0, white=1	high
752	BYSES=(0.0731.32], Sexp=MA or MORE, t_rspnsv=(3.895.84], white=1	high
753	gm_none=1, t_rspnsv=(3.895.84], m_selfcpt=(7.459.94], white=1	high
754	geo_none=0, SATplan=1, female=0, white=1	high
755	Sexp=MA or MORE, gm_none=1, SATplan=1, white=1	high

Table 57 continued

756	BYSES=(0.0731.32], gm_none=1, geo_none=0, grad_eff=Yes Sure will grad	high
757	gm_none=1, geo_1=1, female=0, white=1	high
758	geo_none=0, emph_m=(3.755], SATplan=1, white=1	high
759	al2_none=0, t_rspnsv=(5.847.8], grad_eff=Yes Sure will grad, white=1	high
760	BYSES=(0.0731.32], t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], white=1	high
761	geo_none=0, t_rspnsv=(5.847.8], hispanic=0, black=0	high
762	BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad, SATplan=1, white=1	high
763	gm_none=1, al2_1=1, emph_m=(2.53.75], white=1	high
764	geo_1=1, emph_m=(3.755], hispanic=0, black=0	high
765	al2_2=0, emph_m=(2.53.75], BYTXMIRR=(23.631.8], white=1	high
766	gm_none=1, emph_m=(3.755], t_rspnsv=(5.847.8], SATplan=1	high
767	gm_none=1, geo_none=0, SATplan=1, female=0	high
768	BYGRADS=(3.124], gm_none=1, female=0, white=1	high
769	gm_none=1, al2_1=1, grad_eff=Yes Sure will grad, white=1	high
770	Sexp=4YR DEG, gm_none=1, GPA910_m=(3.124], female=0	high
771	gm_none=1, geo_none=0, SATplan=1, white=1	high
772	al2_none=0, t_rspnsv=(5.847.8], hispanic=0, black=0	high
773	gm_none=1, GPA910_m=(3.124], grad_eff=Yes Sure will grad, white=1	high
774	BYSES=(0.0731.32], gm_none=1, t_rspnsv=(5.847.8], SATplan=1	high
775	Sexp=4YR DEG, GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
776	BYSES=(0.0731.32], Sexp=MA or MORE, grad_eff=Yes Sure will grad, white=1	high
777	Sexp=4YR DEG, gm_none=1, BYTXMIRR=(23.631.8], female=0	high
778	Pexp=4YR DEG, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], female=0	high
779	t_rspnsv=(3.895.84], GPA910_m=(3.124], m_selfcpt=(7.459.94], white=1	high
780	gm_none=1, geo_1=1, female=0, black=0	high
781	t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], female=0, white=1	high
782	emph_m=(3.755], BYTXMIRR=(23.631.8], female=0, white=1	high
783	gm_none=1, geo_1=1, SATplan=1, hispanic=0	high
784	BYGRADS=(3.124], gm_none=1, SATplan=1, black=0	high
785	Pexp=4YR DEG, BYTXMIRR=(23.631.8], female=0, black=0	high
786	gm_none=1, al2_none=0, female=0, white=1	high
787	BYGRADS=(3.124], gm_none=1, t_rspnsv=(3.895.84], female=0	high
788	BYSES=(0.0731.32], gm_half=0, BYTXMIRR=(23.631.8], white=1	high
789	gm_none=1, geo_1=1, t_rspnsv=(3.895.84], white=1	high
790	Sexp=4YR DEG, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], female=0	high
791	emph_m=(3.755], t_rspnsv=(5.847.8], SATplan=1, white=1	high
792	gm_none=1, al2_1=1, grad_eff=Yes Sure will grad, female=0	high
793	geo_2=0, al2_1=1, grad_eff=Yes Sure will grad, white=1	high
794	BYGRADS=(3.124], gm_none=1, t_rspnsv=(3.895.84], white=1	high
795	gm_none=1, t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], white=1	high
796	gm_none=1, geo_1=1, grad_eff=Yes Sure will grad, female=0	high

Table 57 continued

797	gm_none=1, BYTXMIRR=(23.631.8], female=0, black=0	high
798	gm_none=1, GPA910_m=(3.124], grad_eff=Yes Sure will grad, black=0	high
799	gm_none=1, geo_1=1, grad_eff=Yes Sure will grad, white=1	high
800	gm_1=0, geo_1=1, female=0, black=0	high
801	emph_m=(3.755], SATplan=1, female=0, white=1	high
802	Sexp=MA or MORE, gm_none=1, grad_eff=Yes Sure will grad, white=1	high
803	Sexp=4YR DEG, gm_1=0, GPA910_m=(3.124], white=1	high
804	t_rspnsv=(3.895.84], BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad, white=1	high
805	gm_none=1, geo_1=1, hispanic=0, black=0	high
806	BYTXMIRR=(23.631.8], grad_eff=Yes Sure will grad, female=0, white=1	high
807	gm_none=1, emph_m=(3.755], SATplan=1, female=0	high
808	gm_none=1, geo_none=0, female=0, black=0	high
809	gm_none=1, geo_1=1, grad_eff=Yes Sure will grad, black=0	high
810	Sexp=MA or MORE, t_rspnsv=(3.895.84], female=0, white=1	high
811	Sexp=4YR DEG, GPA910_m=(3.124], female=0, white=1	high
812	BYSES=(0.0731.32], gm_none=1, emph_m=(3.755], grad_eff=Yes Sure will grad	high
813	Sexp=4YR DEG, GPA910_m=(3.124], grad_eff=Yes Sure will grad, white=1	high
814	BYSES=(0.0731.32], gm_none=1, SATplan=1, female=0	high
815	BYGRADS=(3.124], gm_2=0, t_rspnsv=(3.895.84], white=1	high
816	Sexp=4YR DEG, m_selfcpt=(7.459.94], female=0, white=1	high
817	BYGRADS=(3.124], grad_eff=Yes Sure will grad, female=0, white=1	high
818	geo_none=0, geo_2=0, female=0, white=1	high
819	BYSES=(0.0731.32], gm_none=1, emph_m=(3.755], SATplan=1	high
820	Pexp=4YR DEG, geo_none=0, emph_m=(2.53.75], female=0	high
821	BYSES=(0.0731.32], gm_1=0, t_rspnsv=(5.847.8], grad_eff=Yes Sure will grad	high
822	BYSES=(0.0731.32], gm_none=1, emph_m=(3.755], female=0	high
823	gm_none=1, grad_eff=Yes Sure will grad, m_selfcpt=(7.459.94], hispanic=0	high
824	grad_eff=Yes Sure will grad, SATplan=1, asian=1	high
825	gm_none=1, emph_m=(3.755], female=0, white=1	high
826	BYSES=(0.0731.32], gm_none=1, SATplan=1, white=1	high
827	gm_none=1, emph_m=(3.755], SATplan=1, white=1	high
828	BYSES=(0.0731.32], gm_1=0, SATplan=1, female=0	high
829	t_rspnsv=(3.895.84], GPA910_m=(3.124], grad_eff=Yes Sure will grad, white=1	high
830	BYSES=(0.0731.32], geo_2=0, SATplan=1, female=0	high
831	BYSES=(0.0731.32], emph_m=(3.755], grad_eff=Yes Sure will grad, female=0	high
832		low

Table 58. RIPPER ruleset (Study 2, 29 possible predictors)

Rule order	Antecedent	Math ach	Coverage (train/test)*	Confidence (train/test)*
1	(BYTXMIRR <= 23.211) and (BYTXMIRR <= 17.999)	Low	.31	.93
2	(BYTXMIRR <= 23.91) and (geo_none = 1)	Low	.12	.75
3	(BYTXMIRR <= 26.54) and (GPA910_m <= 2.791311) and (BYTXMIRR <= 23.211)	Low	.06	.64
4	(BYTXMIRR <= 24.794) and (m_selfcpt <= 6.785321) and (BYSES <= -0.294)	Low	.01	.64
5	(BYTXMIRR <= 25.307) and (BYTXMIRR <= 22.232) and (geo_1 = 0)	Low	.002	.68/.73
6	If none of the rules apply	High	.49	.86

Note: Coverage and support for test set is indicated separately only if the values differ by more than a percentage point.

Table 59. RIPPER ruleset (Study 2, 1933 possible predictors)

Rule order	Antecedent	Math achievement	Coverage (train/test)*	Confidence (train/test)*
1	(BYTXMIRR <= 23.211) and (BYTXMIRR <= 18.004)	Low	.31	.93
2	(BYTXMIRR <= 24.794) and (geo_none = 1) and (F2T3_4 = missing)	Low	.07	.82
3	(BYTXMIRR <= 25.314) and (GPA910_m <= 2.786712) and (BYTXMIRR <= 22.763)	Low	.07	.68
4	(BYTXMIRR <= 25.239) and (m_selfcpt <= 6.785321) and (BYTXMIRR <= 23.115) and (F2T3_4 = missing)	Low	.011/.007	.69/.64
5	(BYTXMIRR <= 25.239) and (geo_none = 1) and (al2_none = 1) and (BYTXMIRR <= 23.116)	Low	.02	.72/.69
6	If none of the rules apply	High	.51/.53	.85/.83

Note: Coverage and confidence for test set is indicated separately only if the values differ by more than a percentage point, or by 20 percent of the larger value.

Table 60. PART ruleset (Study 2, 29 possible predictors)

Rule order	Antecedent	Math ach	Coverage (train/test)*	Confidence (train/test)*
1	BYTXMIRR > 23.91	High	.44	.88
2	BYTXMIRR <= 17.999	Low	.31	.93
3	BYGRADS > 2.4 AND geo_none = 1	Low	.08	.70
4	BYGRADS > 2.4 AND GPA910_m <= 2.791311	Low	.05	.58
5	BYGRADS <= 2.6	Low	.06	.81/.82
6	If none of the rules apply	High	.04	.64

Note: Coverage and confidence for test set is indicated separately only if the values differ by more than a percentage point.

Table 61. PART ruleset (Study 2, 1933 possible predictors)

Rule order	Antecedent	Math ach	Coverage (train/test)*	Confidence (train/test)*
1	BYTXMIRR > 28.597 AND geo_none = 0 AND m_selfcpt > 6.751391	High	.16	.99
2	BYTXMIRR > 30.561 AND F1S20 = COLLEGE PREP	High	.05	.97
3	BYTXMIRR <= 18.004	Low	.31	.93
4	BYTXMIRR > 23.91 AND BYGRADS > 2.7 AND BYTXMIRR > 26.528 AND F2T3_16C = NOT A PROBLEM	High	.05	.94
5	BYTXMIRR > 23.91 AND BYGRADS > 3	High	.08	.82
6	BYTXMIRR <= 24.495 AND geo_none = 1	Low	.13	.74
7	BYTXMIRR > 24.495	High	.09	.70
8	m_selfcpt > 6.770048	High	.05	.63/.59
9	F2T3_16K = missing	Low	.04	.72/.71
10	If none of the rules apply	High	.04	.50/.52

Note: Coverage and confidence for test set is indicated separately only if the values differ by more than a percentage point.

Table 62. C4.5 ruleset (Study 2, 29 possible predictors)

Rule	Antecedent	Math ach	Coverage (train/test)*	Confidence (train/test)*
1	BYTXMIRR <= 23.876, BYTXMIRR <= 17.999	Low	.31	.93
2	BYTXMIRR <= 23.876, BYTXMIRR > 17.999, geo_none = 0, m_selfcpt <= 6.770048, BYSES <= -0.032	Low	.03	.63
3	BYTXMIRR <= 23.876, BYTXMIRR > 17.999, geo_none = 0, m_selfcpt <= 6.770048, BYSES > -0.032	High	.03	.51
4	BYTXMIRR <= 23.876, BYTXMIRR > 17.999, geo_none = 0, m_selfcpt > 6.770048	High	.04	.64
5	BYTXMIRR <= 23.876, BYTXMIRR > 17.999, geo_none = 1	Low	.12	.75
6	BYTXMIRR <= 23.876, BYTXMIRR > 17.999, geo_none = "missing"	Low	.02	.76
7	BYTXMIRR > 23.876	High	.44	.88

Table 63. C4.5 ruleset (Study 2, 1933 possible predictors)

Rule	Antecedent	Math ach	Coverage (train/test)*	Confidence (train/test)*
1	BYTXMIRR <= 23.91, BYTXMIRR <= 18.004	Low	.31	.93
2	BYTXMIRR <= 23.91, BYTXMIRR >18.004, BYGRADS <= 2.4	Low	.05	.82
3	BYTXMIRR <= 23.91, BYTXMIRR > 18.004, BYGRADS > 2.4, geo_none = 0, m_selfcpt <= 6.770048	Low	.05	.55/.53
4	BYTXMIRR <= 23.91, BYTXMIRR > 18.004, BYGRADS > 2.4, geo_none = 0, m_selfcpt > 6.770048	High	.04	.67/.64
5	BYTXMIRR <= 23.91, BYTXMIRR > 18.004, BYGRADS > 2.4, geo_none = 1	Low	.08	.70
6	BYTXMIRR <= 23.91, BYTXMIRR > 18.004, BYGRADS > 2.4, geo_none = missing	Low	.01	.68
7	BYTXMIRR > 23.91, BYTXMIRR <= 28.597	High	.15	.73
8	BYTXMIRR > 23.91, BYTXMIRR > 28.597, BYGRADS <= 2.8	High	.04	.87
9	BYTXMIRR > 23.91, BYTXMIRR > 28.597, BYGRADS > 2.8, m_selfcpt <= 6.751391, BYS8F = missing	Low	<.001	.78/.88
10	BYTXMIRR > 23.91, BYTXMIRR > 28.597, BYGRADS > 2.8, m_selfcpt <= 6.751391, BYS8F = NO	High	.03	.92
11	BYTXMIRR > 23.91, BYTXMIRR > 28.597, BYGRADS > 2.8, m_selfcpt <= 6.751391, BYS8F = YES	High	.04	.96/.94
12	BYTXMIRR > 23.91, BYTXMIRR > 28.597, BYGRADS > 2.8, m_selfcpt > 6.751391	High	.17	.99

Table 64. CART ruleset (Study 2, 29 possible predictors)

Rule	Antecedent	Math ach	Coverage*	Within 5 pts of estimate*	Within 10 pts of estimate*
1	BYTXMIRR>=23.85 BYTXMIRR>=30.51 BYTXMIRR>=35.26 BYGRADS>=3.6	69.74	.054	.71	.97
2	BYTXMIRR>=23.85 BYTXMIRR>=30.51 BYTXMIRR>=35.26 BYGRADS< 3.6	65.99	.031	.56	.90
3	BYTXMIRR>=23.85	63.75	.098	.61	.90

Table 64 continued

	BYTXMIRR>=30.51 BYTXMIRR< 35.26 m_selfcpt>=6.83				
4	BYTXMIRR>=23.85 BYTXMIRR< 30.51 m_selfcpt>=6.735 geo_none=0 BYTXMIRR>=28.16	60.66	.031	.56	.89
5	BYTXMIRR>=23.85 BYTXMIRR>=30.51 BYTXMIRR< 35.26 m_selfcpt< 6.83	58.34	.053	.48	.85
6	BYTXMIRR>=23.85 BYTXMIRR< 30.51 m_selfcpt>=6.735 geo_none=0 BYTXMIRR< 28.16	56.41	.048	.47	.84
7	BYTXMIRR>=23.85 BYTXMIRR< 30.51 m_selfcpt< 6.735 geo_none=0 BYTXMIRR>=26.68	54.63	.038	.53	.87
8	BYTXMIRR>=23.85 BYTXMIRR< 30.51 m_selfcpt>=6.735 geo_none=1	52.93	.033	.50	.82
9	BYTXMIRR>=23.85 BYTXMIRR< 30.51 m_selfcpt< 6.735 geo_none=0 BYTXMIRR< 26.68	50.38	.032	.55	.83
10	BYTXMIRR< 23.85 BYTXMIRR>=16.65 gm_none=1 geo_none=0 GPA910_m>=2.75	50.31	.044	.46	.76
11	BYTXMIRR>=23.85 BYTXMIRR< 30.51 m_selfcpt< 6.735 geo_none=1	48.39	.036	.39	.73
12	BYTXMIRR< 23.85 BYTXMIRR>=16.65 gm_none=1 geo_none=0 GPA910_m< 2.75	46.46	.056	.49	.80
13	BYTXMIRR< 23.85 BYTXMIRR>=16.65 gm_none=1 geo_none=1	43.89	.095	.50	.79
14	BYTXMIRR< 23.85 BYTXMIRR>=16.65 gm_none=0 GPA910_m>=2.25	41.42	.045	.42	.75

Table 64 continued

15	BYTXMIRR< 23.85 BYTXMIRR< 16.65 gm_none=1 geo_none=0	40.78	.035	.42	.72
16	BYTXMIRR< 23.85 BYTXMIRR< 16.65 gm_none=1 geo_none=1 BYTXMIRR>=12.5	37.14	.056	.47	.77
17	BYTXMIRR< 23.85 BYTXMIRR>=16.65 gm_none=0 GPA910_m< 2.25	37.01	.025	.47	.77
18	BYTXMIRR< 23.85 BYTXMIRR< 16.65 gm_none=0 BYTXMIRR>=13.11	33.89	.060	.48	.85
19	BYTXMIRR< 23.85 BYTXMIRR< 16.65 gm_none=1 geo_none=1 BYTXMIRR< 12.5	31.69	.033	.44	.85
20	BYTXMIRR< 23.85 BYTXMIRR< 16.65 gm_none=0 BYTXMIRR< 13.11	29.15	.092	.61	.90

Note: For ease of interpretation, surrogate branches for missing data are excluded from this table. Coverage was the same across training and test sets.

Table 65. CART ruleset (Study 2, 1933 possible predictors)

Rule	**Antecedent**	**Math ach**	**Coverage***	**Within 5 pts of estimate***	**Within 10 pts of estimate***
1	BYTXMIRR>=23.85 BYTXMIRR>=30.64 BYTXMIRR>=35.26	68.45	.087	.68	.95
2	BYTXMIRR>=23.85 BYTXMIRR>=30.64 BYTXMIRR< 35.26 m_selfcpt>=6.83	63.81	.096	.62	.91
3	BYTXMIRR>=23.85 BYTXMIRR>=30.64 BYTXMIRR< 35.26 m_selfcpt< 6.83	58.59	.054	.49	.85
4	BYTXMIRR>=23.85 BYTXMIRR< 30.64 m_selfcpt>=6.78 geo_none=0	58.18	.083	.48/.50	.84
5	BYTXMIRR>=23.85 BYTXMIRR< 30.64 m_selfcpt>=6.78 geo_none=1	53.21	.032	.48	.81/.83

Table 65 continued

6	BYTXMIRR>=23.85 BYTXMIRR< 30.64 m_selfcpt< 6.78 geo_1=1	53.08	.062	.53/.55	.86
7	BYTXMIRR>=23.85 BYTXMIRR< 30.64 m_selfcpt< 6.78 geo_1=0	48.74	.047	.42	.70
8	BYTXMIRR< 23.85 BYTXMIRR>=16.65 gm_none=1 geo_none=0	48.21	.097	.48	.79
9	BYTXMIRR< 23.85 BYTXMIRR>=16.65 gm_none=1 geo_none=1	43.82	.094	.51/.48	.78
10	BYTXMIRR< 23.85 BYTXMIRR< 16.65 gm_none=1 F2HSPROG=ACADEMIC PROGRAM	41.36	.037	.46/.44	.74/.72
11	BYTXMIRR< 23.85 BYTXMIRR>=16.65 gm_none=0	39.96	.071	.40	.74
12	BYTXMIRR< 23.85 BYTXMIRR< 16.65 gm_none=1 F2HSPROG=GEN. HS PRGRM,VOC./TECHNICAL,OTHER SPEC PRGRM,SPEC ED PRGRM,ALT/DO PREVENT.,DON'T KNOW BYTXMIRR>=13.8	37.52	.037	.52	.84
13	BYTXMIRR< 23.85 BYTXMIRR< 16.65 gm_none=0 BYTXMIRR>=13.11	33.96	.062	.47	.84
14	BYTXMIRR< 23.85 BYTXMIRR< 16.65 gm_none=1 F2HSPROG=GEN. HS PRGRM,VOC./TECHNICAL,OTHER SPEC PRGRM,SPEC ED PRGRM,ALT/DO PREVENT.,DON'T KNOW BYTXMIRR< 13.8	33.03	.046	.41/.43	.73/.76
15	BYTXMIRR< 23.85 BYTXMIRR< 16.65 gm_none=0 BYTXMIRR< 13.11	29.07	.094	.61	.90

Note: For ease of interpretation, surrogate branches for missing data are excluded from this table. Coverage and confidence values are indicated separately for training and test sets (training/test) only if they differed by more than a percentage point.

Table 66. C5.0 ruleset (Study 1, 29 possible predictors)

Rule	Antecedent	Math ach	Coverage (train/test)*	Confidence (train/test)*
1	BYTXMIRR > 23.846, BYTXMIRR > 28.597	high	.29	.96
2	BYTXMIRR > 23.846, BYTXMIRR <= 28.597, geo_1 = 1	high	.09	.81
3	BYTXMIRR > 23.846, BYTXMIRR <= 28.597, geo_1 = 0, BYGRADS <= 2.6	low	.02	.50/.52
4	BYTXMIRR > 23.846, BYTXMIRR <= 28.597, geo_1 = 0, BYGRADS > 2.6	high	.04	.70/.71
5	BYTXMIRR <= 23.846, BYTXMIRR <= 17.999	low	.31	.93
6	BYTXMIRR <= 23.846, BYTXMIRR > 17.999, gm_none = 0	low	.05	.82
7	BYTXMIRR <= 23.846, BYTXMIRR > 17.999, gm_none = 1, GPA910_m <= 2.5	low	.08	.64
8	BYTXMIRR <= 23.846, BYTXMIRR > 17.999, gm_none = 1, GPA910_m > 2.5, BYSES <= -0.248	low	.03	.58
9	BYTXMIRR <= 23.846, BYTXMIRR > 17.999, gm_none = 1, GPA910_m > 2.5, BYSES > -0.248	high	.03	.63

Table 67. C5.0 ruleset (Study 2, 1933 possible predictors)

Rule	Antecedent	Math ach	Coverage (train/test)*	Confidence (train/test)*
1	BYTXMIRR > 23.91, BYTXMIRR > 28.597	high	.29	.96
2	BYTXMIRR > 23.91, BYTXMIRR <= 28.597, geo_1 = 1	high	.09	.81
3	BYTXMIRR > 23.91, BYTXMIRR <= 28.597, geo_1 = 0, BYGRADS <= 2.6	low	.02	.52/.50
4	BYTXMIRR > 23.91, BYTXMIRR <= 28.597, geo_1 = 0, BYGRADS > 2.6	high	.04	.71/.68
5	BYTXMIRR <= 23.91, BYTXMIRR <= 18.004	low	.31	.93
6	BYTXMIRR <= 23.91, BYTXMIRR > 18.004, BYGRADS <= 2.4	low	.05	.81
7	BYTXMIRR <= 23.91, BYTXMIRR > 18.004, BYGRADS > 2.4, geo_none = 1	low	.07	.69
8	BYTXMIRR <= 23.91, BYTXMIRR > 18.004, BYGRADS > 2.4, geo_none = 0, GPA910_m <= 2.5	low	.04	.55
9	BYTXMIRR <= 23.91, BYTXMIRR > 18.004, BYGRADS > 2.4, geo_none = 0, GPA910_m > 2.5	high	.04	.62/.61

Table 68. QUEST ruleset (Study 2, 29 possible predictors)

Rule	Antecedent	Math ach	Coverage	Confidence
1	geo_none=0, gm_none=0, BYTXMIRR<=23.531	Low	.03	.85
2	geo_none=0, gm_none=0, BYTXMIRR>23.531	High	.03	.81
3	geo_none=0, gm_none=1, al2_none=0	High	.17	.88
4	geo_none=0, gm_none=1, al2_none=1, BYTXMIRR<=21.222	Low	.07	.69
5	geo_none=0, gm_none=1, al2_none=1, BYTXMIRR>21.222	High	.22	.83
6	geo_none=1, al2_none=0, BYTXMIRR<=22.491	Low	.03	.78
7	geo_none=1, al2_none=0, BYTXMIRR>22.491	High	.05	.82
8	geo_none=1, al2_none=0, gm_none=0	Low	.21	.92
9	geo_none=1, al2_none=0, gm_none=1, BYTXMIRR<=24.942	Low	.16	.82
10	geo_none=1, al2_none=0, gm_none=1, BYTXMIRR>24.942	High	.04	.72

Note: For ease of interpretation, surrogate branches for missing data are excluded from this table. Coverage and confidence were the same across training and test sets.

Table 69. QUEST ruleset (Study 2, 1933 possible predictors)

Rule	Antecedent	Math ach	Coverage	Confidence
1	geo_none=0, BYTXMIRR<=21.534	Low	.11	.72
2	geo_none=0, BYTXMIRR>21.534	High	.41	.87
3	geo_none=1, BYTXMIRR<=24.625	Low	.39	.88
4	geo_none=1, BYTXMIRR>24.625	High	.09	.76

Note: For ease of interpretation, surrogate branches for missing data are excluded from this table. Coverage and confidence were the same across training and test sets.

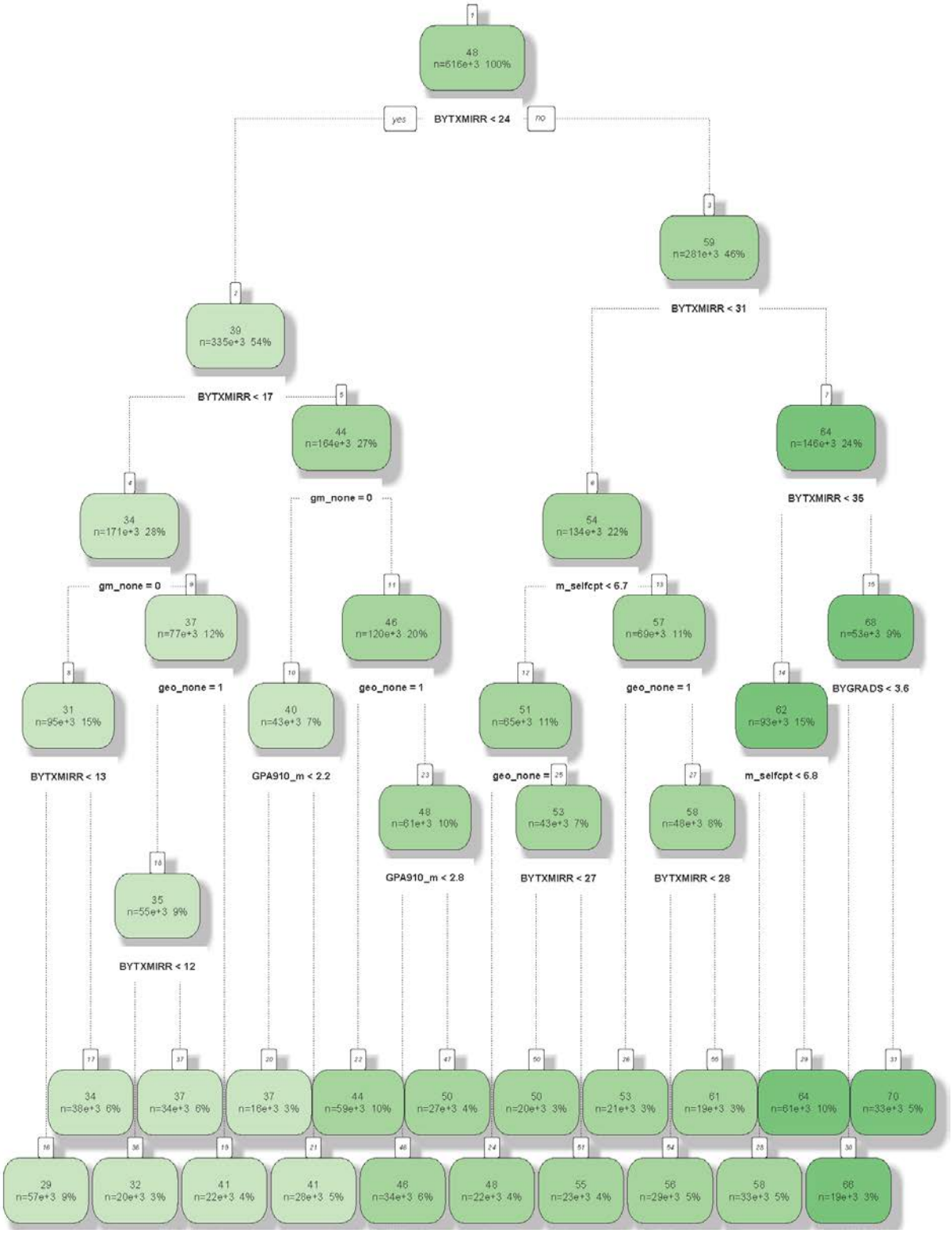


Figure 36. CART tree (Study 2, 29 possible predictors; results shown on training data)

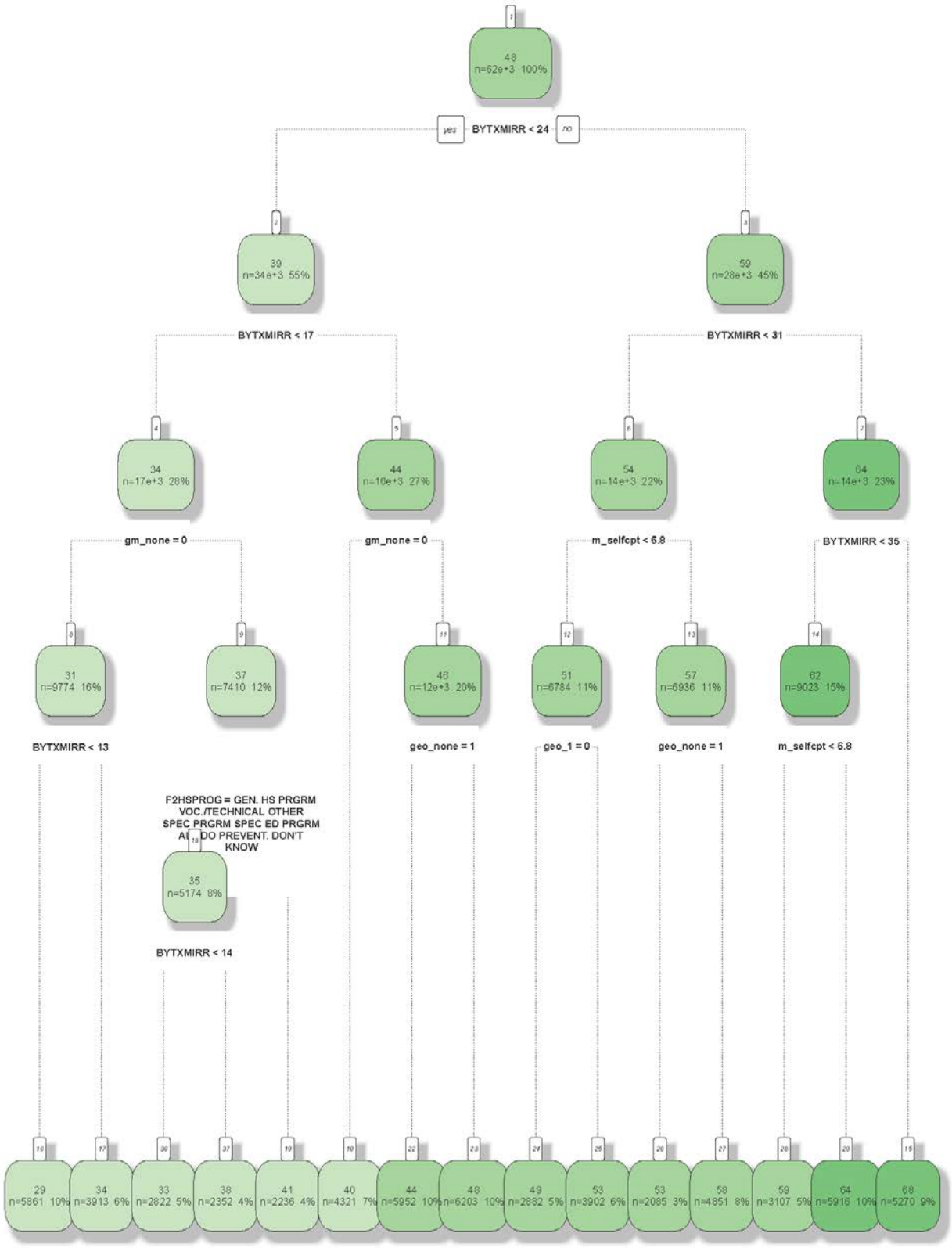


Figure 37. CART tree (Study 2, 1933 possible predictors; results shown on training data)

```
BYTXMIRR > 23.846:
:...BYTXMIRR > 28.597: high (183477.8/10935.3)
:   BYTXMIRR <= 28.597:
:   :...geo_1 = 1: high (59074.5/12000.3)
:       geo_1 = 0:
:       :...BYGRADS <= 2.6: low (14720.2/6993.8)
:           BYGRADS > 2.6: high (25845/8185.8)
BYTXMIRR <= 23.846:
:...BYTXMIRR <= 17.999: low (198875/15198.8)
    BYTXMIRR > 17.999:
    :...gm_none = 0: low (35373.6/6551.1)
        gm_none = 1:
        :...GPA910_m <= 2.5: low (57460.1/19919.1)
            GPA910_m > 2.5:
            :...BYSES <= -0.248: low (17911.7/7511)
                BYSES > -0.248: high (23276.2/8575.8)
```

Figure 38. C5.0 tree (Study 2, 29 possible predictors; results shown on training data)

```
BYTXMIRR > 23.91:
:...BYTXMIRR > 28.597: high (18299.8/1076.8)
:   BYTXMIRR <= 28.597:
:   :...geo_1 = 1: high (5851.9/1165.5)
:       geo_1 = 0:
:       :...BYGRADS <= 2.6: low (1453.1/676.8)
:           BYGRADS > 2.6: high (2556.8/775.3)
BYTXMIRR <= 23.91:
:...BYTXMIRR <= 18.004: low (19919.7/1527.6)
    BYTXMIRR > 18.004:
    :...BYGRADS <= 2.4: low (3057.5/577.1)
        BYGRADS > 2.4:
        :...geo_none = 1: low (4959.3/1503.1)
            geo_none = 0:
            :...GPA910_m <= 2.5: low (3022.6/1303.4)
                GPA910_m > 2.5: high (2552.3/967.3)
```

Figure 39. C5.0 tree (Study 2, 1933 possible predictors; results shown on training data)

```
BYTXMIRR <= 23.876
|   BYTXMIRR <= 17.999: low (128703.0/8616.0)
|   BYTXMIRR > 17.999
|   |   geo_none = 0
|   |   |   m_selfcpt <= 6.770048
|   |   |   |   BYSES <= -0.032: low (12631.0/4633.0)
|   |   |   |   BYSES > -0.032: high (12421.0/5990.0)
|   |   |   m_selfcpt > 6.770048: high (16836.0/5970.0)
|   |   geo_none = 1: low (50219.0/12649.0)
|   |   geo_none = missing: low (7385.0/1802.0)
BYTXMIRR > 23.876: high (182481.0/22045.0)
```

Figure 40. C4.5 tree (Study 2, 29 possible predictors; results shown on training data)

```
BYTXMIRR <= 23.91
|   BYTXMIRR <= 18.004: low (12816.0/861.0)
|   BYTXMIRR > 18.004
|   |   BYGRADS <= 2.4: low (2276.0/398.0)
|   |   BYGRADS > 2.4
|   |   |   geo_none = 0
|   |   |   |   m_selfcpt <= 6.770048: low (2293.0/1047.0)
|   |   |   |   m_selfcpt > 6.770048: high (1572.0/524.0)
|   |   |   geo_none = 1: low (3461.0/1029.0)
|   |   |   geo_none = missing: low (484.0/146.0)
BYTXMIRR > 23.91
|   BYTXMIRR <= 28.597: high (6372.0/1669.0)
|   BYTXMIRR > 28.597
|   |   BYGRADS <= 2.8: high (1730.0/225.0)
|   |   BYGRADS > 2.8
|   |   |   m_selfcpt <= 6.751391
|   |   |   |   BYS8F = missing: low (19.0/5.0)
|   |   |   |   BYS8F = NO: high (1435.0/103.0)
|   |   |   |   BYS8F = YES: high (1757.0/80.0)
|   |   |   m_selfcpt > 6.751391: high (6901.0/76.0)
```

Figure 41. C4.5 tree (Study 2, 1933 possible predictors; results shown on training data)

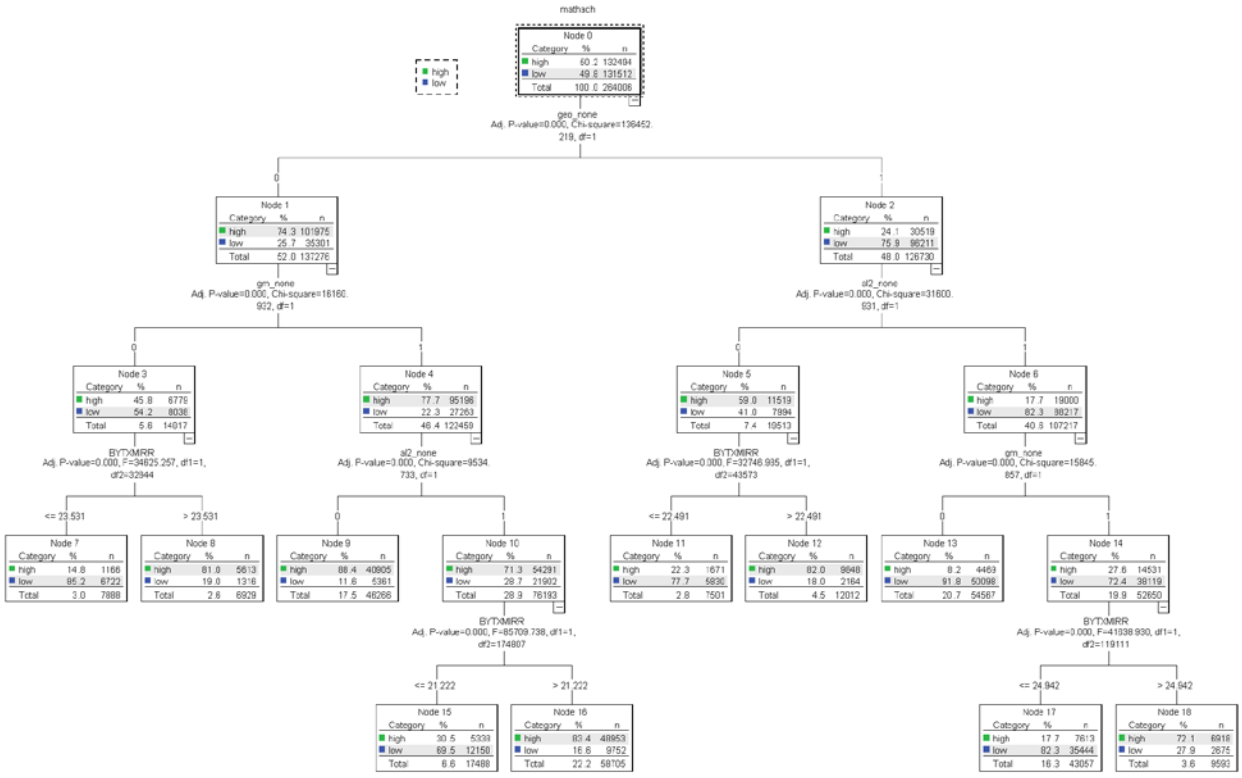


Figure 42. QUEST tree (Study 2, 29 possible predictors; performance on test set)

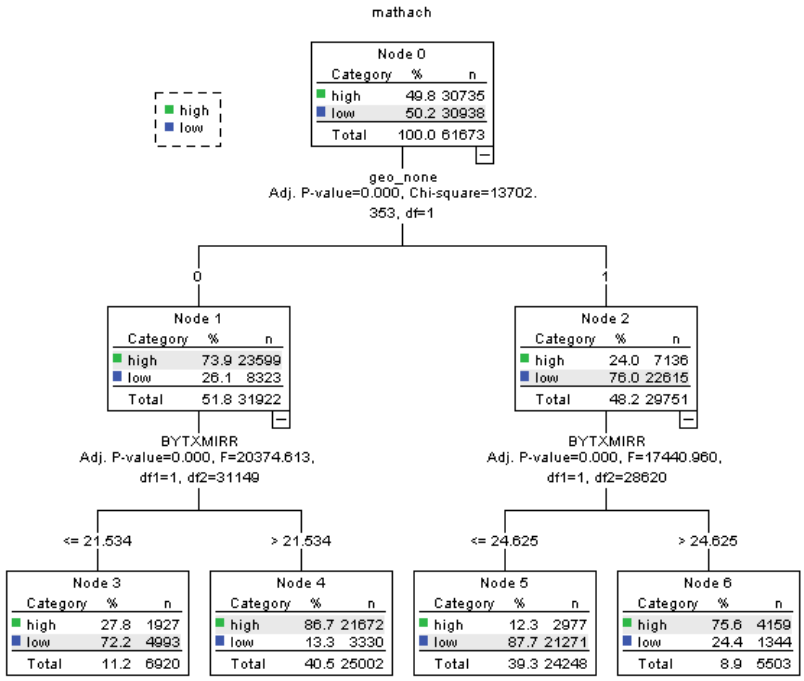


Figure 43. QUEST tree (Study 2, 1933 possible predictors; performance on test set)

Table 70. Categorization of variables included in Study 2 rule induction with 1933 possible predictors

8th grade math scores	MATHEMATICS IRT-ESTIMATED NUMBER RIGHT (BYTXMIRR)
8th grade SES	SOCIO-ECONOMIC STATUS COMPOSITE (BYSES)
Advanced/gifted class in 8th grade	R'S ABILITY GROUP FOR MATHEMATICS (BYS60A), IN ADVANCED, ENRICHED, ACCELERATED MATH (BYS66D), ATTEND ALGEBRA AT LEAST ONCE A WEEK (BYS67C), ENROLLED IN CLASSES FOR GIFTED STUDENTS (BYS68C), CHILD RECVD SERVICES FOR LEARNING PROBLM (BYP48G), CHILD ENROLLED IN GIFTED/TALENTED PROG (BYP51)
Discussed drugs/alcohol abuse in 8th gr	TALK TO COUNSELOR ABT DRUG/ALCOHOL ABUSE (BYS51GA), TALK TO OTH ADULT ABT DRUG/ALCOHOL ABUSE (BYS51GC)
Discussed studies/HS program in 8th gr	TALK TO COUNSELOR ABOUT STUDIES IN CLASS (BYS51EA)
Ever held back	EVER HELD BACK A GRADE IN SCHOOL (BYS74)
Freq of dept meetings 12th gr	FREQUENCY OF DEPARTMENT STAFF MEETINGS (F2T3_4)
Grades in 8th grade	GRADES COMPOSITE (BYGRADS)
Graduation efficacy	Graduation efficacy (grad_eff)
HS program	DESCRIBE PRESENT HIGH SCHOOL PROGRAM (F1S20), RESPONDENT-INDICATED HIGH SCHOOL PROGRAM (F2HSPROG)
Lives w/sisters	R LIVES IN HOUSEHOLD WITH SISTER(S) (BYS8F)
Locus of control	CHANCE AND LUCK IMPORTANT IN MY LIFE (BYS44M)
Math GPA	Math GPA of 9th and 10th grades (GPA910_m)
Math instruction 10th gr	EMPHASIS ON LEARNING MATH FACTS/RULES (F1S31B), OFTEN USE BOOKS OTHR THN MATH TEXT BOOKS (F1S32B)
Math self-concept	Math self-concept (m_selfcpt)
Middle school emphasized sports	SCHOOL EMPHASIZES SPORTS (BYSC47N)
Middle school provided standardized test scores to families	STDIZED TEST RESULTS PROV. TO FAMILIES (BYSC37)

Table 70 continued

Other course-taking	ATTEND ENGLISH AT LEAST ONCE A WEEK (BYS67BA)
Parental education	HIGHEST LEVEL OF EDUCATION R COMPLETED (BYP30)
Parents check homework	HOW OFTEN PARENTS CHECK ON R'S HOMEWORK (BYS38A)
Race	COMPOSITE RACE (RACE), Race is White (white)
School climate 12th gr	THERE IS CHEATING IN SCHOOL (F2S7J), SOME TEACHERS IGNORE CHEATING (F2S7K), DEGREE GANG ACTIVITIES A PROBLEM (F2T3_16C), DEGREE PHYSICAL ABUSE OF TCHRS A PROBLEM (FT3_16K)
School demographics in 8th gr	PERCENT MINORITY IN SCHOOL (G8MINOR), PUPILS ASSIGNED FOR RACIAL/ETHNIC COMP. (BYSC24C)
Self-concept	SELF CONCEPT 2 (BYCNCPT2)
Size of middle school	TOTAL SCHOOL ENROLLMENT COMPOSITE (BYSCENRL), NO. OF FULL TIME REGULAR TEACHERS (BYSC17)
Student at 12th grade	MEMBER 12TH GRADE IN-SCHOOL CLASS 91-92 (G12COHRT)
Student expectation	Student expectations (Sexp)
Survey weight	F2 8TH GRADE PANEL WEIGHT (F2PNLWT)
Taken algebra 2	Have taken no algebra 2 (al2_non)
Taken general math	Have taken no, 1 year, or 2 years of general math (gm_none, gm_1, gm_2).
Taken geometry	Have not taken geometry (geo_none), have taken 1 year of geometry (geo_1)
Teacher demographics in 8th grade	NO. OF BLACK, NON-HISPANIC TEACHERS (BYSC20D), NO. FULL TIME TEACHERS WITH GRAD DEGREE (BYSC21)

APPENDIX E

ASSOCIATION RULES FOR STUDY 1

Table 71. Attribute-values associated with high 12th grade achievement among Black students from low income families identified by association rule mining

Cat	Attribute-value	Variable label	TRP	FPR	PLR	Prec	FOR	RP
S	BYP44=NO	8TH GRADER EVER HELD BACK A GRADE	.91	.6	1.52	.24	.04	5.36
S	BYS74=NO	EVER HELD BACK A GRADE IN SCHOOL	.91	.59	1.54	.24	.04	5.56
S	BYS78A=SELDOM	HOW OFTEN COME TO CLASS W/O PENCIL/PAPER	.59	.39	1.51	.24	.12	1.95
S	BYS44C=STRONGLY DISAGREE	GOOD LUCK MORE IMPORTANT THAN HARD WORK	.57	.34	1.68	.26	.12	2.17
S	BYS67C=ATTEND	ATTEND ALGEBRA AT LEAST ONCE A WEEK	.57	.18	3.17	.4	.1	4.04
S	BYLOCUS1=(.26 31.24]	LOCUS OF CONTROL 1	.55	.3	1.83	.28	.12	2.34
S	BYP53=YES	CHILD ENROLLED IN ALGEBRA COURSE THIS YR	.52	.2	2.6	.35	.11	3.16
S	BYS42B=OVER 5 HRS A DAY	NO. OF HOURS R WATCHES TV ON WEEKENDS	.52	.32	1.63	.25	.13	1.97
S	BYS67BG=ATTEND	ATTEND MUSIC AT LEAST ONCE A WEEK	.51	.31	1.65	.26	.13	1.98
S	BYS60C=HIGH	R'S ABILITY GROUP FOR ENGLISH	.5	.22	2.27	.32	.12	2.73
S	BYLOCU2T=TERTILE 3 HIGH	TERTILE CODING OF VARIABLE BYLOCUS2	.49	.25	1.96	.29	.12	2.34
S	BYS60A=HIGH	R'S ABILITY GROUP FOR MATHEMATICS	.47	.18	2.61	.35	.12	2.97
S	BYP63A=YES	CHILD EVER INVOLVED IN BOY/GIRL SCOUTS	.46	.29	1.59	.25	.14	1.82
S	BYS49=COLL PREP ACADEMIC	WHICH PROGRAM R EXPECTS TO ENROLL IN H.S	.46	.21	2.19	.31	.12	2.51
S	BYS71B=STRONGLY DISAGREE	AFRAID TO ASK QUESTION IN SOCIAL STUDIES	.45	.22	2.05	.3	.13	2.33
S	BYS67B=DO NOT ATTEND	ATTEND REGULAR MATH AT LEAST ONCE A WEEK	.45	.17	2.65	.36	.12	2.93

Table 71 continued

S	BYT2_2.ALL=HIGHER LEVELS	ACHIEVEMENT LEVEL-THIS CLASS VS AVERAGE	.44	.12	3.67	.43	.12	3.7
S	BYLOCUS2=(.33 31.28]	LOCUS OF CONTROL 2	.43	.16	2.69	.36	.12	2.9
S	BYS44G=STRONGLY DISAGREE	PLANS HARDLY WORK OUT, MAKES ME UNHAPPY	.43	.22	1.95	.29	.13	2.19
S	BYPSEPLN=HIGHER SCH AFTR COLL	POST-SECONDARY EDUCATION PLANS	.43	.19	2.26	.32	.13	2.51
S	BYS82A=PARTICIPATED MEMBER	PARTICIPATED IN SCIENCE FAIRS	.42	.21	2	.29	.13	2.22
S	BYLOCU1T=TERTILE 3 HIGH	TERTILE CODING OF VARIABLE BYLOCUS1	.42	.23	1.83	.28	.14	2.03
S	BYS44J=STRONGLY DISAGREE	AT TIMES I THINK I AM NO GOOD AT ALL	.42	.27	1.56	.24	.14	1.72
S	BYP61DB=YES	8TH GRADER GOES TO SCIENCE MUSEUMS	.41	.24	1.71	.26	.14	1.89
S	BYS72B=STRONGLY DISAGREE	AFRAID TO ASK QUESTION IN SCIENCE CLASS	.41	.21	1.95	.29	.13	2.15
S	BYS78B=SELDOM	HOW OFTEN COME TO CLASS WITHOUT BOOKS	.41	.24	1.71	.26	.14	1.89
S	BYS60D=HIGH	R'S ABILITY GROUP FOR SOCIAL STUDIES	.4	.2	2	.29	.14	2.18
S	BYS44M=DISAGREE	CHANCE AND LUCK IMPORTANT IN MY LIFE	.4	.25	1.6	.25	.14	1.75
S	BYS48B=HIGHER SCH AFTR COLL	HOW FAR IN SCHL R'S MOTHER WANTS R TO GO	.4	.24	1.67	.26	.14	1.82
S	BYS52=PROBUSINSSMGRL	KIND OF WORK R EXPECTS TO DO AT AGE 30	.4	.23	1.74	.27	.14	1.9
S	BYT1_2.SCIENCE=NO	STUDENT PERFORMS BELOW ABILITY	.39	.26	1.5	.24	.15	1.62
S	BYS60B=HIGH	R'S ABILITY GROUP FOR SCIENCE	.39	.17	2.29	.32	.13	2.44
S	BYP61CB=YES	8TH GRADER GOES TO ART MUSEUMS	.39	.24	1.63	.25	.14	1.77
S	BYP38B=YES	DID 8TH GRADER ATTEND NURSERY/PRE-SCHOOL	.38	.21	1.81	.27	.14	1.95
S	BYS68A=YES	ENROLLED IN CLASSES FOR GIFTED STUDENTS	.38	.15	2.53	.35	.13	2.62
S	BYS54=NOT WRKD FOR PAY	KIND OF WORK R DOES FOR PAY CURRENT JOB	.38	.22	1.73	.26	.14	1.86
S	BYS41=LESS THAN 1 HOUR	TIME SPENT AFTER SCHL WTH NO ADULT PRSNT	.37	.26	1.42	.23	.15	1.52
S	BYP51=YES	CHILD ENROLLED IN GIFTED/TALENTED PROG	.33	.08	4.13	.46	.13	3.51

Table 71 continued

S	BYS48A=HIGHER SCH AFTR COLL	HOW FAR IN SCHL R'S FATHER WANTS R TO GO	.32	.19	1.68	.26	.15	1.74
S	BYS67AA=ATTEND	ATTEND LABORATORY AT LEAST ONCE A WEEK	.31	.17	1.82	.28	.15	1.86
S	BYT2_2.ENGLISH=HIGHER LEVELS	ACHIEVEMENT LEVEL-THIS CLASS VS AVERAGE	.31	.05	6.2	.56	.13	4.29
S	autonomy=(45]	#N/A	.28	.16	1.75	.27	.15	1.76
S	BYS42A=4-5 HOURS	NO. OF HOURS R WATCHES TV ON WEEKDAYS	.28	.11	2.55	.35	.14	2.4
S	BYS82E=PARTICIPATED MEMBER	PARTICIPATED IN BAND OR ORCHESTRA	.27	.15	1.8	.27	.15	1.8
S	BYP60B=YES	CHILD STUDY MUSIC OUTSIDE REGULAR SCHOOL	.27	.12	2.25	.32	.15	2.17
S	BYS82O=PARTICIPATED MEMBER	PARTICIPATED IN ACADEMIC HONORS SOCIETY	.26	.08	3.25	.4	.14	2.81
S	BYS44M=STRONGLY DISAGREE	CHANCE AND LUCK IMPORTANT IN MY LIFE	.26	.11	2.36	.33	.15	2.23
S	BYT2_2.SCIENCE=HIGHER LEVELS	ACHIEVEMENT LEVEL-THIS CLASS VS AVERAGE	.25	.07	3.57	.43	.14	2.97
S	BYP60H=YES	CHILD STUDY OTHER SKILLS OUTSIDE REG SCH	.2	.1	2	.29	.16	1.88
F	BYP85F=FALSE	CHLD TEST SCORES NOT GOOD ENOUGH QUALIFY	.77	.44	1.75	.27	.08	3.39
F	BYP85E=FALSE	CHILD GRADES NOT HIGH ENOUGH TO QUALIFY	.74	.42	1.76	.27	.09	3.14
F	BYP85G=FALSE	TOO MUCH WORK TO APPLY FOR FINANCIAL AID	.67	.5	1.34	.22	.12	1.81
F	BYS35E=HAVE	R'S FAMILY HAS AN ATLAS	.51	.32	1.59	.25	.13	1.91
F	BYS62=YES	DID PRNTS/GRDNS WANT R TO TAKE ALGEBRA	.47	.26	1.81	.27	.13	2.11
F	BYP85A=FALSE	CHILD WILL BE ABLE TO EARN MONEY FOR ED	.46	.28	1.64	.26	.14	1.89
F	BYS37B=NO	R'S PARENTS SPOKE TO TEACHER/COUNSELOR	.42	.29	1.45	.23	.15	1.59
F	BYS40H=RARELY	NO ONE IS HOME WHEN R RETURNS FROM SCHL	.39	.23	1.7	.26	.14	1.84
F	BYP62B4=YES	R KNOWS PARENT(S) OF CHILD'S 4TH FRIEND	.36	.25	1.44	.23	.15	1.53
F	BYSES=(-.214.792]	SOCIO-ECONOMIC STATUS COMPOSITE	.32	.14	2.29	.32	.14	2.28

Table 71 continued

F	BYP61EA=YES	R GOES TO HISTORY MUSEUMS	.3	.12	2.5	.34	.14	2.41
F	BYP59A=YES	BELONG TO PARENT-TEACHER ORGANIZATION	.3	.18	1.67	.26	.15	1.71
F	BYP52B=VERY IMPORTANT	HOW IMPORT GAINING DEEPR UNDERST OF SUBS	.3	.06	5	.51	.13	3.8
F	BYP61DA=YES	R GOES TO SCIENCE MUSEUMS	.29	.13	2.23	.32	.15	2.18
F	BYP61CA=YES	R GOES TO ART MUSEUMS	.28	.14	2	.29	.15	1.98
F	BYP52D=VERY IMPORTANT	HOW IMPRTNT GREATER INTELLECTL CHALLENGE	.28	.05	5.6	.54	.14	3.95
F	BYS4OCC=CLERICAL	MOTHER/FEMALE GUARDIAN'S OCCUPATION	.25	.09	2.78	.37	.15	2.5
F	BYP76=MASTER'S DEGREE	HOW FAR IN SCHOOL R EXPECT CHILD TO GO	.25	.09	2.78	.37	.15	2.5
F	BYP76=PH.D. M.D.OTHR	HOW FAR IN SCHOOL R EXPECT CHILD TO GO	.24	.06	4	.45	.14	3.15
F	BYPARMAR=DIVORCED	PARENTS' MARITAL STATUS	.23	.15	1.53	.24	.16	1.52
F	BYP7=DIVORCED	R'S CURRENT MARITAL STATUS	.23	.15	1.53	.24	.16	1.52
F	BYSESQ=QUARTILE 3	QUARTILE CODING OF BYSES VARIABLE	.21	.08	2.63	.35	.15	2.33
Sc	BYS59A=AGREE	STUDENTS GET ALONG WELL WITH TEACHERS	.65	.46	1.41	.23	.12	1.91
Sc	BYS59K=STRONGLY DISAGREE	I DON'T FEEL SAFE AT THIS SCHOOL	.51	.29	1.76	.27	.13	2.13
Sc	BYT2_11.ALL=80 - 89%	PERCENTAGE OF TEXTBOOK COVERED IN COURSE	.38	.24	1.58	.25	.15	1.71
Sc	BYS58F=MINOR	VANDALISM OF SCHOOL PROPERTY A PROBLEM	.38	.21	1.81	.27	.14	1.95
Sc	BYP74E=STRONGLY AGREE	MY CHILD ENJOYS SCHOOL	.38	.23	1.65	.26	.14	1.78
Sc	BYT2_20G.ALL=MAJOR TOPIC	EMPHASIS GIVEN TO ALGEBRA	.37	.24	1.54	.24	.15	1.65
Sc	BYP74D=STRONGLY AGREE	MY CHILD IS WORKING HARD AT SCHOOL	.37	.21	1.76	.27	.14	1.88
Sc	BYS58B=MODERATE	STUDENT ABSENTEEISM A PROBLEM AT SCHOOL	.37	.17	2.18	.31	.14	2.28
Sc	sculture=(3.755]	School culture	.37	.18	2.06	.3	.14	2.17
Sc	BYP57F=ONCE OR TWICE	CONTACTED ABOUT SCHOOL FUND RAISING	.36	.24	1.5	.24	.15	1.6
Sc	BYP74C=STRONGLY AGREE	MY CHILD IS CHALLENGED AT SCHOOL	.34	.14	2.43	.34	.14	2.44
Sc	BYT3_26C.ALL=NOT A PROBLEM	DEGREE STUDENT CLASS CUTTING IS PROBLEM	.33	.22	1.5	.24	.15	1.57

Table 71 continued

Sc	BYT2_19.ALL=FIVE OR MORE	NO. OF BOOKS STUDENTS REQUIRED TO READ	.31	.14	2.21	.32	.14	2.2
Sc	BYS58K=MINOR	VERBAL ABUSE OF TEACHERS A PROBLEM	.31	.14	2.21	.32	.14	2.2
Sc	BYT2_17E.ALL=MINOR TOPIC	EMPHASIS GIVEN TO STUDY SKILLS	.3	.19	1.58	.25	.15	1.62
Sc	BYS58A=MODERATE	STUDENT TARDINESS A PROBLEM AT SCHOOL	.27	.16	1.69	.26	.15	1.7
Sc	BYS58C=MODERATE	STUDENTS CUTTING CLASS A PROBLEM AT SCHL	.26	.12	2.17	.31	.15	2.09
Sc	BYT3_23D.ALL=NO	HAS CONTINUING ED CREDIT TRAINING GIFTED	.26	.12	2.17	.31	.15	2.09
Sc	BYP74H=STRONGLY AGREE	SCH PREPARING STUDENTS WELL FOR COLLEGE	.25	.15	1.67	.26	.16	1.66
Sc	BYT2_13E.MATH=NO	DISTRICT COMMITTEE DETERMINE USE OF TEXT	.24	.16	1.5	.24	.16	1.5
Sc	BYS58K=MODERATE	VERBAL ABUSE OF TEACHERS A PROBLEM	.24	.06	4	.45	.14	3.15
Sc	BYT2_13D.ENGLISH=YES	DEPT HEAD HELPED DETERMINE USE OF TEXTBK	.23	.15	1.53	.24	.16	1.52
Sc	BYT2_12A.ENGLISH=STRONGLY DISAGREE	TEXTBOOK LEVEL TOO DIFFICLT FOR STUDENTS	.18	.11	1.64	.25	.16	1.58
P	BYP61EB=YES	8TH GRADER GOES TO HISTORY MUSEUMS	.43	.18	2.39	.33	.13	2.63
P	BYP62A4=YES	4TH FRIEND ATTENDS SAME SCHOOL	.38	.26	1.46	.23	.15	1.57
C	BYS51CC=NO	TALK TO OTHR ADULT ABT IMPROVING SCH WRK	.62	.4	1.55	.24	.12	2.09
C	BYS51AC=NO	TALK TO OTHER ADULT ABOUT H.S. PROGRAMS	.52	.32	1.63	.25	.13	1.97
M	BYP54=SCHOOL PERSONNEL	MOST INFLUENTIAL IN CHILD TAKING ALGEBRA	.44	.27	1.63	.25	.14	1.84
M	BYS65=TEACHERS	WHO HAD THE MST TO SAY ABT R TKNG ALGBR	.33	.2	1.65	.26	.15	1.72

Cat = Category (where S = student, F = family, Sc = School, P = peer, C = community, M = multiple); TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving.

Table 72. Attribute-values associated with high 12th grade achievement among Black students from high income families identified by association rule mining

Cat	Attribute-value	Variable label	TRP	FPR	PLR	Prec	FOR	RP
S	BYS55D=NEVER	PARENTS RECEIVED WARNING ABOUT GRADES	.77	.51	1.51	.44	.19	2.25
S	BYS44D=STRONGLY AGREE	I'M A PERSON OF WORTH, EQUAL OF OTHERS	.71	.48	1.48	.43	.22	1.94
S	BYCNCP2T=TERTILE 3 HIGH	TERTILE CODING OF VARIABLE BYCNCPT2	.64	.43	1.49	.43	.25	1.77
S	BYS44C=STRONGLY DISAGREE	GOOD LUCK MORE IMPORTANT THAN HARD WORK	.63	.37	1.7	.47	.23	2.01
S	BYLOCUS1=(0.263,1.24]	LOCUS OF CONTROL 1	.61	.34	1.79	.48	.23	2.06
S	BYCNCPT2=(0.342,1.23]	SELF CONCEPT 2	.59	.39	1.51	.44	.26	1.7
S	BYLOCU2T=TERTILE 3 HIGH	TERTILE CODING OF VARIABLE BYLOCUS2	.54	.29	1.86	.49	.25	1.96
S	BYLOCU1T=TERTILE 3 HIGH	TERTILE CODING OF VARIABLE BYLOCUS1	.54	.29	1.86	.49	.25	1.96
S	BYS78A=SELDOM	HOW OFTEN COME TO CLASS W/O PENCIL/PAPER	.54	.29	1.86	.49	.25	1.96
S	BYS67C=ATTEND	ATTEND ALGEBRA AT LEAST ONCE A WEEK	.54	.21	2.57	.57	.23	2.47
S	BYS44M=DISAGREE	CHANCE AND LUCK IMPORTANT IN MY LIFE	.53	.25	2.12	.52	.24	2.14
S	BYP53=YES	CHILD ENROLLED IN ALGEBRA COURSE THIS YR	.52	.23	2.26	.54	.24	2.21
S	BYS49=COLL PREP ACADEMIC	WHICH PROGRAM R EXPECTS TO ENROLL IN H.S	.51	.21	2.43	.56	.24	2.3
S	BYS48B=HIGHER SCH AFTR COLL	HOW FAR IN SCHL R'S MOTHER WANTS R TO GO	.5	.29	1.72	.47	.27	1.77
S	BYS67AB=DO NOT ATTEND	ATTEND SCIENCE AT LEAST ONCE A WEEK	.5	.31	1.61	.45	.27	1.67
S	BYS56C=VERY	STUDENTS IN CLASS SEE R AS GOOD STUDENT	.5	.34	1.47	.43	.28	1.54
S	BYP38B=YES	DID 8TH GRADER ATTEND NURSERY/PRE-SCHOOL	.49	.33	1.48	.43	.28	1.54
S	BYS60A=HIGH	R'S ABILITY GROUP FOR MATHEMATICS	.49	.13	3.77	.66	.23	2.85
S	BYS78B=SELDOM	HOW OFTEN COME TO CLASS WITHOUT BOOKS	.49	.28	1.75	.47	.27	1.77
S	BYPSEPLN=HIGHER SCH AFTR COLL	POST-SECONDARY EDUCATION PLANS	.49	.18	2.72	.58	.24	2.41

Table 72 continued

	BYS42B=OVER 5 HRS A DAY	NO. OF HOURS R WATCHES TV ON WEEKENDS	.49	.33	1.48	.43	.28	1.54
S	BYLOCUS2=(0.333,1.28]	LOCUS OF CONTROL 2	.48	.23	2.09	.52	.26	2.01
S	BYS67AD=DO NOT ATTEND	ATTEND EARTH SCIENCE AT LEAST ONCE A WK	.48	.26	1.85	.49	.27	1.83
S	BYP55=YES	CHILD ENROLLED IN FOREIGN LANG COURSE	.43	.17	2.53	.57	.26	2.17
S	BYS48A=HIGHER SCH AFTR COLL	HOW FAR IN SCHL R'S FATHER WANTS R TO GO	.41	.24	1.71	.47	.29	1.64
S	BYS52=PROBUS INSSMGRL	KIND OF WORK R EXPECTS TO DO AT AGE 30	.41	.25	1.64	.46	.29	1.59
S	BYS67B=DO NOT ATTEND	ATTEND REGULAR MATH AT LEAST ONCE A WEEK	.4	.19	2.11	.52	.28	1.88
S	BYS54=NOT WRKD FOR PAY	KIND OF WORK R DOES FOR PAY CURRENT JOB	.4	.2	2	.51	.28	1.82
F	BYP85F=FALSE	CHLD TEST SCORES NOT GOOD ENOUGH QUALIFY	.75	.46	1.63	.46	.19	2.37
F	BYP85E=FALSE	CHILD GRADES NOT HIGH ENOUGH TO QUALIFY	.74	.43	1.72	.47	.19	2.47
F	BYP85G=FALSE	TOO MUCH WORK TO APPLY FOR FINANCIAL AID	.71	.48	1.48	.43	.22	1.94
F	BYP85I=FALSE	DON'T SEE WAY TO GET MONEY FOR COLLEGE	.67	.5	1.34	.41	.25	1.61
F	BYSES=(-0.214,0.792]	SOCIO-ECONOMIC STATUS COMPOSITE	.62	.36	1.72	.47	.23	2.01
F	BYS36C=3 OR MORE TIMES	DISCUSS THNGS STUDIED IN CLASS WTH PRNTS	.6	.41	1.46	.43	.26	1.66
F	BYP85A=FALSE	CHILD WILL BE ABLE TO EARN MONEY FOR ED	.59	.37	1.59	.45	.25	1.8
F	BYS62=YES	DID PRNTS/GRDNS WANT R TO TAKE ALGEBRA	.52	.3	1.73	.47	.26	1.81
F	BYP59A=YES	BELONG TO PARENT-TEACHER ORGANIZATION	.49	.31	1.58	.45	.28	1.63
F	BYS36A=3 OR MORE TIMES	DISCUSS PROGRAMS AT SCHOOL WITH PARENTS	.49	.29	1.69	.47	.27	1.72
F	BYS35I=HAVE	R'S FAMILY HAS AN ELECTRIC DISHWASHER	.48	.3	1.6	.45	.28	1.63
F	BYP85H=FALSE	NOT MUCH INFORMATION ON FINANCIAL AID	.46	.3	1.53	.44	.28	1.55
F	BYSESQ=QUARTILE 4 HIGH	QUARTILE CODING OF BYSES VARIABLE	.39	.12	3.25	.63	.26	2.38
F	BYP70=YES	COMPUTER IN HOME USED FOR ED PURPOSES	.38	.22	1.73	.47	.29	1.62
Sc	BYSC49C=NOT A PROBLEM	DEGREE STUDENT CLASS CUTTING IS A PROB	.64	.35	1.83	.48	.22	2.19

Table 72 continued

Sc	BYP75=VERY SATISFIED	HOW SATISFIED WITH ED CHILD HAS RECEIVED 8TH GRADERS	.61	.35	1.74	.47	.24	2
Sc	BYSC38G=NO	RETAIND:FAILD ANY REQ COURSE	.52	.31	1.68	.46	.26	1.76
Sc	BYP74A=STRONGLY AGREE	THE SCH PLACES HIGH PRIORITY ON LEARNING	.52	.32	1.63	.46	.27	1.71
Sc	BYS59L=DISAGREE	STUDENT DISRUPTIONS INHIBIT LEARNING	.47	.3	1.57	.45	.28	1.59
Sc	BYP74B=STRONGLY AGREE	HOMEWORK ASSIGNED IS WORTHWHILE	.44	.28	1.57	.45	.29	1.56
Sc	BYSC47M=VERY MUCH ACCURATE	TEACHERS RESPOND TO INDIVIDUAL NEEDS	.43	.22	1.95	.5	.27	1.83
Sc	BYP74E=STRONGLY AGREE	MY CHILD ENJOYS SCHOOL	.4	.22	1.82	.48	.28	1.7
Sc	BYSC36C=A LOT	PARENTS INFLUENCE ASSIGNING HS COURSES	.39	.25	1.56	.45	.3	1.51
Sc	BYSC25=YES	SCHOOL HAS FORMAL ADMISSION PROCEDURES	.38	.19	2	.51	.28	1.79
P	BYS63=NEITHER	FRIENDS ENCRG/DISCRG R FROM TAKING ALGBR	.75	.5	1.5	.44	.2	2.13
M	BYS65=I DID	WHO HAD THE MST TO SAY ABT R TKNG ALGBR	.5	.28	1.79	.48	.26	1.82

Cat = Category (where S = student, F = family, Sc = School, P = peer, C = community, M = multiple); TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving.

Table 73. Attribute-values associated with high 12th grade achievement among Black students with low parental education identified by association rule mining

Cat	Attribute-value	Variable label	TRP	FPR	PLR	Prec	FOR	RP
S	BYLOCUS1=(0.263,1.24]	LOCUS OF CONTROL 1	.63	.31	2.03	.19	.06	3.26
S	BYS44C=STRONGLY DISAGREE	GOOD LUCK MORE IMPORTANT THAN HARD WORK	.62	.33	1.88	.18	.06	2.9
S	BYS78C=SELDOM	HOW OFTEN COME TO CLASS WITHOUT HOMEWORK	.54	.35	1.54	.15	.08	2
S	BYS67BG=ATTEND	ATTEND MUSIC AT LEAST ONCE A WEEK	.54	.32	1.69	.16	.07	2.25
S	BYLOCU1T=TERTILE 3 HIGH	TERTILE CODING OF VARIABLE BYLOCUS1	.52	.23	2.26	.21	.07	3.08
S	BYLOCU2T=TERTILE 3 HIGH	TERTILE CODING OF VARIABLE BYLOCUS2	.5	.26	1.92	.18	.07	2.51

Table 73 continued

S	BYS67C=ATTEND	ATTEND ALGEBRA AT LEAST ONCE A WEEK	.5	.17	2.94	.25	.06	3.9
S	BYP53=YES	CHILD ENROLLED IN ALGEBRA COURSE THIS YR	.48	.21	2.29	.21	.07	2.96
S	BYS49=COLL PREP ACADEMIC	WHICH PROGRAM R EXPECTS TO ENROLL IN H.S	.48	.2	2.4	.22	.07	3.11
S	BYS60A=HIGH	R'S ABILITY GROUP FOR MATHEMATICS	.46	.17	2.71	.24	.07	3.41
S	BYS44M=DISAGREE	CHANCE AND LUCK IMPORTANT IN MY LIFE	.46	.24	1.92	.18	.08	2.39
S	BYLOCUS2=(0.333,1.28]	LOCUS OF CONTROL 2	.44	.18	2.44	.22	.07	3.01
S	BYS44G=STRONGLY DISAGREE	PLANS HARDLY WORK OUT, MAKES ME UNHAPPY	.44	.23	1.91	.18	.08	2.34
S	BYS71B=STRONGLY DISAGREE	AFRAID TO ASK QUESTION IN SOCIAL STUDIES	.44	.24	1.83	.17	.08	2.23
S	BYS72B=STRONGLY DISAGREE	AFRAID TO ASK QUESTION IN SCIENCE CLASS	.43	.25	1.72	.17	.08	2.05
S	BYS60C=HIGH	R'S ABILITY GROUP FOR ENGLISH	.41	.21	1.95	.18	.08	2.32
S	BYS44J=STRONGLY DISAGREE	AT TIMES I THINK I AM NO GOOD AT ALL	.41	.26	1.58	.15	.08	1.83
S	BYS67B=DO NOT ATTEND	ATTEND REGULAR MATH AT LEAST ONCE A WEEK	.4	.16	2.5	.22	.08	2.94
S	BYS78B=SELDOM	HOW OFTEN COME TO CLASS WITHOUT BOOKS	.4	.26	1.54	.15	.09	1.76
S	BYS70B=STRONGLY DISAGREE	OFTEN AFRAID TO ASK QUESTIONS IN ENGLISH	.39	.25	1.56	.15	.09	1.78
S	BYT2_2.ALL=HIGHER LEVELS	ACHIEVEMENT LEVEL-THIS CLASS VS AVERAGE	.39	.11	3.55	.29	.07	3.96
S	BYS69B=STRONGLY DISAGREE	AFRAID TO ASK QUESTIONS IN MATH CLASS	.38	.23	1.65	.16	.08	1.88
S	BYPSEPLN=HIGHER SCH AFTR COLL	POST-SECONDARY EDUCATION PLANS	.38	.19	2	.19	.08	2.31
S	BYS54=NOT WRKD FOR PAY	KIND OF WORK R DOES FOR PAY CURRENT JOB	.37	.21	1.76	.17	.08	2.01
S	BYS60D=AREN'T GROUPED	R'S ABILITY GROUP FOR SOCIAL STUDIES	.35	.23	1.52	.15	.09	1.68
S	BYS82A=PARTICIPATED MEMBER	PARTICIPATED IN SCIENCE FAIRS	.33	.23	1.43	.14	.09	1.56
S	BYS60B=HIGH	R'S ABILITY GROUP FOR SCIENCE	.33	.17	1.94	.18	.08	2.15

Table 73 continued

S	BYS68A=YES	ENROLLED IN CLASSES FOR GIFTED STUDENTS	.32	.15	2.13	.2	.08	2.34
S	BYS71C=DISAGREE	SOC. STUDIES WILL BE USEFUL IN MY FUTURE	.28	.19	1.47	.14	.09	1.56
S	BYP51=YES	CHILD ENROLLED IN GIFTED/TALENTED PROG	.26	.07	3.71	.3	.08	3.57
S	BYP55=YES	CHILD ENROLLED IN FOREIGN LANG COURSE	.25	.11	2.27	.21	.09	2.35
S	BYS67BE=ATTEND	ATTEND FOREIGN LANG AT LEAST ONCE A WEEK	.23	.12	1.92	.18	.09	1.98
S	BYT2_2.ENGLISH=HIGHER LEVELS	ACHIEVEMENT LEVEL-THIS CLASS VS AVERAGE	.22	.05	4.4	.34	.09	3.89
S	BYT2_2.SCIENCE=HIGHER LEVELS	ACHIEVEMENT LEVEL-THIS CLASS VS AVERAGE	.19	.06	3.17	.27	.09	2.96
F	BYP85F=FALSE	CHLD TEST SCORES NOT GOOD ENOUGH QUALIFY	.73	.43	1.7	.16	.05	3.16
F	BYP85E=FALSE	CHILD GRADES NOT HIGH ENOUGH TO QUALIFY	.69	.42	1.64	.16	.06	2.74
F	BYP85A=FALSE	CHILD WILL BE ABLE TO EARN MONEY FOR ED	.47	.3	1.57	.15	.08	1.91
F	BYS62=YES	DID PRNTS/GRDNS WANT R TO TAKE ALGEBRA	.44	.27	1.63	.16	.08	1.95
F	BYSES=(-0.214,0.792]	SOCIO-ECONOMIC STATUS COMPOSITE	.42	.19	2.21	.2	.08	2.66
F	BYS40H=RARELY	NO ONE IS HOME WHEN R RETURNS FROM SCHL	.34	.21	1.62	.16	.09	1.79
F	BYSESQ=QUARTILE 3	QUARTILE CODING OF BYSES VARIABLE	.33	.14	2.36	.21	.08	2.59
F	BYS4OCC=CLERICAL	MOTHER/FEMALE GUARDIAN'S OCCUPATION	.3	.13	2.31	.21	.08	2.48
F	BYP61CA=YES	R GOES TO ART MUSEUMS	.28	.18	1.56	.15	.09	1.65
F	BYP61DA=YES	R GOES TO SCIENCE MUSEUMS	.28	.17	1.65	.16	.09	1.76
F	BYS35I=HAVE	R'S FAMILY HAS AN ELECTRIC DISHWASHER	.27	.18	1.5	.15	.09	1.58
F	BYP34B=CLERICAL	DESCRIPTION OF CURRENT JOB	.24	.11	2.18	.2	.09	2.24
F	BYP52B=VERY IMPORTANT	HOW IMPORT GAINING DEEPR UNDERST OF SUBS	.24	.05	4.8	.36	.08	4.22
F	BYP76=PH.D. M.D.OTHR	HOW FAR IN SCHOOL R EXPECT CHILD TO GO	.23	.07	3.29	.27	.09	3.15
F	BYP52D=VERY IMPORTANT	HOW IMPRTNT GREATER INTELLECTL CHALLENGE	.22	.04	5.5	.39	.09	4.53
F	BYS40A=RARELY	MOTHER HOME WHEN R RETURNS FROM SCHOOL	.22	.13	1.69	.16	.09	1.74
F	BYS34B=JUNIOR COLLEGE	MOTHER'S HIGHEST LEVEL OF EDUCATION	.2	.12	1.67	.16	.09	1.7

Table 73 continued

F	BYP62B3=NO	R KNOWS PARENT(S) OF CHILD'S 3RD FRIEND	.19	.1	1.9	.18	.09	1.91
Sc	BYSC49C=NOT A PROBLEM	DEGREE STUDENT CLASS CUTTING IS A PROB	.48	.31	1.55	.15	.08	1.9
Sc	BYS59K=STRONGLY DISAGREE	I DON'T FEEL SAFE AT THIS SCHOOL	.47	.29	1.62	.16	.08	1.99
Sc	sculture=(3.755]	School culture	.39	.25	1.56	.15	.09	1.78
Sc	BYS58F=MINOR	VANDALISM OF SCHOOL PROPERTY A PROBLEM	.35	.21	1.67	.16	.09	1.86
Sc	BYP74E=STRONGLY AGREE	MY CHILD ENJOYS SCHOOL	.35	.23	1.52	.15	.09	1.68
Sc	BYP74D=STRONGLY AGREE	MY CHILD IS WORKING HARD AT SCHOOL	.33	.2	1.65	.16	.09	1.82
Sc	BYS58E=MINOR	ROBBERY OR THEFT A PROBLEM AT SCHOOL	.33	.21	1.57	.15	.09	1.72
Sc	BYP57F=ONCE OR TWICE	CONTACTED ABOUT SCHOOL FUND RAISING	.33	.23	1.43	.14	.09	1.56
Sc	BYP74C=STRONGLY AGREE	MY CHILD IS CHALLENGED AT SCHOOL	.32	.14	2.29	.21	.08	2.5
Sc	BYS58A=MODERATE	STUDENT TARDINESS A PROBLEM AT SCHOOL	.32	.19	1.68	.16	.09	1.84
Sc	BYS58B=MODERATE	STUDENT ABSENTEEISM A PROBLEM AT SCHOOL	.32	.17	1.88	.18	.09	2.07
Sc	BYS58D=MODERATE	PHYSICAL CONFLICTS AMONG STUD A PROBLEM	.31	.18	1.72	.17	.09	1.87
Sc	BYS58K=MINOR	VERBAL ABUSE OF TEACHERS A PROBLEM	.3	.15	2	.19	.09	2.16
Sc	BYP74F=STRONGLY AGREE	STANDARDS SET BY THE SCHL ARE REALISTIC	.27	.17	1.59	.15	.09	1.68
Sc	BYT3_23D.ALL=NO	HAS CONTINUING ED CREDIT TRAINING GIFTED	.26	.13	2	.19	.09	2.1
Sc	BYT3_23E.ALL=NO	NO SPECIAL TRAINING TEACHING GIFTED	.25	.11	2.27	.21	.09	2.35
Sc	BYT3_23B.ALL=NO	HAS UNDERGRAD CREDIT TRAINING GIFTED	.24	.13	1.85	.18	.09	1.92
Sc	BYS58C=MODERATE	STUDENTS CUTTING CLASS A PROBLEM AT SCHL	.23	.13	1.77	.17	.09	1.83
Sc	BYT3_23C.ALL=NO	HAS GRADUATE CREDIT TRAINING GIFTED	.23	.11	2.09	.19	.09	2.14
Sc	BYSC25=YES	SCHOOL HAS FORMAL ADMISSION PROCEDURES	.23	.12	1.92	.18	.09	1.98
Sc	BYS58E=MODERATE	ROBBERY OR THEFT A PROBLEM AT SCHOOL	.22	.13	1.69	.16	.09	1.74
P	BYP61EB=YES	8TH GRADER GOES TO HISTORY MUSEUMS	.44	.26	1.69	.16	.08	2.03
C	BYS51HC=NO	TALK TO OTHR ADULT ABT PERSONAL PROBLEMS	.66	.46	1.43	.14	.07	2.1

Table 73 continued

M	BYP56=EIGHTH GRADER	INFLUENTIAL IN CHILD TAKING FOREIGN LANG	.31	.21	1.48	.15	.09	1.59
M	hw_outsch=7-9 HOURS	Hours of homework outside of school	.2	.1	2	.19	.09	2.02

Cat = Category (where S = student, F = family, Sc = school, P = peer, C = community, M = multiple); TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving.

Table 74. Attribute-values associated with high 12th grade achievement among Black students with high parental education identified by association rule mining

Cat	Attribute-value	Variable label	TRP	FPR	PLR	Prec	FOR	RP
S	BYS48B=HIGHER SCH AFTR COLL	HOW FAR IN SCHL R'S MOTHER WANTS R TO GO	.57	.36	1.58	.57	.36	1.59
S	BYS67AB=DO NOT ATTEND	ATTEND SCIENCE AT LEAST ONCE A WEEK	.55	.3	1.83	.6	.35	1.73
S	BYS67C=ATTEND	ATTEND ALGEBRA AT LEAST ONCE A WEEK	.55	.3	1.83	.6	.35	1.73
S	BYPSEPLN=HIGHER SCH AFTR COLL	POST-SECONDARY EDUCATION PLANS	.55	.36	1.53	.56	.37	1.52
S	BYS60A=HIGH	R'S ABILITY GROUP FOR MATHEMATICS	.48	.22	2.18	.64	.36	1.81
S	BYS67B=DO NOT ATTEND	ATTEND REGULAR MATH AT LEAST ONCE A WEEK	.43	.25	1.72	.59	.39	1.52
F	BYP59A=YES	BELONG TO PARENT-TEACHER ORGANIZATION	.7	.49	1.43	.54	.33	1.65
F	BYS35H=HAVE	R'S FAMILY HAS A COMPUTER	.66	.42	1.57	.57	.33	1.73
F	BYS36A=3 OR MORE TIMES	DISCUSS PROGRAMS AT SCHOOL WITH PARENTS	.55	.35	1.57	.57	.37	1.55
F	BYP30=MA+	HIGHEST LEVEL OF EDUCATION R COMPLETED	.45	.27	1.67	.58	.39	1.51
Sc	BYP75=VERY SATISFIED	HOW SATISFIED WITH ED CHILD HAS RECEIVED	.65	.34	1.91	.61	.31	2.01
Sc	BYSC45D=YES	8TH GRADE FOREIGN LANG COURSES OFFERED	.58	.34	1.71	.59	.35	1.69
Sc	BYP57G=ONCE OR TWICE	CONTACTED ABOUT INFO FOR SCHOOL RECORDS	.57	.34	1.68	.58	.35	1.66

Cat = Category (where S = student, F = family, Sc = school); TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving.

Table 75. Additional conditions that increase associations between 12th grade achievement and 8th grade higher level math course-taking by income and parental education subgroups (Study 1)

Grp	Additional condition	TPR	FPR	PLR	Prec	FOR	RP
LP	8th grader feels safe in school	.25	.06	4.27	.33	.08	3.92
LP	Attends algebra at least once a week	.30	.04	6.73	.44	.08	5.62
LP	Child seldom/never needs help with homework	.24	.05	4.91	.36	.08	4.3
LP	Expects college prep in high school	.26	.05	4.75	.35	.08	4.27
LP	Expects college prep in high school	.27	.05	5.3	.38	.08	4.66
LP	Family has no specific place for study	.35	.08	4.44	.34	.07	4.51
LP	Has not talked to other adult about discipline problems	.40	.09	4.59	.35	.07	4.94
LP	Has not talked to other adult about improving school work	.33	.08	4.24	.33	.08	4.26
LP	Has not talked to other adult about personal problems	.37	.08	4.56	.34	.07	4.7
LP	High ability grp (Eng)	.23	.04	5.41	.38	.08	4.52
LP	High locus of control	.29	.05	5.34	.38	.08	4.76
LP	High locus of control	.33	.07	4.7	.35	.08	4.57
LP	High locus of control	.27	.05	5.53	.39	.08	4.82
LP	High locus of control	.29	.04	6.96	.44	.08	5.63
LP	High locus of control	.29	.05	5.89	.4	.08	5.08
LP	High locus of control	.34	.06	5.78	.4	.07	5.33
LP	Older sibling never home when R returns from school	.3	.07	4.33	.33	.08	4.18
LP	Other adult relative never at home	.25	.06	4.09	.32	.08	3.82
LP	Other adult relative never at home	.28	.06	4.34	.33	.08	4.09
LP	Peers expect college	.37	.08	4.56	.34	.07	4.72
LP	Third friend attends same school	.26	.06	4.44	.34	.08	4.08
LI	23.5-32.8 Ss enrolled in class	.36	.08	4.23	.47	.13	3.66
LI	8th grader does not attend accelerated/advanced English class	.32	.07	4.66	.49	.13	3.72
LI	8th grader feels safe in school	.32	.05	6.6	.58	.13	4.45
LI	8th grader strongly disagrees that good luck is more important than hard work	.32	.07	4.8	.5	.13	3.78
LI	8th grader strongly disagrees that good luck is more important than hard work	.34	.07	4.94	.51	.13	3.93
LI	8th graders somewhat agrees that students in class sees them as important	.43	.09	4.8	.5	.12	4.35
LI	Adult neighbor never at home when R returns from school	.42	.09	4.6	.49	.12	4.19
LI	Attends algebra at least once a week	.35	.05	7.52	.61	.12	4.89
LI	BOTH (Algebra & algebra)	.45	.1	4.43	.48	.11	4.25
LI	BOTH (Algebra & High ability group for math)	.39	.06	6	.56	.12	4.61

Table 75 continued

LI	BOTH (Algebra & not regular math)	.38	.07	5.35	.53	.12	4.29
LI	BOTH (Algebra & not regular math)	.43	.1	4.41	.48	.12	4.13
LI	Drug use not a problem in school	.31	.05	5.69	.54	.13	4.1
LI	Family has rules about programs that child may watch	.39	.08	4.8	.5	.12	4.14
LI	Friends neither encourage/discourage student from taking algebra	.42	.1	4.14	.46	.12	3.92
LI	Has not talked to counselor about drug/alcohol abuse	.57	.14	4.11	.46	.09	4.87
LI	Has not talked to other adult about discipline problems	.43	.1	4.5	.48	.12	4.19
LI	Has not talked to other adult about drug/alcohol abuse	.47	.12	4.05	.46	.11	4.13
LI	Has not talked to other adult about HS programs	.32	.03	9.32	.66	.13	5.14
LI	Has not talked to other adult about improving school work	.38	.08	4.57	.49	.12	3.92
LI	Has not talked to other adult about personal problems	.4	.08	4.8	.5	.12	4.19
LI	High ability grp (Eng)	.35	.06	5.96	.55	.13	4.39
LI	High ability grp (Eng)	.31	.05	6.68	.58	.13	4.43
LI	High ability grp (Eng)	.36	.05	7.4	.61	.12	4.91
LI	High ability grp (Sci)	.31	.06	5.49	.53	.13	4.03
LI	High locus of control	.33	.04	7.77	.62	.13	4.84
LI	High locus of control	.34	.06	5.79	.55	.13	4.27
LI	High locus of control	.27	.02	11.2	.7	.13	5.19
LI	High locus of control	.33	.06	5.63	.54	.13	4.17
LI	High self-concept	.36	.08	4.67	.49	.13	3.89
LI	High self-concept	.33	.07	4.94	.51	.13	3.88
LI	Never had anything stolen at school	.29	.06	4.64	.49	.14	3.6
LI	Never held back	.55	.12	4.49	.48	.1	4.99
LI	No specific place for study	.38	.08	4.8	.5	.12	4.04
LI	Not been late for school in past 4 weeks	.39	.09	4.19	.47	.12	3.81
LI	Not contacted about behavior in school	.44	.1	4.51	.48	.11	4.24
LI	Not the case that relatives will pay for 8th grader's college	.43	.09	4.7	.49	.12	4.29
LI	P regularly talks to child about school experiences	.45	.1	4.43	.48	.11	4.25
LI	P strongly agrees that school places high priority on learning	.32	.05	6.34	.57	.13	4.37
LI	P very satisfied w child's education	.42	.08	5.42	.53	.12	4.6
LI	Parent agrees that parents work together supporting school policy	.4	.07	5.45	.53	.12	4.5
LI	Parent agrees that parents work together supporting school policy	.34	.08	4.42	.48	.13	3.68
LI	Parent does not think 8th grader's grades will not be high enough to qualify for financial aid	.45	.09	4.8	.5	.11	4.47

Table 75 continued

LI	Parent does not think 8th grader's grades will not be high enough to qualify for financial aid	.36	.07	4.8	.5	.13	3.95
LI	Parent does not think 8th grader's test scores will not be high enough to qualify for financial aid	.48	.1	4.71	.5	.11	4.6
LI	Parent does not think 8th grader's test scores will not be high enough to qualify for financial aid	.38	.08	4.92	.51	.12	4.15
LI	Parent does not think student will be able to earn own money for postsecondary ed	.27	.05	5.6	.54	.14	3.9
LI	Parent does not think student will be able to earn own money for postsecondary ed	.29	.07	4.36	.48	.14	3.47
LI	Parent has not contacted school about behavior	.41	.09	4.49	.48	.12	4.07
LI	Parent influence in assigning HS courses is moderate	.3	.06	5.13	.52	.13	3.84
LI	Parent influence in assigning HS courses is moderate	.24	.06	4.28	.47	.14	3.28
LI	Strongly agrees that on the whole s/he is satisfied with him/herself	.29	.07	4.36	.48	.14	3.47
LI	Student newspaper available to 8th grader	.41	.09	4.69	.49	.12	4.18
LI	Younger sibling never home when 8th grader gets home from school	.33	.06	5.44	.53	.13	4.09
LI	Younger sibling never home when 8th grader gets home from school	.33	.07	4.53	.49	.13	3.7
HI	8th grader disagrees that they don't have much to be proud of	.37	.06	5.68	.75	.26	2.89
HI	8th grader frequently discusses school activities with parents (student report)	.38	.05	7.57	.8	.25	3.16
HI	8th grader strongly agrees that they are a person of worth, equal to others	.34	.07	4.86	.71	.27	2.67
HI	Adult neighbor never at home when R returns from school	.37	.06	5.67	.75	.26	2.89
HI	Agrees that discipline is fair	.34	.06	6.18	.76	.26	2.88
HI	Agrees that discipline is fair	.41	.09	4.29	.69	.25	2.73
HI	Agrees that English will be useful for future	.26	.04	7.49	.79	.28	2.81
HI	Agrees that social studies will be useful in future	.29	.05	5.82	.75	.28	2.7
HI	Agrees that teachers are interested in students	.34	.06	5.67	.74	.27	2.8
HI	Agrees that the teaching is good	.3	.06	4.63	.7	.28	2.53
HI	Alcohol not a problem at school	.26	.06	4.37	.69	.29	2.41
HI	BOTH (Algebra & High ability group for math)	.33	.05	6	.76	.27	2.83
HI	BOTH (Algebra & not regular math)	.29	.07	4.16	.68	.28	2.42
HI	BOTH (Algebra and high ability group in math)	.38	.09	4.46	.7	.26	2.69
HI	Not in remedial English	.46	.1	4.35	.69	.24	2.9
HI	Cutting class is not a problem	.33	.06	5.08	.72	.27	2.68
HI	Debate/speech not available to 8th graders	.37	.08	4.61	.7	.26	2.7

Table 75 continued

HI	Did not participate in varsity sport	.27	.05	4.94	.72	.28	2.53
HI	Disagrees that chance and luck are important in life	.24	.03	6.94	.78	.29	2.71
HI	Disagrees that disruptions inhibit learning	.25	.05	4.59	.7	.29	2.43
HI	Does NOT attend art at least 1/wk	.35	.08	4.66	.71	.27	2.66
HI	Does NOT attend Biology at least once a week	.35	.08	4.37	.69	.27	2.59
HI	Does NOT attend computer ed at least once a week	.3	.05	6.02	.76	.27	2.75
HI	Does NOT attend earth science at least once a week	.29	.04	7.28	.79	.28	2.87
HI	Does NOT attend earth science at least once a week	.26	.05	4.77	.71	.29	2.48
HI	Does not attend sex ed at least once a week	.37	.08	4.61	.7	.26	2.7
HI	Teacher does not have BA in Education	.39	.09	4.32	.69	.26	2.68
HI	Drama club available to 8th graders	.31	.06	5.18	.73	.27	2.65
HI	Enrolled in gifted/talented program	.23	.03	7.77	.8	.29	2.77
HI	Expects college prep	.29	.04	7.28	.79	.28	2.87
HI	False that there is not enough information on financial aid	.23	.03	9.32	.83	.29	2.87
HI	Family has electric dishwasher	.28	.06	5.12	.73	.28	2.58
HI	Family has electric dishwasher	.28	.06	5.12	.73	.28	2.58
HI	Family has typewriter	.42	.1	4.4	.69	.25	2.79
HI	Family includes mother & father (vs missing one/both, or having a guardian for one/both)	.29	.06	4.85	.71	.28	2.55
HI	First friend attends same school	.35	.08	4.66	.71	.27	2.66
HI	Foreign language course offered in 8th grade	.37	.05	7.38	.79	.25	3.11
HI	Foreign language course offered in 8th grade	.26	.06	4.77	.71	.29	2.48
HI	Friends neither encourage/discourage student from taking algebra	.44	.09	4.6	.7	.24	2.9
HI	Full year of PE required	.32	.05	5.83	.75	.27	2.78
HI	Has NOT participate in scouting	.42	.09	4.39	.69	.25	2.79
HI	Has NOT participated in chorus	.35	.08	4.11	.68	.27	2.53
HI	Has NOT participated in neighborhood clubs/programs	.39	.09	4.32	.69	.26	2.68
HI	Has NOT participated in summer programs	.29	.07	4.16	.68	.28	2.42
HI	Has not talked to counselor about courses at school	.25	.05	5.05	.72	.29	2.5
HI	Has not talked to counselor about discipline problems	.45	.1	4.47	.7	.24	2.9
HI	Has not talked to counselor about drug/alcohol abuse	.47	.09	5.18	.73	.23	3.13
HI	Has not talked to counselor about jobs/careers after HS	.36	.06	5.99	.76	.26	2.91
HI	Has not talked to counselor about personal problems	.44	.09	4.86	.71	.24	2.96

Table 75 continued

HI	Has not talked to other adult about discipline problems	.33	.07	4.72	.71	.27	2.62
HI	Has not talked to other adult about drug/alcohol abuse	.43	.08	5.34	.73	.24	3.02
HI	Has not talked to other adult about personal problems	.28	.06	4.69	.71	.28	2.5
HI	Has not talked to teacher about discipline problems	.39	.08	5.18	.73	.25	2.86
HI	Has not talked to teacher about drug/alcohol abuse	.45	.1	4.7	.71	.24	2.96
HI	HAS talked to teacher about improving school work	.36	.08	4.49	.7	.26	2.64
HI	HAS talked to teacher about studies in class	.39	.08	4.85	.71	.26	2.8
HI	High locus of control	.27	.06	4.53	.7	.29	2.45
HI	High locus of control	.35	.07	4.66	.71	.27	2.65
HI	High locus of control	.3	.06	5.47	.74	.28	2.68
HI	High locus of control	.38	.09	4.21	.68	.26	2.63
HI	High locus of control	.33	.07	4.72	.71	.27	2.62
HI	High self-concept	.36	.08	4.79	.71	.26	2.71
HI	High self-concept	.33	.07	4.72	.71	.27	2.62
HI	High SES	.26	.02	1.49	.84	.28	3.01
HI	High SES	.35	.08	4.66	.71	.27	2.66
HI	High SES	.28	.05	5.63	.74	.28	2.65
HI	Lives in household with father	.29	.06	4.48	.7	.28	2.48
HI	Low percentage of Ss in remedial reading	.45	.1	4.25	.69	.24	2.84
HI	Low percentage of Ss in special education	.35	.06	5.83	.75	.26	2.86
HI	LT 1 hr outside of school hours spent in record keeping	.21	.05	4.75	.71	.3	2.38
HI	Math club available to 8th graders	.24	.05	4.42	.69	.29	2.38
HI	Minor action for first occurrence of cheating	.42	.1	4.18	.68	.25	2.73
HI	Never been held back	.47	.1	4.66	.71	.23	3.02
HI	Never held back	.42	.1	4.18	.68	.25	2.73
HI	No disruptions	.33	.08	4.13	.68	.27	2.49
HI	No specific amount of instructional time required for family life/sex ed	.31	.06	5.65	.74	.27	2.73
HI	No time outside of school hours spent coordinating curriculum	.32	.06	4.93	.72	.27	2.63
HI	Not at all accurate that teachers have negative attitude about students	.27	.05	5.44	.74	.28	2.6
HI	Not been late for school in past 4 weeks	.3	.05	6.02	.76	.27	2.75
HI	Not enrolled in bilingual education	.48	.12	4.14	.68	.23	2.91
HI	Not missed school in past 4 weeks	.37	.08	4.61	.7	.26	2.7
HI	Not single parent	.33	.06	5.5	.74	.27	2.75
HI	Not the case that parent can't see a way to get money for college	.33	.06	5.08	.72	.27	2.68

Table 75 continued

HI	Not too much work to apply for college financial aid	.39	.04	8.63	.82	.25	3.29
HI	Once in a while feels bored at school	.31	.07	4.78	.71	.28	2.58
HI	Other adult relative never at home	.28	.03	9.39	.83	.28	3
HI	P regularly talks to child about school experiences	.36	.08	4.49	.7	.26	2.64
HI	P very satisfied w child's education	.33	.07	4.72	.71	.27	2.62
HI	Parent does not think 8th grader's grades will not be high enough to qualify for financial aid	.38	.06	6.31	.76	.25	3.01
HI	Parent does not think 8th grader's test scores will not be high enough to qualify for financial aid	.37	.05	6.71	.78	.26	3.03
HI	Parent does not think family income would be too high for college financial aid	.27	.04	6.04	.76	.28	2.68
HI	Parent does not think student will be able to earn own money for postsecondary ed	.28	.04	7.04	.78	.28	2.82
HI	Parent does not think that they would be able to pay for 8th graders' postsecondary ed w/o assistance	.29	.06	4.86	.71	.28	2.55
HI	Parent knows first name of 8th grader's friends	.42	.1	4.18	.68	.25	2.73
HI	Parent/guardians wanted 8th grader to take Algebra	.33	.07	5.08	.72	.27	2.68
HI	Physical abuse of teachers is not a problem	.42	.09	4.64	.7	.25	2.84
HI	Physical abuse of teachers is not a problem	.46	.11	4.15	.68	.24	2.85
HI	Plans to attend public school	.4	.1	4.19	.68	.26	2.68
HI	R knows parent(s) of child's first friend	.35	.08	4.11	.68	.27	2.53
HI	Race of teacher is White, non-Hispanic	.32	.06	5.34	.73	.27	2.7
HI	Religious organization NOT available to 8th graders	.43	.09	4.75	.71	.24	2.9
HI	School is departmentalized	.43	.1	4.27	.69	.25	2.79
HI	Science fair available to 8th graders	.34	.07	4.85	.71	.27	2.67
HI	Second friend attends same school	.31	.06	5.18	.73	.27	2.65
HI	Seldom comes to class without pencil/paper	.3	.05	6.02	.76	.27	2.75
HI	Seldom comes to class without books	.23	.02	11.65	.86	.29	2.98
HI	Student newspaper available to 8th grader	.38	.08	5.05	.72	.26	2.81
HI	Student newspaper available to 8th grader	.35	.08	4.11	.68	.27	2.53
HI	Suspension for first occurrence of alcohol possession	.4	.09	4.68	.71	.25	2.79
HI	Suspension for first occurrence of alcohol use	.35	.08	4.66	.71	.27	2.65
HI	Suspension for first occurrence of illegal drug use	.33	.06	5.08	.72	.27	2.68
HI	Suspension for first occurrence of injury to other students	.29	.05	5.3	.73	.28	2.63
HI	Suspension for first occurrence of weapon possession	.28	.06	5.12	.72	.28	2.58
HI	Suspension for repeated occurrence of smoking	.41	.08	5.1	.72	.25	2.91

Table 75 continued

HI	Teachers have a lot of influence in assigning HS courses	.31	.06	4.78	.71	.28	2.58
HI	Transfers not allowed	.41	.09	4.8	.71	.25	2.85
HI	Ts agree that textbooks interesting to most Ss	.26	.06	4.37	.69	.29	2.41
HI	Vandalism is a minor problem	.28	.06	4.69	.71	.28	2.5
HI	Very much accurate the students are expected to do HW	.34	.08	4.53	.7	.27	2.6
HI	Younger sibling never home when 8th grader gets home from school	.32	.08	4.27	.69	.27	2.5

Grp =Group (LP = low parental education, HP = high parental education, LI = low-income, HI = high income); TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving.

APPENDIX F

ASSOCIATION RULES FOR STUDY 2

Table 76. Attribute-values associated with higher than expected 12th grade math achievement identified by association rule mining

Cat	Attribute-value	Variable label	TRP	FPR	PLR	Prec	FOR	RP
0-16	BYP85F=FALSE	CHLD TEST SCORES NOT GOOD ENOUGH QUALIFY	.49	.30	1.67	.16	.07	2.12
0-16	BYP85E=FALSE	CHILD GRADES NOT HIGH ENOUGH TO QUALIFY	.46	.28	1.61	.15	.08	1.95
0-16	BYP76=4-5YR COLLEG PGM	HOW FAR IN SCHOOL R EXPECT CHILD TO GO	.41	.25	1.62	.15	.08	1.90
0-16	Pexp=4YR DEG	Parent expectations after high school	.41	.25	1.62	.15	.08	1.90
0-16	F1C13C=75% TO 100%	% IN ACADEMIC COUNSELING PROGRAM STUDENTS ARE	.41	.24	1.72	.16	.08	2.02
0-16	F1C93E=VERY ACCURATE	EXPECTED TO DO HOMEWORK	.39	.25	1.57	.15	.08	1.79
0-16	BYSC46V=NO	CHEERLEADING, ETC. AVAIL TO 8TH GRADERS NOT MUCH	.34	.22	1.57	.15	.09	1.73
0-16	BYP85H=FALSE	INFORMATION ON FINANCIAL AID	.34	.21	1.60	.15	.09	1.77
0-16	F2T3_7F=AGREE	NECESSARY MATERIALS READILY AVAILABLE	.33	.20	1.67	.16	.08	1.85
0-16	BYS62=YES	DID PRNTS/GRDNS WANT R TO TAKE ALGEBRA	.33	.20	1.69	.16	.08	1.87
0-16	geo_none=0	0 years of Geometry	.33	.17	1.95	.18	.08	2.16
0-16	F2HSPROG=ACADEMIC PROGRAM	RESPONDENT-INDICATED HIGH SCHOOL PROGRAM	.29	.17	1.67	.16	.09	1.80
0-16	BYT3_29.SOC.STUDIES.HISTORY=SEVERAL TIMES	HOW OFTEN SUPERVISOR OBSERVED TEACHING	.27	.17	1.62	.15	.09	1.73
17-23	F2HSPROG=ACADEMIC PROGRAM	RESPONDENT-INDICATED HIGH SCHOOL PROGRAM	.54	.33	1.63	.17	.08	2.13
17-23	F2T3_7F=AGREE	NECESSARY MATERIALS READILY AVAILABLE	.54	.30	1.63	.17	.09	2.01
17-23	F2T3_7H=AGREE	GRADING PRACTICES CONSISTENT AND FAIR	.52	.33	1.55	.16	.08	1.94

Table 76 continued

17-23	F2T3_5A=AGREE	ENCOURAGED TO EXPERIMENT WITH TEACHING	.50	.29	1.75	.18	.08	2.23
17-23	F2T3_16D=MINOR PROBLEM	DEGREE ROBBERY OR THEFT A PROBLEM	.43	.28	1.55	.16	.09	1.81
17-23	F2T3_16M=MINOR PROBLEM	DGREE STUS UNDR INFL DRUGS/ALCHL A PRBLM	.43	.26	1.65	.17	.09	1.94
17-23	F2T3_16I=MINOR PROBLEM	DEGREE USE OF ILLEGAL DRUGS A PROBLEM	.42	.25	1.66	.17	.09	1.93
17-23	F2T3_5J=DISAGREE	ROUTINE DEPT DUTIES INTERFERE W/TEACHING	.39	.23	1.66	.17	.09	1.89
17-23	F2T3_5G=AGREE	DEPT COMMITTED TO AP AND HONORS COURSES	.34	.19	1.77	.18	.09	1.95
17-23	F2T3_16J=MINOR PROBLEM	DEGREE POSSESSION OF WEAPONS A PROBLEM	.33	.17	1.89	.19	.09	2.07
17-23	F2T3_6D=AGREE	DEPT CHAIR CONSULTS STAFF BEFOR DECISION	.32	.19	1.69	.18	.10	1.85
17-23	F2T4_19G=YES	HIGHER ORDER THINKING SKILLS DISCUSSED	.32	.18	1.75	.18	.10	1.90
17-23	F2T4_19F=YES	COOPERATIVE LEARNING DISCUSSED	.30	.17	1.80	.19	.10	1.94
17-23	al2_none=0	0 years of Algebra 2	.30	.16	1.91	.19	.10	2.05
24-30	F1S29D=ONCE A WEEK	WRITE RPTS OF LABORATORY WORK IN SCIENCE	.49	.32	1.53	.14	.08	1.89
24-30	al2_none=0	0 years of Algebra 2	.49	.26	1.87	.17	.07	2.41
24-30	al2_1=1	1 year of Algebra 2	.43	.21	2.01	.18	.07	2.46
24-30	BYP55=YES	CHILD ENROLLED IN FOREIGN LANG COURSE	.42	.26	1.61	.15	.08	1.91
24-30	BYS57C=ONCE OR TWICE	SOMEONE THREATENED TO HURT R AT SCHOOL	.29	.19	1.54	.14	.09	1.65
24-30	BYP8=1940 - 1944	R'S YEAR OF BIRTH	.26	.17	1.57	.15	.09	1.67

Cat = 8th grade math achievement cateogory; TPR = True positive rate, or P(A|B); FPR = False positive rate, or P(A|¬B); PLR = positive likelihood ratio or TPR/FPR; Precision = P(B|A); FOR = False omission rate, or P(B|¬A); RP = relative probability = Precision/FOR, where P(A) is probability that rule antecedent applies, and P(B) is probability that student is high achieving.

APPENDIX G

ANALYSIS FLOW & SYNTAX FOR STUDY 1 DATA MINING

Note: See Table 18 for list and references to R packages for each rule induction algorithm. Other R packages used for data preparation and visualization include plyr (Wickham, 2011), caret (Wing et al., 2016), sas7bdat (Shotwell, 2014), dplyr (Wickham & Francois, 2016), magrittr (Bache & Wickham, 2014), methods (R Core Team, 2016), rattle (Williams, 2011) and haven (Wickham & Miller, 2016).

Create large dataset for data mining

1. Using SAS software, create NELS datafiles (SAS format) that saves values as data rather than as labels.
2. In R, save each of the files above as an R datafile, then for each,
    a. Convert missing values to NA
    b. Set numeric variables to numeric
    c. Set ordinal variables to ordered factors
    d. Set categorical variables to factors
    e. Save file as e.g., nels_by_stu_clean1.Rda, nels_by_sch_clean1.Rda, etc.
3. Open dataset that was used for replication study.
4. Convert missing to NA
5. Recode ordinal variables so they are correctly configured as ordered factors (vs being nominal)
6. Append other variables onto this dataset using student or school ID (making sure to avoid merging in duplicate variables along the way to reduce clutter). Datasets appended (located in C:/Users/Emi/Dropbox/Dissertation/Data/NELS88_Large_dataset_creation) include:
    a. nels_by_stu_clean1.Rda
    b. nels_by_sch_clean1.Rda
    c. nels_by_par_clean1.Rda
    d. nels_by_tea_clean1_wide.Rda*
    e. nels_f1_stu_clean1.Rda
    f. nels_f1_sch_clean1.Rda
    g. nels_f1_tea_clean1_wide.Rda*
    h. nels_f2_stu1_clean1.Rda
    i. nels_f2_stu2_clean1.Rda
    j. nels_f2_par_clean1.Rda
    k. nels_f2_sch_clean1.Rda
    l. nels_f2_tea_clean1.Rda

*note: For the teacher datasets, when more than one teacher rated the student/school, I took the average for numeric variables, and the mode for categorical variables. If there were two or more modes, I took the first value. Also, after merging this dataset in, I correctly designated the variable types.

7. Drop all variables beginning with F1 and F2 (7967 variables → 2077 variables).
8. Drop variable that are irrelevant or correlate with the outcome. (2077→ 1994)
9. Drop near zero and near-zero variance predictors (1994→1530)
10. Drop duplicate predictors (1530→1469)
11. Drop unused levels.
12. Drop variables with >=95% missing. (1162+/1223 missing; 1469→1373)
13. Examined the 33 categorical variables with 10 or more levels and reduced levels for 31 of them.
14. Save dataset (nels_thomasTLA_final1a.Rda).

Ruleset mining

C45, PART, RIPPER, C5.0, boosted C5.0, CART, bagged CART, Random Forest, QUEST

1. Open large dataset created above (nels_thomasTLA_final1a.Rda).
2. [For bagged CART] Retain variables with fewer than 50% NAs (1373 → 706 variables)
3. Shuffle the data
4. [For C5.0, rCBA] Recode all numeric variables in dataset into ordered factors (split into 4 groups)
5. [For rCBA] Get rid of all commas in the dataset
6. [For randomForest] Substitute missing values (mean for numeric, and "missing" category for factors)
7. [For randomForest] Convert every factor to character, all NA to "missing", then characters back to factor.
8. Shuffle the dataset
9. Stratified random sampling into training and test sets (70-30 split, stratified by outcome)
10. [For C4.5, PART, RIPPER, randomForest] Substitute missing values (mean for numeric, and "missing" category for factors)
11. Grow final C4.5 tree (first exploring different parameters and check prediction on holdout data.)

```
##FINAL C4.5 model
#adjusted the cost matrix to be 3:1--penalty for misclassifying the high achievers
#use reduced error pruning (R=TRUE; 1/4 used for pruning).
#min in leaf = 2% of high achievers = 4
model1<- CostSensitiveClassifier(highach ~ ., data = training.data3,
            control = Weka_control(`cost-matrix` = matrix(c(0, 3, 1, 0), ncol = 2),
                                   W = list(J48, M=3, R=TRUE, N=3, S=FALSE,
                                   Q=1234)))
model1
summary(model1)
## predict on holdout data
```

```
model1p <- predict(model1, test.data3)
table(test.data$highach, predicted = model1p)
plot(model1p)
```

12. Grow PART

```
model2<- CostSensitiveClassifier(highach ~ ., data = training.data3,
        control = Weka_control(`cost-matrix` = matrix(c(0, 3, 1, 0), ncol = 2),
                W = list(PART, M=3, R=TRUE, N=3, S=FALSE,
                Q=1234)))
model2
summary(model2)
## predict on holdout data
model2p <- predict(model2, test.data3)
table(test.data$highach, predicted = model2p)
plot(model2p)
```

13. Grow RIPPER

```
model2<- CostSensitiveClassifier(highach ~ ., data = training.data3,
        control = Weka_control(`cost-matrix` = matrix(c(0, 3, 1, 0), ncol = 2),
                W = "weka.classifiers.rules.JRip",
                S = 1234))
model2
summary(model2)
## predict on holdout data
model2p <- predict(model2, test.data3)
table(test.data$highach, predicted = model2p)
plot(model2p)
```

14. Grow C5.0 and boosted C5.0

```
# Cost matrix
costmatrix1<- matrix (c(0, 1, 3, 0), nrow = 2, ncol = 2)
dimnames(costmatrix1) <- list(c("No", "Yes"), c("No", "Yes"))

# C50 tree (create without boosting)
library(C50)
#Find out where the predictor is
match("highach",names(training.data))

#Final non-boosted model (after trying many non-boosted models) (only for final,
change sample from .5 to 0)
model1 <- C5.0(training.data[, -8], training.data[, 8],
                trials=1, rules=FALSE,
                control = C5.0Control(subset = TRUE, bands = 0, winnow = TRUE,
                noGlobalPruning = FALSE, CF = 0.05, minCases = 2, sample = 0,
                fuzzyThreshold = FALSE, seed = 1234, label = "highach"),
                costs=costmatrix1)
```

```
summary(model1)
## predict on holdout data
model1p <- predict(model1, test.data)
model1p
table(test.data$highach,predicted = model1p)
plot(model1p)

## Final boosted model  (change "sample" from .5 to 0)
model2 <- C5.0(training.data[, -8], training.data[, 8],trials=20,
                    control = C5.0Control(subset = TRUE, bands = 0, winnow = FALSE,
                    noGlobalPruning = FALSE, CF = 0.05, minCases = 2, sample = 0,
                    fuzzyThreshold = FALSE, seed = 1234, label = "highach"),
                    costs=costmatrix1)
summary(model2)

## predict on holdout data
model2p <- predict(model2, test.data)
model2p
table(test.data$highach,predicted = model2p)
plot(model2p)
```

15. Grow CART

```
library(rpart)
Model_rpart1 <- rpart(highach ~ . ,
                method = "class",
                data = training.data,
                xval = 10,
                control = rpart.control(minbucket=0, cp=0),
                parms=list(split = "gini", loss=matrix(c(0, 1, 3, 0), byrow=TRUE, ncol
= 2)),
)

#print(Model_rpart1) # results
printcp(Model_rpart1) # display the results
plotcp(Model_rpart1) # visualize cross-validation results, see where to prune
# summary(Model_rpart1) # detailed summary of splits
pModel_rpart1 <-prune(Model_rpart1, cp=0.06) #prune tree based on plotcp
print(pModel_rpart1)
summary(pModel_rpart1)
pred <- predict(pModel_rpart1, test.data, type="class")
summary(pred)

### plot tree the prettiest way
#function that wraps strings at desired width (adjust "width=")
```

```
split.fun <- function(x, labs, digits, varlen, faclen)
{
 # replace commas with spaces (needed for strwrap)
 labs <- gsub(",", " ", labs)
 for(i in 1:length(labs)) {
  # split labs[i] into multiple lines
  labs[i] <- paste(strwrap(labs[i], width=25), collapse="\n")
 }
 labs
}
#Tweak adjusts the font size, gap & space adjust space between boxes
#& between text & box.
library(rattle)
fancyRpartPlot(pModel_rpart1, sub="CART on Thomas' data (highach vs not)",
            split.fun=split.fun, tweak=1, gap=1, space=1
            )


# create confusion matrix
CM1 <- table(test.data$highach, pred)
rownames(CM1) <- paste("Actual", rownames(CM1), sep = ":")
colnames(CM1) <- paste("Pred", colnames(CM1), sep = ":")
print(CM1)

#Ruleset
library(rattle)
asRules(pModel_rpart1, compact=FALSE)
```

16. Grow bagged CART

```
library(ipred)
model1<- bagging(highach~., data=training.data, control=rpart.control(minbucket=0,
cp=0))
print(model1)
pred <- predict(model1, test.data, type="class")
# create confusion matrix
CM1 <- table(test.data$highach, pred)
rownames(CM1) <- paste("Actual", rownames(CM1), sep = ":")
colnames(CM1) <- paste("Pred", colnames(CM1), sep = ":")
print(CM1)
```

17. Grow Random Forest

```
library(randomForest)
set.seed(1234)
model1 <- randomForest(highach ~ ., training.data, importance=TRUE)
```

```
print(model1)
## predict on holdout data
model1p <- predict(model1, test.data)
table(test.data$highach, predicted = model1p)
plot(model1p)

##Variable importance
#p.6 (https://cran.r-project.org/web/packages/randomForest/randomForest.pdf)
varImpPlot(model1, n.var=30, main="Thomas_TLA_RF \nImportant Variables")
out1 <- importance(model1, type=1)
write.csv(out1, file = "TLA_RF_VarImp1.csv")
out2 <- importance(model1, type=2)
write.csv(out2, file = "TLA_RF_VarImp2.csv")
```

18. Grow QUEST (in SPSS) (Note: Use training and test sets created for CART. Use haven package to save R dataset as SPSS dataset)

```
* Decision Tree.
TREE highach [n] BY par_ed [o] G8MINOR [o] G8LUNCH [o] hw_sch [o]
hw_outsch [o] safety [o] income91 [o] female [n] peerexcl [n] admintel [n]
SINPAR [n] privsch [n] religsch [n] public [n] urban [n] hhressc [s] autonomy [s]
pinvolve [s] pexpcol [n] goodpeer [s] badpeer [s] activity [s] sculture [s] climate [s]
diversassg [n] unsafe [n] disrupt [n] BYS4A [n] BYS4OCC [n] BYS5A [n] BYS7A
[n] BYS7OCC [n] BYS8A [n] BYS8B [n] BYS8C [n] BYS8D [n] BYS8E [n]
BYS8F [n] BYS8G [n] BYS8H [n] BYS12 [n] BYS14 [n] BYS15 [n] BYS16 [n]
BYS21 [n] BYS22 [n] BYS23 [n] BYS24 [n] BYS25A [o] BYS25B [o] BYS25C
[o] BYS25D [o] BYS26A [o] BYS26B [o] BYS26C [o] BYS26D [o] BYS26E [o]
BYS26F [o] BYS26G [o] BYS26H [o] BYS26I [o] BYS27A [o] BYS27B [o]
BYS27C [o] BYS27D [o] BYS28B1 [n] BYS28B3 [n] BYS28D1 [n] BYS28D3
[n] BYS28E1 [n] BYS28E3 [n] BYS28F1 [n] BYS28F3 [n] BYS29 [n] BYS32 [o]
BYS33 [o] BYS34A [n] BYS34B [n] BYS35A [n] BYS35B [n] BYS35C [n]
BYS35D [n] BYS35E [n] BYS35G [n] BYS35H [n] BYS35I [n] BYS35J [n]
BYS35K [n] BYS35L [n] BYS35M [n] BYS35N [n] BYS35O [n] BYS35P [n]
BYS36A [o] BYS36B [o] BYS36C [o] BYS37A [n] BYS37B [n] BYS37C [n]
BYS37D [n] BYS38A [o] BYS38B [o] BYS38C [o] BYS38D [o] BYS39A [n]
BYS39B [n] BYS39C [n] BYS40A [o] BYS40B [o] BYS40C [o] BYS40E [o]
BYS40F [o] BYS40G [o] BYS40H [o] BYS41 [o] BYS42A [o] BYS42B [o]
BYS44A [o] BYS44B [o] BYS44C [o] BYS44D [o] BYS44E [o] BYS44F [o]
BYS44G [o] BYS44H [o] BYS44I [o] BYS44J [o] BYS44K [o] BYS44L [o]
BYS44M [o] BYS46 [o] BYS47 [o] BYS48A [n] BYS48B [n] BYS49 [n]
BYS50A [o] BYS50B [o] BYS50C [o] BYS50D [o] BYS50E [o] BYS50F [o]
BYS51AA [n] BYS51AB [n] BYS51AC [n] BYS51BA [n] BYS51BB [n] BYS51BC
[n] BYS51CA [n] BYS51CB [n] BYS51CC [n] BYS51DA [n] BYS51DB [n]
BYS51DC [n] BYS51EA [n] BYS51EB [n] BYS51EC [n] BYS51FA [n]
BYS51FB [n] BYS51FC [n] BYS51GA [n] BYS51GB [n] BYS51GC [n] BYS51HA
```

[n] BYS51HB [n] BYS51HC [n] BYS52 [n] BYS53 [o] BYS54 [n] BYS55A [o] BYS55B [o] BYS55C [o] BYS55D [o] BYS55E [o] BYS55F [o] BYS56A [o] BYS56B [o] BYS56C [o] BYS56D [o] BYS56E [o] BYS57A [o] BYS57B [o] BYS57C [o] BYS58A [o] BYS58B [o] BYS58C [o] BYS58D [o] BYS58E [o] BYS58F [o] BYS58G [o] BYS58H [o] BYS58I [o] BYS58J [o] BYS58K [o] BYS59A [o] BYS59B [o] BYS59C [o] BYS59D [o] BYS59E [o] BYS59F [o] BYS59G [o] BYS59H [o] BYS59I [o] BYS59J [o] BYS59K [o] BYS59L [o] BYS59M [o] BYS60A [n] BYS60B [n] BYS60C [n] BYS60D [n] BYS61 [n] BYS62 [n] BYS63 [n] BYS64 [n] BYS65 [n] BYS66A [n] BYS66B [n] BYS66C [n] BYS66D [n] BYS67A [n] BYS67B [n] BYS67C [n] BYS67AA [n] BYS67AB [n] BYS67AC [n] BYS67AD [n] BYS67BA [n] BYS67BB [n] BYS67BC [n] BYS67BD [n] BYS67BE [n] BYS67BF [n] BYS67BG [n] BYS67BH [n] BYS67CA [n] BYS67CB [n] BYS67CC [n] BYS67CD [n] BYS67CE [n] BYS67DA [n] BYS67DB [n] BYS67DC [n] BYS67DD [n] BYS68A [n] BYS68B [n] BYS69A [o] BYS69B [o] BYS69C [o] BYS70A [o] BYS70B [o] BYS70C [o] BYS71A [o] BYS71B [o] BYS71C [o] BYS72A [o] BYS72B [o] BYS72C [o] BYS73 [o] BYS74 [n] BYS74A [n] BYS74B [n] BYS74C [n] BYS74D [n] BYS74E [n] BYS74F [n] BYS74G [n] BYS74H [n] BYS74I [n] BYS75 [o] BYS76 [o] BYS77 [o] BYS78A [o] BYS78B [o] BYS78C [o] BYS79A [o] BYS79B [o] BYS79C [o] BYS79D [o] BYS79E [o] BYS80 [o] BYS82A [n] BYS82B [n] BYS82C [n] BYS82D [n] BYS82E [n] BYS82F [n] BYS82G [n] BYS82I [n] BYS82J [n] BYS82K [n] BYS82L [n] BYS82M [n] BYS82N [n] BYS82O [n] BYS82P [n] BYS82Q [n] BYS82R [n] BYS82S [n] BYS82T [n] BYS82U [n] BYS83A [n] BYS83B [n] BYS83C [n] BYS83D [n] BYS83E [n] BYS83F [n] BYS83G [n] BYS83H [n] BYS83I [n] BYS83J [n] G8TYPE [n] G8CTRL [n] BYSCENRL [o] G8ENROL [o] G8URBAN [n] G8REGON [n] NOMSECT [n] SEX [n] HANDPAST [n] BYLOCUS1 [s] BYLOCU1T [o] BYLOCUS2 [s] BYLOCU2T [o] BYCNCPT1 [s] BYCNCP1T [o] BYCNCPT2 [s] BYCNCP2T [o] BYSES [s] BYSESQ [o] BYPARED [n] BYFAMSIZ [o] BYFCOMP [n] BYPARMAR [n] BYFAMINC [o] BYPSEPLN [n] BYHOMEWK [o] BYLM [n] BYSC6 [o] BYSC7 [s] BYSC9H [s] BYSC11 [s] BYSC12 [s] BYSC13A [o] BYSC13B [o] BYSC13C [o] BYSC13D [o] BYSC13E [o] BYSC14 [o] BYSC15 [o] BYSC16B [s] BYSC16C [s] BYSC16E [s] BYSC16F [s] BYSC16G [s] BYSC17 [o] BYSC18 [n] BYSC19 [o] BYSC20B [o] BYSC20C [o] BYSC20D [o] BYSC20E [o] BYSC21 [s] BYSC22 [s] BYSC23 [n] BYSC24A [n] BYSC24B [n] BYSC24C [n] BYSC24E [n] BYSC24F [n] BYSC25 [n] BYSC26 [o] BYSC27 [o] BYSC28A [o] BYSC28B [o] BYSC28C [o] BYSC28D [o] BYSC28E [o] BYSC28F [o] BYSC28G [o] BYSC28H [o] BYSC29 [n] BYSC30 [n] BYSC31 [o] BYSC32 [o] BYSC33 [s] BYSC34 [o] BYSC35 [n] BYSC36A [o] BYSC36B [o] BYSC36C [o] BYSC36D [o] BYSC37 [o] BYSC38A [n] BYSC38B [n] BYSC38C [n] BYSC38D [n] BYSC38E [n] BYSC38F [n] BYSC38G [n] BYSC39D [n] BYSC39E [n] BYSC39F [n] BYSC39G [n] BYSC39H [n] BYSC39I [n] BYSC39J [n] BYSC39K [n] BYSC39L [n] BYSC39M [n] BYSC40 [n] BYSC41A [n]

BYSC41B [n] BYSC41C [n] BYSC41D [n] BYSC41E [n] BYSC41F [n] BYSC41G [n] BYSC41H [n] BYSC41I [n] BYSC42 [n] BYSC43 [n] BYSC44B [n] BYSC44C [n] BYSC44D [n] BYSC44E [n] BYSC44F [n] BYSC44G [n] BYSC44H [n] BYSC44I [n] BYSC45A [n] BYSC45B1 [n] BYSC45B2 [n] BYSC45B3 [n] BYSC45B4 [n] BYSC45C2 [n] BYSC45D [n] BYSC46A [n] BYSC46B [n] BYSC46C [n] BYSC46D [n] BYSC46E [n] BYSC46F [n] BYSC46G [n] BYSC46H [n] BYSC46I [n] BYSC46J [n] BYSC46K [n] BYSC46L [n] BYSC46M [n] BYSC46N [n] BYSC46O [n] BYSC46P [n] BYSC46Q [n] BYSC46R [n] BYSC46S [n] BYSC46T [n] BYSC46U [n] BYSC46V [n] BYSC47A [o] BYSC47B [o] BYSC47C [o] BYSC47D [o] BYSC47E [o] BYSC47F [o] BYSC47G [o] BYSC47H [o] BYSC47I [o] BYSC47J [o] BYSC47K [o] BYSC47L [o] BYSC47M [o] BYSC47N [o] BYSC47O [o] BYSC48A [n] BYSC48B [n] BYSC48C [n] BYSC48D [n] BYSC48E [n] BYSC48F [n] BYSC48H [n] BYSC48I [n] BYSC48J [n] BYSC49A [o] BYSC49B [o] BYSC49C [o] BYSC49D [o] BYSC49E [o] BYSC49F [o] BYSC49G [o] BYSC49H [o] BYSC49I [o] BYSC49J [o] BYSC49K [o] BYSC50AA [o] BYSC50AB [o] BYSC50AC [o] BYSC50AD [o] BYSC50AE [o] BYSC50AF [o] BYSC50AG [o] BYSC50AH [o] BYSC50AI [o] BYSC50AJ [o] BYSC50AK [o] BYSC50AL [o] BYSC50AM [o] BYSC50BA [o] BYSC50BB [o] BYSC50BC [o] BYSC50BD [o] BYSC50BE [o] BYSC50BF [o] BYSC50BG [o] BYSC50BH [o] BYSC50BI [o] BYSC50BJ [o] BYSC50BK [o] BYSC50BL [o] BYSC50BM [o] G8SUBS [n] BYSCORG2 [n] BYRATIO [o] BYP1A1 [n] BYP1A2 [n] BYP2 [o] BYP3A [o] BYP3B [o] BYP4 [o] BYP5A [o] BYP5B [o] BYP6 [o] BYP7 [n] BYP8 [o] BYP9 [n] BYP11 [n] BYP14 [n] BYP22A [n] BYP29 [n] BYP30 [o] BYP31 [o] BYP32 [n] BYP33A [n] BYP33B [n] BYP34A [n] BYP34B [n] BYP35 [n] BYP36A [n] BYP36B [n] BYP37A [n] BYP37B [n] BYP38A [n] BYP38B [n] BYP38C [n] BYP38D [n] BYP39 [o] BYP40 [o] BYP44 [n] BYP45A [n] BYP45B [n] BYP45C [n] BYP46B [n] BYP46C [n] BYP46D [n] BYP46E [n] BYP46F [n] BYP46G [n] BYP46H [n] BYP46I [n] BYP47J [n] BYP50 [n] BYP51 [n] BYP52A [o] BYP52B [o] BYP52C [o] BYP52D [o] BYP52E [o] BYP53 [n] BYP54 [n] BYP55 [n] BYP56 [n] BYP57A [o] BYP57B [o] BYP57C [o] BYP57D [o] BYP57E [o] BYP57F [o] BYP57G [o] BYP57H [o] BYP58A [o] BYP58B [o] BYP58C [o] BYP58D [o] BYP58E [o] BYP58F [o] BYP59A [n] BYP59B [n] BYP59C [n] BYP59D [n] BYP59E [n] BYP60A [n] BYP60B [n] BYP60C [n] BYP60E [n] BYP60F [n] BYP60G [n] BYP60H [n] BYP61AA [n] BYP61AB [n] BYP61BA [n] BYP61BB [n] BYP61CA [n] BYP61CB [n] BYP61DA [n] BYP61DB [n] BYP61EA [n] BYP61EB [n] BYP62 [n] BYP62A1 [n] BYP62B1 [n] BYP62A2 [n] BYP62B2 [n] BYP62A3 [n] BYP62B3 [n] BYP62A4 [n] BYP62B4 [n] BYP62A5 [n] BYP62B5 [n] BYP63A [n] BYP63B [n] BYP63D [n] BYP63E [n] BYP63F [n] BYP63G [n] BYP63H [n] BYP63I [n] BYP64A [n] BYP64B [n] BYP64C [n] BYP64D [n] BYP65A [n] BYP65B [n] BYP65C [n] BYP66 [o] BYP67 [o] BYP68 [o] BYP69 [o] BYP70 [n] BYP71 [n] BYP72A [o] BYP72B [o] BYP72C [o] BYP72E [o] BYP72F [o] BYP72G [o] BYP72H [o] BYP73 [n] BYP74A [o] BYP74B [o]

BYP74C [o] BYP74D [o] BYP74E [o] BYP74F [o] BYP74G [o] BYP74H [o] BYP74I [o] BYP74J [o] BYP74K [o] BYP75 [o] BYP76 [n] BYP77 [n] BYP78 [n] BYP79 [o] BYP80 [o] BYP81 [o] BYP82A [n] BYP82B [n] BYP82C [n] BYP82AA [o] BYP82BA [n] BYP82BB [n] BYP82BD [n] BYP82BE [n] BYP82BF [n] BYP82BH [n] BYP82BI [n] BYP82BJ [n] BYP82BK [n] BYP82BL [n] BYP83 [n] BYP84 [n] BYP84AA [n] BYP84AB [n] BYP84AC [n] BYP84AD [n] BYP84AE [n] BYP84AF [n] BYP84AG [n] BYP84B [o] BYP84C [o] BYP84D [n] BYP85A [n] BYP85B [n] BYP85C [n] BYP85D [n] BYP85E [n] BYP85F [n] BYP85G [n] BYP85H [n] BYP85I [n] BYP85J [n] BYT1_2.ENGLISH [n] BYT1_3.ENGLISH [n] BYT1_4.ENGLISH [n] BYT1_5.ENGLISH [n] BYT1_6.ENGLISH [n] BYT1_7.ENGLISH [n] BYT1_8.ENGLISH [n] BYT2_2.ENGLISH [n] BYT2_3.ENGLISH [s] BYT2_7H.ENGLISH [s] BYT2_7M.ENGLISH [s] BYT2_8A.ENGLISH [o] BYT2_8B.ENGLISH [o] BYT2_8C.ENGLISH [o] BYT2_9A.ENGLISH [o] BYT2_9B.ENGLISH [o] BYT2_9C.ENGLISH [o] BYT2_9D.ENGLISH [o] BYT2_11.ENGLISH [o] BYT2_12A.ENGLISH [o] BYT2_12B.ENGLISH [o] BYT2_12C.ENGLISH [o] BYT2_12D.ENGLISH [o] BYT2_12E.ENGLISH [o] BYT2_12F.ENGLISH [o] BYT2_13A.ENGLISH [n] BYT2_13B.ENGLISH [n] BYT2_13C.ENGLISH [n] BYT2_13D.ENGLISH [n] BYT2_13E.ENGLISH [n] BYT2_13F.ENGLISH [n] BYT2_13G.ENGLISH [n] BYT2_14.ENGLISH [o] BYT2_15.ENGLISH [s] BYT2_16A.ENGLISH [o] BYT2_16B.ENGLISH [o] BYT2_16C.ENGLISH [o] BYT2_16D.ENGLISH [o] BYT2_16E.ENGLISH [o] BYT2_16F.ENGLISH [o] BYT2_16G.ENGLISH [o] BYT3_1.ENGLISH [n] BYT3_2.ENGLISH [n] BYT3_3Y.ENGLISH [o] BYT3_4.ENGLISH [o] BYT3_5.ENGLISH [o] BYT3_7A.ENGLISH [n] BYT3_7B.ENGLISH [n] BYT3_7C.ENGLISH [n] BYT3_7D.ENGLISH [n] BYT3_8.ENGLISH [n] BYT3_9A1.ENGLISH [n] BYT3_9A2.ENGLISH [n] BYT3_9B1.ENGLISH [n] BYT3_9B2.ENGLISH [n] BYT3_9C1.ENGLISH [n] BYT3_9C2.ENGLISH [n] BYT3_9F2.ENGLISH [n] BYT3_9G1.ENGLISH [n] BYT3_9G2.ENGLISH [n] BYT3_10A.ENGLISH [n] BYT310A1.ENGLISH [n] BYT310A2.ENGLISH [n] BYT310B1.ENGLISH [n] BYT310B2.ENGLISH [n] BYT310G1.ENGLISH [n] BYT3_11B.ENGLISH [o] BYT3_12C.ENGLISH [o] BYT3_13.ENGLISH [n] BYT3_14A.ENGLISH [n] BYT3_14B.ENGLISH [n] BYT3_14D.ENGLISH [n] BYT3_14G.ENGLISH [n] BYT3_14M.ENGLISH [n] BYT3_15.ENGLISH [n] BYT3_16A.ENGLISH [o] BYT3_16B.ENGLISH [o] BYT3_16C.ENGLISH [o] BYT3_16D.ENGLISH [o] BYT3_17A.ENGLISH [n] BYT3_17B.ENGLISH [n] BYT3_17C.ENGLISH [n] BYT3_17D.ENGLISH [n] BYT3_17E.ENGLISH [n] BYT3_19.ENGLISH [o] BYT3_20A.ENGLISH [n] BYT3_20B.ENGLISH [n] BYT3_20C.ENGLISH [n] BYT3_20D.ENGLISH [n] BYT3_20E.ENGLISH [n] BYT3_21.ENGLISH [n] BYT3_22.ENGLISH [s] BYT3_23A.ENGLISH [n] BYT3_23B.ENGLISH [n] BYT3_23C.ENGLISH [n] BYT3_23D.ENGLISH [n] BYT3_24.ENGLISH [n] BYT3_25A.ENGLISH [o] BYT3_25B.ENGLISH [o] BYT3_25C.ENGLISH [o] BYT3_25E.ENGLISH [o] BYT3_25F.ENGLISH [o] BYT3_26A.ENGLISH [o]

BYT3_26B.ENGLISH [o] BYT3_26C.ENGLISH [o] BYT3_26D.ENGLISH [o]
BYT3_26E.ENGLISH [o] BYT3_26F.ENGLISH [o] BYT3_26G.ENGLISH [o]
BYT3_26H.ENGLISH [o] BYT3_26I.ENGLISH [o] BYT3_26J.ENGLISH [o]
BYT3_26K.ENGLISH [o] BYT3_27.ENGLISH [n] BYT3_28.ENGLISH [o]
BYT3_29.ENGLISH [o] BYT3_30A.ENGLISH [o] BYT3_30B.ENGLISH [o]
BYT3_30C.ENGLISH [o] BYT3_30D.ENGLISH [o] BYT3_30E.ENGLISH [o]
BYT3_30F.ENGLISH [o] BYT3_30G.ENGLISH [o] BYT3_30H.ENGLISH [o]
BYT3_31.ENGLISH [o] BYT3_32.ENGLISH [o] BYT3_33.ENGLISH [o]
BYT1_2.SCIENCE [n] BYT1_3.SCIENCE [n] BYT1_4.SCIENCE [n]
BYT1_5.SCIENCE [n] BYT1_6.SCIENCE [n] BYT1_7.SCIENCE [n]
BYT1_8.SCIENCE [n] BYT2_2.SCIENCE [n] BYT2_3.SCIENCE [s]
BYT2_7H.SCIENCE [s] BYT2_7M.SCIENCE [s] BYT2_8A.SCIENCE [o]
BYT2_8B.SCIENCE [o] BYT2_8C.SCIENCE [o] BYT2_9A.SCIENCE [o]
BYT2_9B.SCIENCE [o] BYT2_9C.SCIENCE [o] BYT2_9D.SCIENCE [o]
BYT2_11.SCIENCE [o] BYT2_12A.SCIENCE [o] BYT2_12B.SCIENCE [o]
BYT2_12C.SCIENCE [o] BYT2_12D.SCIENCE [o] BYT2_12E.SCIENCE [o]
BYT2_12F.SCIENCE [o] BYT2_13A.SCIENCE [n] BYT2_13B.SCIENCE [n]
BYT2_13C.SCIENCE [n] BYT2_13D.SCIENCE [n] BYT2_13E.SCIENCE [n]
BYT2_13F.SCIENCE [n] BYT2_13G.SCIENCE [n] BYT2_14.SCIENCE [o]
BYT2_15.SCIENCE [s] BYT2_16A.SCIENCE [o] BYT2_16B.SCIENCE [o]
BYT2_16C.SCIENCE [o] BYT2_16D.SCIENCE [o] BYT2_16E.SCIENCE [o]
BYT2_16F.SCIENCE [o] BYT2_16G.SCIENCE [o] BYT3_1.SCIENCE [n]
BYT3_2.SCIENCE [n] BYT3_3Y.SCIENCE [o] BYT3_4.SCIENCE [o]
BYT3_5.SCIENCE [o] BYT3_7A.SCIENCE [n] BYT3_7B.SCIENCE [n]
BYT3_7C.SCIENCE [n] BYT3_7D.SCIENCE [n] BYT3_8.SCIENCE [n]
BYT3_9A1.SCIENCE [n] BYT3_9A2.SCIENCE [n] BYT3_9B2.SCIENCE [n]
BYT3_9C1.SCIENCE [n] BYT3_9C2.SCIENCE [n] BYT3_9D2.SCIENCE [n]
BYT3_9E1.SCIENCE [n] BYT3_9E2.SCIENCE [n] BYT3_9G1.SCIENCE [n]
BYT3_10A.SCIENCE [n] BYT310A1.SCIENCE [n] BYT310A2.SCIENCE [n]
BYT310C2.SCIENCE [n] BYT310E1.SCIENCE [n] BYT310E2.SCIENCE [n]
BYT310G1.SCIENCE [n] BYT3_11B.SCIENCE [o] BYT3_12C.SCIENCE [o]
BYT3_13.SCIENCE [n] BYT3_19.SCIENCE [o] BYT3_20A.SCIENCE [n]
BYT3_20B.SCIENCE [n] BYT3_20C.SCIENCE [n] BYT3_20D.SCIENCE [n]
BYT3_20E.SCIENCE [n] BYT3_21.SCIENCE [n] BYT3_26A.SCIENCE [o]
BYT3_26B.SCIENCE [o] BYT3_26C.SCIENCE [o] BYT3_26D.SCIENCE [o]
BYT3_26E.SCIENCE [o] BYT3_26F.SCIENCE [o] BYT3_26G.SCIENCE [o]
BYT3_26H.SCIENCE [o] BYT3_26I.SCIENCE [o] BYT3_26J.SCIENCE [o]
BYT3_26K.SCIENCE [o] BYT3_27.SCIENCE [n] BYT3_28.SCIENCE [o]
BYT3_29.SCIENCE [o] BYT3_30A.SCIENCE [o] BYT3_30B.SCIENCE [o]
BYT3_30C.SCIENCE [o] BYT3_30D.SCIENCE [o] BYT3_30E.SCIENCE [o]
BYT3_30F.SCIENCE [o] BYT3_30G.SCIENCE [o] BYT3_30H.SCIENCE [o]
BYT3_31.SCIENCE [o] BYT3_32.SCIENCE [o] BYT3_33.SCIENCE [o]
BYT1_2.SOC.STUDIES.HISTORY [n] BYT1_3.SOC.STUDIES.HISTORY [n]

BYT1_4.SOC.STUDIES.HISTORY [n] BYT1_5.SOC.STUDIES.HISTORY [n]
BYT1_6.SOC.STUDIES.HISTORY [n] BYT1_7.SOC.STUDIES.HISTORY [n]
BYT1_8.SOC.STUDIES.HISTORY [n] BYT2_2.SOC.STUDIES.HISTORY [n]
BYT2_3.SOC.STUDIES.HISTORY [s] BYT2_6.SOC.STUDIES.HISTORY [s]
BYT2_7H.SOC.STUDIES.HISTORY [s] BYT2_7M.SOC.STUDIES.HISTORY [s]
BYT2_8A.SOC.STUDIES.HISTORY [o] BYT2_8B.SOC.STUDIES.HISTORY [o]
BYT2_8C.SOC.STUDIES.HISTORY [o] BYT2_9A.SOC.STUDIES.HISTORY
[o] BYT2_9B.SOC.STUDIES.HISTORY [o] BYT2_9C.SOC.STUDIES.HISTORY
[o] BYT2_9D.SOC.STUDIES.HISTORY [o]
BYT2_11.SOC.STUDIES.HISTORY [o] BYT2_12A.SOC.STUDIES.HISTORY [o]
BYT2_12B.SOC.STUDIES.HISTORY [o] BYT2_12C.SOC.STUDIES.HISTORY
[o] BYT2_12D.SOC.STUDIES.HISTORY [o]
BYT2_12E.SOC.STUDIES.HISTORY [o] BYT2_12F.SOC.STUDIES.HISTORY
[o] BYT2_13A.SOC.STUDIES.HISTORY [n]
BYT2_13B.SOC.STUDIES.HISTORY [n] BYT2_13C.SOC.STUDIES.HISTORY
[n] BYT2_13D.SOC.STUDIES.HISTORY [n]
BYT2_13E.SOC.STUDIES.HISTORY [n] BYT2_13F.SOC.STUDIES.HISTORY
[n] BYT2_13G.SOC.STUDIES.HISTORY [n]
BYT2_14.SOC.STUDIES.HISTORY [o] BYT2_15.SOC.STUDIES.HISTORY [s]
BYT2_16A.SOC.STUDIES.HISTORY [o] BYT2_16B.SOC.STUDIES.HISTORY
[o] BYT2_16C.SOC.STUDIES.HISTORY [o]
BYT2_16D.SOC.STUDIES.HISTORY [o] BYT2_16E.SOC.STUDIES.HISTORY
[o] BYT2_16F.SOC.STUDIES.HISTORY [o]
BYT2_16G.SOC.STUDIES.HISTORY [o] BYT3_1.SOC.STUDIES.HISTORY
[n] BYT3_2.SOC.STUDIES.HISTORY [n] BYT3_3Y.SOC.STUDIES.HISTORY
[o] BYT3_4.SOC.STUDIES.HISTORY [o] BYT3_5.SOC.STUDIES.HISTORY
[o] BYT3_6.SOC.STUDIES.HISTORY [o] BYT3_7B.SOC.STUDIES.HISTORY
[n] BYT3_7C.SOC.STUDIES.HISTORY [n] BYT3_7D.SOC.STUDIES.HISTORY
[n] BYT3_8.SOC.STUDIES.HISTORY [n]
BYT3_9A1.SOC.STUDIES.HISTORY [n] BYT3_9A2.SOC.STUDIES.HISTORY
[n] BYT3_9B1.SOC.STUDIES.HISTORY [n]
BYT3_9B2.SOC.STUDIES.HISTORY [n] BYT3_9C1.SOC.STUDIES.HISTORY
[n] BYT3_9C2.SOC.STUDIES.HISTORY [n]
BYT3_9G1.SOC.STUDIES.HISTORY [n] BYT3_9G2.SOC.STUDIES.HISTORY
[n] BYT3_10A.SOC.STUDIES.HISTORY [n]
BYT310A1.SOC.STUDIES.HISTORY [n] BYT310A2.SOC.STUDIES.HISTORY
[n] BYT310B1.SOC.STUDIES.HISTORY [n]
BYT310C1.SOC.STUDIES.HISTORY [n] BYT310C2.SOC.STUDIES.HISTORY
[n] BYT310G1.SOC.STUDIES.HISTORY [n]
BYT310G2.SOC.STUDIES.HISTORY [n] BYT3_11B.SOC.STUDIES.HISTORY
[o] BYT3_12C.SOC.STUDIES.HISTORY [o]
BYT3_13.SOC.STUDIES.HISTORY [n] BYT3_14A.SOC.STUDIES.HISTORY [n]
BYT3_14B.SOC.STUDIES.HISTORY [n] BYT3_14C.SOC.STUDIES.HISTORY

[n] BYT3_14D.SOC.STUDIES.HISTORY [n]
BYT3_14G.SOC.STUDIES.HISTORY [n] BYT3_14M.SOC.STUDIES.HISTORY
[n] BYT3_15.SOC.STUDIES.HISTORY [n]
BYT3_16A.SOC.STUDIES.HISTORY [o] BYT3_16B.SOC.STUDIES.HISTORY
[o] BYT3_16C.SOC.STUDIES.HISTORY [o]
BYT3_16D.SOC.STUDIES.HISTORY [o] BYT3_17A.SOC.STUDIES.HISTORY
[n] BYT3_17B.SOC.STUDIES.HISTORY [n]
BYT3_17C.SOC.STUDIES.HISTORY [n] BYT3_17D.SOC.STUDIES.HISTORY
[n] BYT3_17E.SOC.STUDIES.HISTORY [n]
BYT3_19.SOC.STUDIES.HISTORY [o] BYT3_20A.SOC.STUDIES.HISTORY [n]
BYT3_20B.SOC.STUDIES.HISTORY [n] BYT3_20C.SOC.STUDIES.HISTORY
[n] BYT3_20D.SOC.STUDIES.HISTORY [n]
BYT3_20E.SOC.STUDIES.HISTORY [n] BYT3_21.SOC.STUDIES.HISTORY
[n] BYT3_22.SOC.STUDIES.HISTORY [s] BYT3_23A.SOC.STUDIES.HISTORY
[n] BYT3_23B.SOC.STUDIES.HISTORY [n]
BYT3_23C.SOC.STUDIES.HISTORY [n] BYT3_23D.SOC.STUDIES.HISTORY
[n] BYT3_23E.SOC.STUDIES.HISTORY [n]
BYT3_25A.SOC.STUDIES.HISTORY [o] BYT3_25B.SOC.STUDIES.HISTORY
[o] BYT3_25C.SOC.STUDIES.HISTORY [o]
BYT3_25D.SOC.STUDIES.HISTORY [o] BYT3_25E.SOC.STUDIES.HISTORY
[o] BYT3_25F.SOC.STUDIES.HISTORY [o]
BYT3_26A.SOC.STUDIES.HISTORY [o] BYT3_26B.SOC.STUDIES.HISTORY
[o] BYT3_26C.SOC.STUDIES.HISTORY [o]
BYT3_26D.SOC.STUDIES.HISTORY [o] BYT3_26E.SOC.STUDIES.HISTORY
[o] BYT3_26F.SOC.STUDIES.HISTORY [o]
BYT3_26G.SOC.STUDIES.HISTORY [o] BYT3_26H.SOC.STUDIES.HISTORY
[o] BYT3_26I.SOC.STUDIES.HISTORY [o]
BYT3_26J.SOC.STUDIES.HISTORY [o] BYT3_26K.SOC.STUDIES.HISTORY [o]
BYT3_27.SOC.STUDIES.HISTORY [n] BYT3_28.SOC.STUDIES.HISTORY [o]
BYT3_29.SOC.STUDIES.HISTORY [o] BYT3_30A.SOC.STUDIES.HISTORY
[o] BYT3_30B.SOC.STUDIES.HISTORY [o]
BYT3_30C.SOC.STUDIES.HISTORY [o] BYT3_30D.SOC.STUDIES.HISTORY
[o] BYT3_30E.SOC.STUDIES.HISTORY [o]
BYT3_30F.SOC.STUDIES.HISTORY [o] BYT3_30G.SOC.STUDIES.HISTORY
[o] BYT3_30H.SOC.STUDIES.HISTORY [o] BYT3_31.SOC.STUDIES.HISTORY
[o] BYT3_32.SOC.STUDIES.HISTORY [o] BYT3_33.SOC.STUDIES.HISTORY
[o] BYT1_2.MATH [n] BYT1_3.MATH [n] BYT1_4.MATH [n] BYT1_5.MATH
[n] BYT1_6.MATH [n] BYT1_7.MATH [n] BYT1_8.MATH [n] BYT2_2.MATH
[n] BYT2_3.MATH [s] BYT2_7H.MATH [s] BYT2_7M.MATH [s]
BYT2_8A.MATH [o] BYT2_8B.MATH [o] BYT2_8C.MATH [o]
BYT2_9A.MATH [o] BYT2_9B.MATH [o] BYT2_9C.MATH [o]
BYT2_9D.MATH [o] BYT2_11.MATH [o] BYT2_12A.MATH [o]
BYT2_12B.MATH [o] BYT2_12C.MATH [o] BYT2_12D.MATH [o]

BYT2_12E.MATH [o] BYT2_12F.MATH [o] BYT2_13A.MATH [n] BYT2_13B.MATH [n] BYT2_13C.MATH [n] BYT2_13D.MATH [n] BYT2_13E.MATH [n] BYT2_13F.MATH [n] BYT2_13G.MATH [n] BYT2_14.MATH [o] BYT2_15.MATH [s] BYT2_16A.MATH [o] BYT2_16B.MATH [o] BYT2_16C.MATH [o] BYT2_16D.MATH [o] BYT2_16E.MATH [o] BYT2_16F.MATH [o] BYT2_16G.MATH [o] BYT3_1.MATH [n] BYT3_2.MATH [n] BYT3_3Y.MATH [o] BYT3_4.MATH [o] BYT3_5.MATH [o] BYT3_7A.MATH [n] BYT3_7B.MATH [n] BYT3_7C.MATH [n] BYT3_7D.MATH [n] BYT3_8.MATH [n] BYT3_9A1.MATH [n] BYT3_9A2.MATH [n] BYT3_9B2.MATH [n] BYT3_9C1.MATH [n] BYT3_9C2.MATH [n] BYT3_9D1.MATH [n] BYT3_9D2.MATH [n] BYT3_9E1.MATH [n] BYT3_9E2.MATH [n] BYT3_9G1.MATH [n] BYT3_9G2.MATH [n] BYT3_10A.MATH [n] BYT310A1.MATH [n] BYT310A2.MATH [n] BYT310C2.MATH [n] BYT310D1.MATH [n] BYT310D2.MATH [n] BYT310G1.MATH [n] BYT310G2.MATH [n] BYT3_11B.MATH [o] BYT3_12C.MATH [o] BYT3_13.MATH [n] BYT3_19.MATH [o] BYT3_20A.MATH [n] BYT3_20B.MATH [n] BYT3_20C.MATH [n] BYT3_20D.MATH [n] BYT3_20E.MATH [n] BYT3_21.MATH [n] BYT3_22.MATH [s] BYT3_23A.MATH [n] BYT3_23B.MATH [n] BYT3_23C.MATH [n] BYT3_23E.MATH [n] BYT3_25A.MATH [o] BYT3_25B.MATH [o] BYT3_25C.MATH [o] BYT3_25D.MATH [o] BYT3_25E.MATH [o] BYT3_25F.MATH [o] BYT3_26A.MATH [o] BYT3_26B.MATH [o] BYT3_26C.MATH [o] BYT3_26D.MATH [o] BYT3_26E.MATH [o] BYT3_26F.MATH [o] BYT3_26G.MATH [o] BYT3_26H.MATH [o] BYT3_26I.MATH [o] BYT3_26J.MATH [o] BYT3_26K.MATH [o] BYT3_27.MATH [n] BYT3_28.MATH [o] BYT3_29.MATH [o] BYT3_30A.MATH [o] BYT3_30B.MATH [o] BYT3_30C.MATH [o] BYT3_30D.MATH [o] BYT3_30E.MATH [o] BYT3_30F.MATH [o] BYT3_30G.MATH [o] BYT3_30H.MATH [o] BYT3_31.MATH [o] BYT3_32.MATH [o] BYT3_33.MATH [o] BYT2_3.ALL [s] BYT2_6.ALL [s] BYT2_7H.ALL [s] BYT2_7M.ALL [s] BYT2_15.ALL [s] BYT3_22.ALL [s] BYT2_2.ALL [n] BYT3_2.ALL [n] BYT3_8.ALL [n] BYT3_15.ALL [n] BYT3_27.ALL [n] BYT1_2.ALL [n] BYT1_3.ALL [n] BYT1_6.ALL [n] BYT1_8.ALL [n] BYT2_13A.ALL [n] BYT2_13B.ALL [n] BYT2_13C.ALL [n] BYT2_13D.ALL [n] BYT2_13E.ALL [n] BYT2_13F.ALL [n] BYT2_21.ALL [n] BYT2_27A.ALL [n] BYT2_27B.ALL [n] BYT3_1.ALL [n] BYT3_7A.ALL [n] BYT3_7B.ALL [n] BYT3_7C.ALL [n] BYT3_7D.ALL [n] BYT3_9A1.ALL [n] BYT3_9C2.ALL [n] BYT3_10A.ALL [n] BYT310A1.ALL [n] BYT310A2.ALL [n] BYT310C1.ALL [n] BYT310C2.ALL [n] BYT310E1.ALL [n] BYT310G1.ALL [n] BYT3_14A.ALL [n] BYT3_14B.ALL [n] BYT3_14D.ALL [n] BYT3_14G.ALL [n] BYT3_14M.ALL [n] BYT3_17A.ALL [n] BYT3_17B.ALL [n] BYT3_17C.ALL [n] BYT3_17D.ALL [n] BYT3_17E.ALL [n] BYT3_20A.ALL [n] BYT3_20B.ALL [n]

BYT3_20E.ALL [n] BYT3_21.ALL [n] BYT3_23A.ALL [n] BYT3_23B.ALL [n] BYT3_23C.ALL [n] BYT3_23D.ALL [n] BYT3_23E.ALL [n] BYT2_8A.ALL [o] BYT2_8B.ALL [o] BYT2_8C.ALL [o] BYT2_9B.ALL [o] BYT2_9C.ALL [o] BYT2_9D.ALL [o] BYT2_11.ALL [o] BYT2_12A.ALL [o] BYT2_12B.ALL [o] BYT2_12C.ALL [o] BYT2_12D.ALL [o] BYT2_12E.ALL [o] BYT2_12F.ALL [o] BYT2_14.ALL [o] BYT2_16A.ALL [o] BYT2_16B.ALL [o] BYT2_16C.ALL [o] BYT2_16D.ALL [o] BYT2_16E.ALL [o] BYT2_16F.ALL [o] BYT2_16G.ALL [o] BYT2_17A.ALL [o] BYT2_17B.ALL [o] BYT2_17C.ALL [o] BYT2_17D.ALL [o] BYT2_17E.ALL [o] BYT2_17F.ALL [o] BYT2_18A.ALL [o] BYT2_18B.ALL [o] BYT2_18C.ALL [o] BYT2_18D.ALL [o] BYT2_18E.ALL [o] BYT2_18F.ALL [o] BYT2_18G.ALL [o] BYT2_19.ALL [o] BYT2_20A.ALL [o] BYT2_20B.ALL [o] BYT2_20C.ALL [o] BYT2_20D.ALL [o] BYT2_20E.ALL [o] BYT2_20F.ALL [o] BYT2_20G.ALL [o] BYT2_20H.ALL [o] BYT2_20I.ALL [o] BYT2_20J.ALL [o] BYT2_22.ALL [o] BYT2_23A.ALL [o] BYT2_23B.ALL [o] BYT2_23C.ALL [o] BYT2_23D.ALL [o] BYT2_23E.ALL [o] BYT2_23F.ALL [o] BYT2_23G.ALL [o] BYT2_23H.ALL [o] BYT2_24A.ALL [o] BYT2_24B.ALL [o] BYT2_24C.ALL [o] BYT2_24D.ALL [o] BYT2_24E.ALL [o] BYT2_24F.ALL [o] BYT2_24G.ALL [o] BYT2_24H.ALL [o] BYT2_24I.ALL [o] BYT2_24J.ALL [o] BYT2_24K.ALL [o] BYT2_24L.ALL [o] BYT2_24M.ALL [o] BYT2_24N.ALL [o] BYT2_24O.ALL [o] BYT2_24P.ALL [o] BYT2_24Q.ALL [o] BYT2_25.ALL [o] BYT2_26.ALL [o] BYT2_28.ALL [o] BYT2_29.ALL [o] BYT3_3Y.ALL [o] BYT3_4.ALL [o] BYT3_5.ALL [o] BYT3_11B.ALL [o] BYT3_12C.ALL [o] BYT3_16A.ALL [o] BYT3_16B.ALL [o] BYT3_16C.ALL [o] BYT3_16D.ALL [o] BYT3_19.ALL [o] BYT3_25A.ALL [o] BYT3_25B.ALL [o] BYT3_25C.ALL [o] BYT3_25D.ALL [o] BYT3_25E.ALL [o] BYT3_25F.ALL [o] BYT3_26A.ALL [o] BYT3_26B.ALL [o] BYT3_26C.ALL [o] BYT3_26D.ALL [o] BYT3_26E.ALL [o] BYT3_26F.ALL [o] BYT3_26G.ALL [o] BYT3_26H.ALL [o] BYT3_26I.ALL [o] BYT3_26J.ALL [o] BYT3_26K.ALL [o] BYT3_28.ALL [o] BYT3_29.ALL [o] BYT3_30A.ALL [o] BYT3_30B.ALL [o] BYT3_30C.ALL [o] BYT3_30D.ALL [o] BYT3_30E.ALL [o] BYT3_30F.ALL [o] BYT3_30G.ALL [o] BYT3_30H.ALL [o] BYT3_31.ALL [o] BYT3_32.ALL [o] BYT3_33.ALL [o] FAMCOMP [n] G8CTRL1 [n] G8CTRL2 [n] G10CTRL1 [n] G10URBAN [n] G10REGON [n] G10ENROL [n] G10COHRT [n] G12COHRT [n] G12CTRL1 [n] G12URBN3 [n] G12REGON [n] TRNURBN3 [n] TRNREGON [n]
/TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO
/DEPCATEGORIES USEVALUES=[1.00 2.00] TARGET=[2.00]
/PRINT MODELSUMMARY CLASSIFICATION RISK
/GAIN CATEGORYTABLE=YES TYPE=[NODE] SORT=DESCENDING CUMULATIVE=NO
/RULES NODES=TERMINAL SYNTAX=GENERIC LABELS=YES OUTFILE='C:\Users\Emi\Desktop\rules.txt'

```
  /METHOD TYPE=QUEST MAXSURROGATES=AUTO PRUNE=SE(1)
  /GROWTHLIMIT MAXDEPTH=AUTO MINPARENTSIZE=5
 MINCHILDSIZE=2
  /VALIDATION TYPE=SPLITSAMPLE(training) OUTPUT=BOTHSAMPLES
  /QUEST ALPHASPLIT=0.05
  /COSTS CUSTOM= 1.00 1.00 [0] 1.00 2.00 [1]  2.00 1.00 [3] 2.00 2.00 [0]
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.
```

19. Need to follow up with each ruleset, to calculate relevant accuracy measures for each rule.

Association Rule Mining

1. Open large dataset created above (nels_thomasTLA_final1a.Rda).
2. Shuffle the data
3. Recode all numeric variables in dataset into ordered factors (split into 4 groups)
4. Get rid of all commas in the dataset
5. Retain only those with lower parental education
6. Stratified random sampling into training and test sets (70-30 split, stratified by outcome)
7. Stratified random sampling of training set into generation and screening sets (50-50 split, stratified by outcome)
8. Split each of the datasets (generation, screening, training, test) by outcome (high achievers vs not)
9. Convert datasets (that are dataframes) into transactional form
10. Examine descriptive statistics
11. Set minimum support at 25% (since many questions are asked in a 4 point-scale)
12. Remove unneeded datasets to save memory
13. Run association rules using the generation set (for high achievers)

```
rules1 <- apriori(gen_in_t,
        parameter=list(support= sup, confidence=1, minlen=2, maxlen=3,
maxtime = 50),
        appearance = list(rhs=c("highach=Yes"), default="lhs"),
        control=list(memopt=TRUE, load=FALSE)
)
print("Summary of generation rules")
summary(rules1)
```

14. Calculate support in screening set, and only retain subset with support >=.25.

```
quality(rules1)<-cbind(quality(rules1),
          support_s=interestMeasure(rules1,
                    measure="support",
                    transactions=scr_in_t,
                    reuse=FALSE
          ))
```

```
rules1<- subset(rules1, support_s >=sup)

print("Summary of rules that have required stats in screening set as well")
summary(rules1)
```

15. Calculate total training set coverage

```
quality(rules1)<-cbind(quality(rules1),
                coverage_tr_in=interestMeasure(rules1,
                                measure="coverage",
                                transactions=train_in_t,
                                reuse=FALSE
                ))

quality(rules1)<-cbind(quality(rules1),
                coverage_tr_out=interestMeasure(rules1,
                                 measure="coverage",
                                 transactions=train_out_t,
                                 reuse=FALSE
                ))
```

16. Calculate total test set coverage

```
quality(rules1)<-cbind(quality(rules1),
                coverage_ts_in=interestMeasure(rules1,
                                measure="coverage",
                                transactions=test_in_t,
                                reuse=FALSE
                ))
quality(rules1)<-cbind(quality(rules1),
                coverage_ts_out=interestMeasure(rules1,
                                 measure="coverage",
                                 transactions=test_out_t,
                                 reuse=FALSE
                ))

print("Summary of rules")
summary(rules1)
```

17. Save rules as a dataframe and write as CSV

```
rules1.df <- as(rules1, "data.frame")
saveRDS(rules1.df, file="TL_LoParEd_Hiach_take2.Rda")
write.csv(rules1.df, file = "TL_LoParEd_Hiach_take2.csv")
```

18. Need to conduct additional analyses using the output to identify which rules are most relevant.

Run rCBA

Note: This only worked when the dataset was small.

1. Open small dataset. Make sure missing values are correctly designated and everything us coded as factors. (Need to recode all numeric variables into factors.
2. Get rid of all commas in the dataset
3. Shuffle the dataset
4. Stratified random sampling into training and test sets (70-30 split, stratified by outcome)
5. Conduct Apriori (see steps above)
6. Convert rules to data frame
   training.df <- as(rules1b, "data.frame")
7. Conduct CBA
   options(java.parameters = "-Xmx40g")
   library(rCBA)
   prunedRulesFrame <- pruning(training.data, training.df, method="m2cba")
   print(nrow(prunedRulesFrame))
   #Save output
   out2 <- capture.output(print(nrow(prunedRulesFrame)))
   cat("n_pruned_rules", out2, file="TSA_v3_prunedRulesFrame.txt", sep="\n",
   append=FALSE)
8. Evaluate how good the ruleset is on the test set
   ##Note, with the following program, no commas allowed in the dataset (or rules)
   ##Name the CBA output data frame as "prunedRulesFrame" (as I've done above)
   ##Name the to-be-evaluated-dataframe "eval.df" (as I've done above)
   ##Results in dataset called "Output" with whether each of the rules applied and what the final prediction was

```
prunedRulesFrame$ant<-gsub("} => .*|[{]", '', prunedRulesFrame$rules)
prunedRulesFrame$consq<-gsub(".*=> |[{}]", '', prunedRulesFrame$rules)
varname1 <- paste("Rule", row.names(prunedRulesFrame), sep = "_")
results <- data.frame(setNames(replicate(length(varname1), numeric(0), simplify = F),
varname1))

for (j in 1:((NROW(prunedRulesFrame))-1)) {
x0<-prunedRulesFrame[j, "consq"]
x1<-strsplit(prunedRulesFrame[j, "ant"], ",")[[1]]
x2<-unname(mapply(sub, "=", "zzz", x1))
value<-unname(mapply(sub, "zzz.*", "", x2))
attribute<-unname(mapply(sub, ".*zzz", "", x2))
df1<-cbind(value, attribute)
eval.df$pred<-NROW(df1)
myList<-list()
for (i in 1:NROW(df1)) myList[[paste('A', i)]] <-
  ifelse(eval.df[,print(match(df1[i,1], colnames(eval.df)))]==df1[i,2], 1, 0)
df2 <- data.frame(matrix(unlist(myList), nrow=NROW(eval.df), byrow=FALSE))
df3 <- merge(df2, eval.df, by=0)
```

```
if (NROW(df1)>1) df3$sumX <-rowSums(df3[,2:(NROW(df1)+1)])
   else (df3$sumX <- df3[,2])
results[1:NROW(df3),j] <- ifelse(df3$sumX==NROW(df1), x0, 0)
}
results[is.na(results)] <-0 #convert missing into 0 for the results
#find CBA prediction
results$prediction<-0
for (j in 1:NROW(results)) {
  for (i in 1:(NROW(prunedRulesFrame)-1)) {
    if (results[j, i] != 0) {results[j,"prediction"]<-results[j,i]; break}
  }
}
#Create final dataset
output <- merge(df3, results, by=0)
########################
table(output$highach, output$prediction, exclude=NULL)
write.csv(output, file = "TSA_testset_rCBA_pred.csv")
write.csv(prunedRulesFrame, file = "TSA_rCBA_rules.csv")
######################
```

APPENDIX H

ANALYSIS FLOW & SYNTAX FOR STUDY 2 DATA MINING

Note: See Table 18 for list and references to R packages for each rule induction algorithm. Other R packages used for data preparation and visualization include plyr (Wickham, 2011), caret (Wing et al., 2016), sas7bdat (Shotwell, 2014), dplyr (Wickham & Francois, 2016), magrittr (Bache & Wickham, 2014), methods (R Core Team, 2016), rattle (Williams, 2011) and haven (Wickham & Miller, 2016).

Create dataset that is expanded by weights

1. In SAS, create SAS datafiles of each NELS dataset that saves values as data rather than as labels.
2. In R, save each of the files above as an R datafile, then for each,
    a. Convert missing values to NA
    b. Set numeric variables to numeric
    c. Set ordinal variables to ordered factors
    d. Set categorical variables to factors
3. Dichotomize dependent variable where at/above median (48.55) is high achievement
4. Append other data onto b&m dataset from replication round (outcome = 12^{th} grade math)
5. Select variables I want from F1 & F2, plus STU_ID, F2PNLWT and F22XMIRR. (subset 1, 1058 variables)
6. Make another subset that doesn't have any F1/F2 vars (subset 2, 1994 variables)
7. Merge subset 1 & subset 2 based on STU_ID (3039 variables)
8. Delete duplicated variables and unused levels (nels_bm_final1.Rda , 2751 variables)
9. Divide each weight by min (2.391) or min*10 (23.91) and round to a whole number, making sure that values less than 0 round up to 1.
10. Expand cases according to weight
11. Remove Zero and Near Zero-Variance Predictors (2754-->2303)
12. Categorize dependent variable (at or above median is "high" achievement)
13. Cut down 4 extra variables that somehow snuck in
14. Examined the 36 categorical variables with 10 or more levels and reduced levels for 31 of them.
15. Save dataset (nels_bla_expanded_v1.Rda, or nels_bla_expanded_tenth_v1.Rda)

Create dataset that is not yet expanded by weights for association rule mining

1. Open nels_bm_final1.Rda (intermediate dataset created above)

2. Trim the dataset to only the 1933+ predictors that resulted from the data cleaning process above.
3. Categorize dependent variable (at or above median is "high" achievement)
4. Reduced levels for 31 out of 36 categorical variables that had 10 or more levels.
5. Save dataset (nels_bla_NOTexpanded_v1.Rda)

Ruleset mining

C45, PART, RIPPER, C5.0, boosted C5.0, CART, bagged CART, Random Forest, QUEST

1. Open large dataset created above (nels_bla_expanded_v1.Rda, or nels_bla_expanded_tenth_v1.Rda for C45, PART, RIPPER, C5.0, CART, bagged CART, Random Forest).
2. [For C45, PART, RIPPER, C5.0] Delete STU_ID and F22XMIRR, set mathach to factor.
3. [Random Forest] Delete STU_ID and F22XMIRR, set mathach to factor.
4. [For CART, bagged CART] Calculate mean and median of 12th grade math score (F22XMIRR); Delete STU_ID and mathach.
5. [For bagged CART] Retain variables with fewer than 20% NAs (1934 variables --> 849 variables) (note: 40% cut-off, that got number of variables down to 1384, didn't work)
6. Shuffle the dataset
7. [For C4.5, PART, RIPPER, Random Forest] Substitute missing values (mean for numeric, and "missing" category for factors)
8. [For C4.5, PART, RIPPER, Random Forest] Convert every factor to character, all NA to "missing", then characters back to factor.
9. Stratified random sampling into training and test sets (70-30 split, stratified by outcome) [For CART, bagged CART] Calculate minimum cases I want in each terminal node (2% of sample)

10. [For bagged CART] Retain variables with fewer than 50% NAs (1373 → 706 variables)
11. Shuffle the data
12. [For C5.0, rCBA] Recode all numeric variables in dataset into ordered factors (split into 4 groups)
13. [For rCBA] Get rid of all commas in the dataset
14. [For randomForest] Substitute missing values (mean for numeric, and "missing" category for factors)

15. Grow final C4.5 tree (created without boosting or cost-adjustment)

```
options( java.parameters = "-Xmx6g")
library(RWeka)
# C45 tree
options( java.parameters = "-Xmx6g")
library(RWeka)
model1<- J48(mathach ~ ., data = training.data3,
        control = Weka_control(S=FALSE, M=1232, R=TRUE, N=3, Q=1234))
```

```
model1
summary(model1)
# predict on holdout data
model1p <- predict(model1, test.data3)
table(test.data3$mathach, predicted = model1p)
plot(model1p)
```

16. Grow PART

```
model1<- PART(mathach ~ ., data = training.data3,
          control = Weka_control(S=FALSE, M=1232, R=TRUE, N=3, Q=1234))
model1
summary(model1)
## predict on holdout data
model1p <- predict(model1, test.data3)
table(test.data3$mathach, predicted = model1p)
plot(model1p)
```

17. Grow RIPPER

```
model1<- JRip(mathach ~ ., data = training.data3,
        control = Weka_control(N=1232, S=1234))
model1
summary(model1)
## predict on holdout data
model1p <- predict(model1, test.data3)
table(test.data3$mathach, predicted = model1p)
plot(model1p)
```

18. Grow C5.0 and boosted C5.0

```
#Find out where the outcome is
match("mathach",names(training.data))

##Final non-boosted model (after trying many non-boosted models) (only for final,
change sample from .5 to 0)
model1 <- C5.0(training.data[, c(-23)], training.data[, 23],
           trials=1, rules=FALSE,
           control = C5.0Control(subset = FALSE,
                       bands = 0,
                       winnow = FALSE,
                       noGlobalPruning = FALSE,
                       CF = 0.25,
                       minCases = 1232,
                       sample = 0,
                       fuzzyThreshold = FALSE,
```

```
                        seed = 1234,
                        label = "mathach")
                )
        model1
        summary(model1)

        ## predict on holdout data
        model1p <- predict(model1, test.data)
        model1p
        table(test.data$mathach,predicted = model1p)

        ## Final boosted model  (change "sample" from .5 to 0)
        model2 <- C5.0(training.data[, c(-23)], training.data[, 23],
                trials=20,
                control = C5.0Control(subset = FALSE,
                        bands = 0,
                        winnow = FALSE,
                        noGlobalPruning = FALSE,
                        CF = 0.25,
                        minCases = 1232,
                        sample = 0,
                        fuzzyThreshold = FALSE,
                        seed = 1234,
                        label = "mathach"))
        model2
        summary(model2)

        ## predict on holdout data
        model2p <- predict(model2, test.data)
        model2p
        table(test.data$mathach,predicted = model2p)
```

19. Grow CART

```
        library(rpart)
        Model_rpart1 <- rpart(F22XMIRR ~ . ,
                        method = "anova",
                        data = training.data,
                        xval = 10,
                        control = rpart.control(minbucket=mb, cp=0),
                        parms=list(split = "gini"))

        #print(Model_rpart1) # results
        printcp(Model_rpart1) # display the results
```

```
plotcp(Model_rpart1) # visualize cross-validation results, see where to prune

# summary(Model_rpart1) # detailed summary of splits
pModel_rpart1 <-prune(Model_rpart1, cp=0.00205512) #prune tree based on plotcp
print(pModel_rpart1)
summary(pModel_rpart1)

### Calculate predicted value, Mean-squared error and R^2 in testset
test.data$pred <- predict(pModel_rpart1, test.data, type="vector")
summary(test.data$pred)
#squared error
test.data$err_sq <- ((test.data$F22XMIRR - test.data$pred)^2)
#SSE
SSE <- sum(test.data$err_sq)
SSE
#mean of mathach (from earlier)
m
#SST
test.data$SST <- (test.data$F22XMIRR - m)^2
SST<- sum(test.data$SST)
SST
#R-squared
1-SSE/SST
#MSE
MSE <- SSE/nrow(test.data)
MSE
#RMSE
RMSE <- MSE^(.5)
RMSE
#Confusion matrix
test.data$mathach <- ifelse(test.data$F22XMIRR >= 48.57,
                        c("high"), c("low"))
test.data$predach <- ifelse(test.data$pred >= 48.57,
                            c("high"), c("low"))
table(test.data$predach, test.data$mathach)
```

20. Grow bagged CART

```
library(rpart)
library(ipred)
model1<- bagging(F22XMIRR~., data=training.data)

model1<- bagging(F22XMIRR~.,
                 data=training.data,
```

```
            control=rpart.control(minbucket=mb, cp=0.00205512))
print(model1)
pred <- predict(model1, test.data)


### Calculate predicted value, Mean-squared error and R^2 in testset
test.data$pred <- predict(model1, test.data, type="prob")
summary(test.data$pred)

#squared error (weignted)
test.data$err_sq <- ((test.data$F22XMIRR - test.data$pred)^2)
#SSE
SSE <- sum(test.data$err_sq)
print("SSE")
SSE
#weighted mean of mathach (calculated earlier)
m
#SST
test.data$SST <- ((test.data$F22XMIRR - m)^2)
SST<- sum(test.data$SST)
print("SST")
SST
#R-squared
print("R-squared")
1-SSE/SST
#MSE
MSE <- SSE/(nrow(test.data))
print("MSE")
MSE
#RMSE
RMSE <- MSE^(.5)
print("RMSE")
RMSE
#Adj R-squared
print("adj R-sq")
1-(SSE*(nrow(test.data)-1)/(SST*(nrow(test.data)-1340)))

#Confusion matrix
test.data$mathach <- ifelse(test.data$F22XMIRR >= 48.57,
                    c("high"), c("low"))
test.data$predach <- ifelse(test.data$pred >= 48.57,
                    c("high"), c("low"))
```

```
table(test.data$predach, test.data$mathach)
```

21. Grow Random Forest

```
#find column numbers for variables I want to exclude
match("mathach",names(BM_lg.datar))
match("STU_ID",names(BM_lg.datar))

library(randomForest)
set.seed(1234)
model1 <- randomForest(F22XMIRR ~ .,
                       training.data[,c(-1, -24)],
                       nodesize=1232, #default is 5 for Reg T!
                       ntree=500, #default is 500
                       importance=TRUE)
print(model1)

### Calculate predicted value, Mean-squared error and R^2 in testset
test.data$pred <- predict(model1, test.data)
summary(test.data$pred)
#squared error (weignted)
test.data$err_sq <- ((test.data$F22XMIRR - test.data$pred)^2)
#SSE
SSE <- sum(test.data$err_sq)
SSE
#weighted mean of mathach
m <- mean(test.data$F22XMIRR)
m
#SST
test.data$SST <- ((test.data$F22XMIRR - m)^2)
SST<- sum(test.data$SST)
SST
#R-squared
1-SSE/SST
#MSE
MSE <- SSE/(nrow(test.data))
MSE
#RMSE
RMSE <- MSE^(.5)
RMSE
#Adj R-squared
1-(SSE*(nrow(test.data)-1)/(SST*(nrow(test.data)-15)))
```

```
#CM
test.data$mathach <- ifelse(test.data$F22XMIRR >= 48.57,
                        c("high"), c("low"))
test.data$predach <- ifelse(test.data$pred >= 48.57,
                        c("high"), c("low"))

table(test.data$predach, test.data$mathach)


##Variable importance
#p.6 (https://cran.r-project.org/web/packages/randomForest/randomForest.pdf)
varImpPlot(model1, main="BM_Small_RF \nImportant Variables")
out1 <- importance(model1, type=1)
write.csv(out1, file = "BM_lg_RF_VarImp1.csv")
out2 <- importance(model1, type=2)
write.csv(out2, file = "BM_lg_RF_VarImp2.csv")
```

22. Grow QUEST (in SPSS) (Note: Use training and test sets created for CART. Use haven package to save R dataset as SPSS dataset)

```
* Decision Tree.
TREE mathach [n] BY BYSES [s] Pexp [o] Sexp [o] BYGRADS [s] gm_none [n]
gm_1 [n] gm_2 [n] geo_none [n] geo_1 [n] al2_none [n] al2_1 [n] emph_m [s]
t_rspnsv [s]  BYTXMIRR [s] GPA910_m [s] grad_eff [o] SATplan [n] m_selfcpt [s]
female [n] hispanic [n] black [n] white [n]  BYS4A [n] BYS4OCC [n] BYS7OCC [n]
BYS8A [n] BYS8B [n] BYS8C [n] BYS8E [n] BYS8F [n] BYS8G [n] BYS8H [n]
BYS12 [n] BYS14 [n] BYS15 [n] BYS17 [n] BYS21 [n] BYS31A [n] BYS32 [o]
BYS33 [o] BYS34A [n] BYS34B [n] BYS35A [n] BYS35B [n] BYS35C [n]
BYS35D [n] BYS35E [n] BYS35G [n] BYS35H [n] BYS35I [n] BYS35J [n]
BYS35L [n] BYS35M [n] BYS35N [n] BYS35P [n] BYS36A [o] BYS36B [o]
BYS36C [o] BYS37A [n] BYS37B [n] BYS37C [n] BYS37D [n] BYS38A [o]
BYS38B [o] BYS38C [o] BYS38D [o] BYS39A [n] BYS39B [n] BYS39C [n]
BYS40A [o] BYS40B [o] BYS40C [o] BYS40E [o] BYS40F [o] BYS40G [o]
BYS40H [o] BYS41 [o] BYS42A [o] BYS42B [o] BYS44A [o] BYS44B [o]
BYS44C [o] BYS44D [o] BYS44E [o] BYS44F [o] BYS44G [o] BYS44H [o]
BYS44I [o] BYS44J [o] BYS44K [o] BYS44L [o] BYS44M [o] BYS46 [o] BYS47
[o] BYS48A [n] BYS48B [n] BYS49 [n] BYS50A [o] BYS50B [o] BYS50C [o]
BYS50D [o] BYS50E [o] BYS50F [o] BYS51AA [n] BYS51AB [n] BYS51AC [n]
BYS51BA [n] BYS51BB [n] BYS51BC [n] BYS51CA [n] BYS51CB [n] BYS51CC
[n] BYS51DA [n] BYS51DB [n] BYS51DC [n] BYS51EA [n] BYS51EB [n]
BYS51EC [n] BYS51FA [n] BYS51FB [n] BYS51FC [n] BYS51GA [n] BYS51GB
[n] BYS51GC [n] BYS51HA [n] BYS51HB [n] BYS51HC [n] BYS52 [n] BYS53 [o]
BYS54 [n] BYS55A [o] BYS55B [o] BYS55C [o] BYS55D [o] BYS55E [o]
BYS55F [o] BYS56A [o] BYS56B [o] BYS56C [o] BYS56D [o] BYS56E [o]
```

BYS57A [o] BYS57B [o] BYS57C [o] BYS58A [o] BYS58B [o] BYS58C [o] BYS58D [o] BYS58E [o] BYS58F [o] BYS58G [o] BYS58H [o] BYS58I [o] BYS58J [o] BYS58K [o] BYS59A [o] BYS59B [o] BYS59C [o] BYS59D [o] BYS59E [o] BYS59F [o] BYS59G [o] BYS59H [o] BYS59I [o] BYS59J [o] BYS59K [o] BYS59L [o] BYS59M [o] BYS60A [n] BYS60B [n] BYS60C [n] BYS60D [n] BYS61 [n] BYS62 [n] BYS63 [n] BYS64 [n] BYS65 [n] BYS66A [n] BYS66B [n] BYS66C [n] BYS66D [n] BYS67A [n] BYS67B [n] BYS67C [n] BYS67AA [n] BYS67AB [n] BYS67AC [n] BYS67AD [n] BYS67BA [n] BYS67BB [n] BYS67BC [n] BYS67BD [n] BYS67BE [n] BYS67BF [n] BYS67BG [n] BYS67BH [n] BYS67CA [n] BYS67CB [n] BYS67CC [n] BYS67CD [n] BYS67DA [n] BYS67DB [n] BYS67DC [n] BYS67DD [n] BYS68A [n] BYS69A [o] BYS69B [o] BYS69C [o] BYS70A [o] BYS70B [o] BYS70C [o] BYS71A [o] BYS71B [o] BYS71C [o] BYS72A [o] BYS72B [o] BYS72C [o] BYS73 [o] BYS74 [n] BYS75 [o] BYS76 [o] BYS77 [o] BYS78A [o] BYS78B [o] BYS78C [o] BYS79A [o] BYS79B [o] BYS79C [o] BYS79D [o] BYS79E [o] BYS80 [o] BYS82A [n] BYS82B [n] BYS82C [n] BYS82D [n] BYS82E [n] BYS82F [n] BYS82G [n] BYS82K [n] BYS82L [n] BYS82M [n] BYS82N [n] BYS82O [n] BYS82P [n] BYS82Q [n] BYS82R [n] BYS82S [n] BYS82T [n] BYS83A [n] BYS83B [n] BYS83C [n] BYS83D [n] BYS83E [n] BYS83F [n] BYS83G [n] BYS83H [n] BYS83I [n] BYS83J [n] G8TYPE [n] G8CTRL [n] BYSCENRL [o] G8ENROL [o] G8URBAN [n] G8REGON [n] G8MINOR [o] G8LUNCH [o] NOMSECT [n] SEX [n] RACE [n] HISP [n] HANDPAST [n] BYLOCUS1 [s] BYLOCU1T [o] BYLOCUS2 [s] BYLOCU2T [o] BYCNCPT1 [s] BYCNCP1T [o] BYCNCPT2 [s] BYCNCP2T [o] BYSESQ [o] BYPARED [n] BYFAMSIZ [o] BYFCOMP [n] BYPARMAR [n] BYFAMINC [o] BYHMLANG [n] BYPSEPLN [n] BYHOMEWK [o] BYLM [n] BYSC6 [o] BYSC7 [s] BYSC9H [s] BYSC11 [s] BYSC12 [s] BYSC13A [o] BYSC13B [o] BYSC13C [o] BYSC13D [o] BYSC13E [o] BYSC14 [o] BYSC15 [o] BYSC16B [s] BYSC16C [s] BYSC16E [s] BYSC16F [s] BYSC16G [s] BYSC17 [o] BYSC18 [n] BYSC19 [o] BYSC20A [o] BYSC20B [o] BYSC20C [o] BYSC20D [o] BYSC20E [o] BYSC21 [s] BYSC22 [s] BYSC23 [n] BYSC24A [n] BYSC24B [n] BYSC24C [n] BYSC24E [n] BYSC24F [n] BYSC25 [n] BYSC29 [n] BYSC30 [n] BYSC35 [n] BYSC36A [o] BYSC36B [o] BYSC36C [o] BYSC36D [o] BYSC37 [o] BYSC38A [n] BYSC38B [n] BYSC38C [n] BYSC38D [n] BYSC38F [n] BYSC38G [n] BYSC39C [n] BYSC39D [n] BYSC39E [n] BYSC39F [n] BYSC39G [n] BYSC39H [n] BYSC39I [n] BYSC39J [n] BYSC39K [n] BYSC39L [n] BYSC39M [n] BYSC40 [n] BYSC41A [n] BYSC41B [n] BYSC41C [n] BYSC41D [n] BYSC41E [n] BYSC41F [n] BYSC41G [n] BYSC41H [n] BYSC41I [n] BYSC43 [n] BYSC44A [n] BYSC44B [n] BYSC44C [n] BYSC44D [n] BYSC44E [n] BYSC44F [n] BYSC44G [n] BYSC44H [n] BYSC44I [n] BYSC45A [n] BYSC45B1 [n] BYSC45B2 [n] BYSC45B3 [n] BYSC45B4 [n] BYSC45C2 [n] BYSC45D [n] BYSC46A [n] BYSC46B [n] BYSC46C [n] BYSC46D [n] BYSC46E [n] BYSC46F [n] BYSC46G [n] BYSC46H [n] BYSC46I [n] BYSC46J [n] BYSC46K [n] BYSC46L [n] BYSC46M [n] BYSC46N [n]

BYSC46O [n] BYSC46P [n] BYSC46Q [n] BYSC46R [n] BYSC46S [n] BYSC46T [n] BYSC46U [n] BYSC46V [n] BYSC47A [o] BYSC47B [o] BYSC47C [o] BYSC47D [o] BYSC47E [o] BYSC47F [o] BYSC47G [o] BYSC47H [o] BYSC47I [o] BYSC47J [o] BYSC47K [o] BYSC47L [o] BYSC47M [o] BYSC47N [o] BYSC47O [o] BYSC48A [n] BYSC48B [n] BYSC48C [n] BYSC48D [n] BYSC48E [n] BYSC48F [n] BYSC48G [n] BYSC48H [n] BYSC48I [n] BYSC48J [n] BYSC49A [o] BYSC49B [o] BYSC49C [o] BYSC49D [o] BYSC49E [o] BYSC49F [o] BYSC49G [o] BYSC49H [o] BYSC49I [o] BYSC49J [o] BYSC49K [o] BYSC50AA [o] BYSC50AB [o] BYSC50AC [o] BYSC50AD [o] BYSC50AE [o] BYSC50AF [o] BYSC50AG [o] BYSC50AH [o] BYSC50AI [o] BYSC50AJ [o] BYSC50AK [o] BYSC50AL [o] BYSC50AM [o] BYSC50BA [o] BYSC50BB [o] BYSC50BC [o] BYSC50BD [o] BYSC50BE [o] BYSC50BF [o] BYSC50BG [o] BYSC50BH [o] BYSC50BI [o] BYSC50BJ [o] BYSC50BK [o] BYSC50BL [o] BYSC50BM [o] G8SUBS [n] BYSCORG2 [n] BYRATIO [o] BYP2 [o] BYP3A [o] BYP3B [o] BYP4 [o] BYP5A [o] BYP5B [o] BYP6 [o] BYP7 [n] BYP8 [o] BYP9 [n] BYP10 [n] BYP11 [n] BYP14 [n] BYP22A [n] BYP29 [n] BYP30 [o] BYP31 [o] BYP32 [n] BYP34A [n] BYP34B [n] BYP35 [n] BYP37A [n] BYP37B [n] BYP38A [n] BYP38B [n] BYP38C [n] BYP38D [n] BYP39 [o] BYP40 [o] BYP44 [n] BYP47G [n] BYP48D [n] BYP48G [n] BYP50 [n] BYP51 [n] BYP53 [n] BYP54 [n] BYP55 [n] BYP56 [n] BYP57A [o] BYP57B [o] BYP57C [o] BYP57D [o] BYP57E [o] BYP57F [o] BYP57G [o] BYP57H [o] BYP58A [o] BYP58B [o] BYP58C [o] BYP58D [o] BYP58E [o] BYP58F [o] BYP59A [n] BYP59B [n] BYP59C [n] BYP59D [n] BYP59E [n] BYP60A [n] BYP60B [n] BYP60C [n] BYP60E [n] BYP60G [n] BYP60H [n] BYP61AA [n] BYP61AB [n] BYP61BA [n] BYP61BB [n] BYP61CA [n] BYP61CB [n] BYP61DA [n] BYP61DB [n] BYP61EA [n] BYP61EB [n] BYP62 [n] BYP62A1 [n] BYP62B1 [n] BYP62A2 [n] BYP62B2 [n] BYP62A3 [n] BYP62B3 [n] BYP62A4 [n] BYP62B4 [n] BYP62A5 [n] BYP62B5 [n] BYP63A [n] BYP63B [n] BYP63D [n] BYP63E [n] BYP63F [n] BYP63G [n] BYP63H [n] BYP63I [n] BYP64A [n] BYP64B [n] BYP64C [n] BYP64D [n] BYP65A [n] BYP65B [n] BYP65C [n] BYP66 [o] BYP67 [o] BYP68 [o] BYP69 [o] BYP70 [n] BYP71 [n] BYP72A [o] BYP72B [o] BYP72C [o] BYP72E [o] BYP72F [o] BYP72G [o] BYP72H [o] BYP73 [n] BYP74A [o] BYP74B [o] BYP74C [o] BYP74D [o] BYP74E [o] BYP74F [o] BYP74G [o] BYP74H [o] BYP74I [o] BYP74J [o] BYP74K [o] BYP75 [o] BYP76 [n] BYP77 [n] BYP78 [n] BYP80 [o] BYP81 [o] BYP82A [n] BYP82B [n] BYP82C [n] BYP83 [n] BYP84 [n] BYP84AA [n] BYP84AB [n] BYP84AC [n] BYP84AD [n] BYP84AE [n] BYP84AF [n] BYP84AG [n] BYP84B [o] BYP84C [o] BYP84D [n] BYP85A [n] BYP85B [n] BYP85C [n] BYP85D [n] BYP85E [n] BYP85F [n] BYP85G [n] BYP85H [n] BYP85I [n] BYP85J [n] BYT1_2.ENGLISH [n] BYT1_3.ENGLISH [n] BYT1_4.ENGLISH [n] BYT1_6.ENGLISH [n] BYT1_7.ENGLISH [n] BYT1_8.ENGLISH [n] BYT1_11.ENGLISH [n] BYT2_2.ENGLISH [n] BYT2_3.ENGLISH [s] BYT2_7H.ENGLISH [s] BYT2_7M.ENGLISH [s] BYT2_8A.ENGLISH [o] BYT2_8B.ENGLISH [o] BYT2_8C.ENGLISH [o]

BYT2_9A.ENGLISH [o] BYT2_9B.ENGLISH [o] BYT2_9C.ENGLISH [o]
BYT2_9D.ENGLISH [o] BYT2_11.ENGLISH [o] BYT2_12A.ENGLISH [o]
BYT2_12B.ENGLISH [o] BYT2_12C.ENGLISH [o] BYT2_12D.ENGLISH [o]
BYT2_12E.ENGLISH [o] BYT2_12F.ENGLISH [o] BYT2_13A.ENGLISH [n]
BYT2_13B.ENGLISH [n] BYT2_13C.ENGLISH [n] BYT2_13D.ENGLISH [n]
BYT2_13E.ENGLISH [n] BYT2_13F.ENGLISH [n] BYT2_13G.ENGLISH [n]
BYT2_14.ENGLISH [o] BYT2_15.ENGLISH [s] BYT2_16A.ENGLISH [o]
BYT2_16B.ENGLISH [o] BYT2_16C.ENGLISH [o] BYT2_16D.ENGLISH [o]
BYT2_16E.ENGLISH [o] BYT2_16F.ENGLISH [o] BYT2_16G.ENGLISH [o]
BYT3_1.ENGLISH [n] BYT3_2.ENGLISH [n] BYT3_3Y.ENGLISH [o]
BYT3_4.ENGLISH [o] BYT3_5.ENGLISH [o] BYT3_6.ENGLISH [o]
BYT3_7A.ENGLISH [n] BYT3_7B.ENGLISH [n] BYT3_7C.ENGLISH [n]
BYT3_7D.ENGLISH [n] BYT3_8.ENGLISH [n] BYT3_9A1.ENGLISH [n]
BYT3_9A2.ENGLISH [n] BYT3_9B1.ENGLISH [n] BYT3_9B2.ENGLISH [n]
BYT3_9C1.ENGLISH [n] BYT3_9C2.ENGLISH [n] BYT3_9F2.ENGLISH [n]
BYT3_9G1.ENGLISH [n] BYT3_9G2.ENGLISH [n] BYT3_10A.ENGLISH [n]
BYT3_11B.ENGLISH [o] BYT3_12C.ENGLISH [o] BYT3_13.ENGLISH [n]
BYT3_19.ENGLISH [o] BYT3_20A.ENGLISH [n] BYT3_20B.ENGLISH [n]
BYT3_20C.ENGLISH [n] BYT3_20D.ENGLISH [n] BYT3_20E.ENGLISH [n]
BYT3_20F.ENGLISH [n] BYT3_21.ENGLISH [n] BYT3_26A.ENGLISH [o]
BYT3_26B.ENGLISH [o] BYT3_26C.ENGLISH [o] BYT3_26D.ENGLISH [o]
BYT3_26E.ENGLISH [o] BYT3_26F.ENGLISH [o] BYT3_26G.ENGLISH [o]
BYT3_26H.ENGLISH [o] BYT3_26I.ENGLISH [o] BYT3_26J.ENGLISH [o]
BYT3_26K.ENGLISH [o] BYT3_27.ENGLISH [n] BYT3_28.ENGLISH [o]
BYT3_29.ENGLISH [o] BYT3_30A.ENGLISH [o] BYT3_30B.ENGLISH [o]
BYT3_30C.ENGLISH [o] BYT3_30D.ENGLISH [o] BYT3_30E.ENGLISH [o]
BYT3_30F.ENGLISH [o] BYT3_30G.ENGLISH [o] BYT3_30H.ENGLISH [o]
BYT3_31.ENGLISH [o] BYT3_32.ENGLISH [o] BYT1_2.SCIENCE [n]
BYT1_3.SCIENCE [n] BYT1_4.SCIENCE [n] BYT1_6.SCIENCE [n]
BYT1_7.SCIENCE [n] BYT1_8.SCIENCE [n] BYT2_2.SCIENCE [n]
BYT2_3.SCIENCE [s] BYT2_7H.SCIENCE [s] BYT2_7M.SCIENCE [s]
BYT2_8A.SCIENCE [o] BYT2_8B.SCIENCE [o] BYT2_8C.SCIENCE [o]
BYT2_9A.SCIENCE [o] BYT2_9B.SCIENCE [o] BYT2_9C.SCIENCE [o]
BYT2_9D.SCIENCE [o] BYT2_11.SCIENCE [o] BYT2_12A.SCIENCE [o]
BYT2_12B.SCIENCE [o] BYT2_12C.SCIENCE [o] BYT2_12D.SCIENCE [o]
BYT2_12E.SCIENCE [o] BYT2_12F.SCIENCE [o] BYT2_13A.SCIENCE [n]
BYT2_13B.SCIENCE [n] BYT2_13C.SCIENCE [n] BYT2_13D.SCIENCE [n]
BYT2_13E.SCIENCE [n] BYT2_13F.SCIENCE [n] BYT2_13G.SCIENCE [n]
BYT2_14.SCIENCE [o] BYT2_15.SCIENCE [s] BYT2_16A.SCIENCE [o]
BYT2_16B.SCIENCE [o] BYT2_16C.SCIENCE [o] BYT2_16D.SCIENCE [o]
BYT2_16E.SCIENCE [o] BYT2_16F.SCIENCE [o] BYT2_16G.SCIENCE [o]
BYT3_1.SCIENCE [n] BYT3_2.SCIENCE [n] BYT3_3Y.SCIENCE [o]
BYT3_4.SCIENCE [o] BYT3_5.SCIENCE [o] BYT3_7A.SCIENCE [n]

BYT3_7B.SCIENCE [n] BYT3_7C.SCIENCE [n] BYT3_7D.SCIENCE [n]
BYT3_8.SCIENCE [n] BYT3_9A1.SCIENCE [n] BYT3_9A2.SCIENCE [n]
BYT3_9B2.SCIENCE [n] BYT3_9C1.SCIENCE [n] BYT3_9C2.SCIENCE [n]
BYT3_9D2.SCIENCE [n] BYT3_9E1.SCIENCE [n] BYT3_9E2.SCIENCE [n]
BYT3_9G1.SCIENCE [n] BYT3_9G2.SCIENCE [n] BYT3_10A.SCIENCE [n]
BYT3_11B.SCIENCE [o] BYT3_12C.SCIENCE [o] BYT3_13.SCIENCE [n]
BYT3_19.SCIENCE [o] BYT3_20A.SCIENCE [n] BYT3_20B.SCIENCE [n]
BYT3_20C.SCIENCE [n] BYT3_20D.SCIENCE [n] BYT3_20E.SCIENCE [n]
BYT3_21.SCIENCE [n] BYT3_26A.SCIENCE [o] BYT3_26B.SCIENCE [o]
BYT3_26C.SCIENCE [o] BYT3_26D.SCIENCE [o] BYT3_26E.SCIENCE [o]
BYT3_26F.SCIENCE [o] BYT3_26G.SCIENCE [o] BYT3_26H.SCIENCE [o]
BYT3_26I.SCIENCE [o] BYT3_26J.SCIENCE [o] BYT3_26K.SCIENCE [o]
BYT3_27.SCIENCE [n] BYT3_28.SCIENCE [o] BYT3_29.SCIENCE [o]
BYT3_30A.SCIENCE [o] BYT3_30B.SCIENCE [o] BYT3_30C.SCIENCE [o]
BYT3_30D.SCIENCE [o] BYT3_30E.SCIENCE [o] BYT3_30F.SCIENCE [o]
BYT3_30G.SCIENCE [o] BYT3_30H.SCIENCE [o] BYT3_31.SCIENCE [o]
BYT3_32.SCIENCE [o] BYT1_2.SOC.STUDIES.HISTORY [n]
BYT1_3.SOC.STUDIES.HISTORY [n] BYT1_4.SOC.STUDIES.HISTORY [n]
BYT1_6.SOC.STUDIES.HISTORY [n] BYT1_7.SOC.STUDIES.HISTORY [n]
BYT1_8.SOC.STUDIES.HISTORY [n] BYT2_2.SOC.STUDIES.HISTORY [n]
BYT2_3.SOC.STUDIES.HISTORY [s] BYT2_7H.SOC.STUDIES.HISTORY [s]
BYT2_7M.SOC.STUDIES.HISTORY [s] BYT2_8A.SOC.STUDIES.HISTORY [o]
BYT2_8B.SOC.STUDIES.HISTORY [o] BYT2_8C.SOC.STUDIES.HISTORY [o]
BYT2_9A.SOC.STUDIES.HISTORY [o] BYT2_9B.SOC.STUDIES.HISTORY [o]
BYT2_9C.SOC.STUDIES.HISTORY [o] BYT2_9D.SOC.STUDIES.HISTORY [o]
BYT2_11.SOC.STUDIES.HISTORY [o] BYT2_12A.SOC.STUDIES.HISTORY [o]
BYT2_12B.SOC.STUDIES.HISTORY [o] BYT2_12C.SOC.STUDIES.HISTORY
[o] BYT2_12D.SOC.STUDIES.HISTORY [o]
BYT2_12E.SOC.STUDIES.HISTORY [o] BYT2_12F.SOC.STUDIES.HISTORY
[o] BYT2_13A.SOC.STUDIES.HISTORY [n]
BYT2_13B.SOC.STUDIES.HISTORY [n] BYT2_13C.SOC.STUDIES.HISTORY
[n] BYT2_13D.SOC.STUDIES.HISTORY [n]
BYT2_13E.SOC.STUDIES.HISTORY [n] BYT2_13F.SOC.STUDIES.HISTORY
[n] BYT2_13G.SOC.STUDIES.HISTORY [n] BYT2_14.SOC.STUDIES.HISTORY
[o] BYT2_15.SOC.STUDIES.HISTORY [s] BYT2_16A.SOC.STUDIES.HISTORY
[o] BYT2_16B.SOC.STUDIES.HISTORY [o]
BYT2_16C.SOC.STUDIES.HISTORY [o] BYT2_16D.SOC.STUDIES.HISTORY
[o] BYT2_16E.SOC.STUDIES.HISTORY [o]
BYT2_16F.SOC.STUDIES.HISTORY [o] BYT2_16G.SOC.STUDIES.HISTORY
[o] BYT3_1.SOC.STUDIES.HISTORY [n] BYT3_2.SOC.STUDIES.HISTORY [n]
BYT3_3Y.SOC.STUDIES.HISTORY [o] BYT3_4.SOC.STUDIES.HISTORY [o]
BYT3_5.SOC.STUDIES.HISTORY [o] BYT3_7A.SOC.STUDIES.HISTORY [n]
BYT3_7B.SOC.STUDIES.HISTORY [n] BYT3_7C.SOC.STUDIES.HISTORY [n]

BYT3_7D.SOC.STUDIES.HISTORY [n] BYT3_8.SOC.STUDIES.HISTORY [n]
BYT3_9A1.SOC.STUDIES.HISTORY [n] BYT3_9A2.SOC.STUDIES.HISTORY
[n] BYT3_9B1.SOC.STUDIES.HISTORY [n]
BYT3_9B2.SOC.STUDIES.HISTORY [n] BYT3_9C1.SOC.STUDIES.HISTORY
[n] BYT3_9C2.SOC.STUDIES.HISTORY [n]
BYT3_9E2.SOC.STUDIES.HISTORY [n] BYT3_9G1.SOC.STUDIES.HISTORY
[n] BYT3_9G2.SOC.STUDIES.HISTORY [n]
BYT3_10A.SOC.STUDIES.HISTORY [n] BYT3_11B.SOC.STUDIES.HISTORY
[o] BYT3_12C.SOC.STUDIES.HISTORY [o] BYT3_13.SOC.STUDIES.HISTORY
[n] BYT3_19.SOC.STUDIES.HISTORY [o] BYT3_20A.SOC.STUDIES.HISTORY
[n] BYT3_20B.SOC.STUDIES.HISTORY [n]
BYT3_20C.SOC.STUDIES.HISTORY [n] BYT3_20D.SOC.STUDIES.HISTORY
[n] BYT3_20E.SOC.STUDIES.HISTORY [n] BYT3_21.SOC.STUDIES.HISTORY
[n] BYT3_26A.SOC.STUDIES.HISTORY [o]
BYT3_26B.SOC.STUDIES.HISTORY [o] BYT3_26C.SOC.STUDIES.HISTORY
[o] BYT3_26D.SOC.STUDIES.HISTORY [o]
BYT3_26E.SOC.STUDIES.HISTORY [o] BYT3_26F.SOC.STUDIES.HISTORY
[o] BYT3_26G.SOC.STUDIES.HISTORY [o]
BYT3_26H.SOC.STUDIES.HISTORY [o] BYT3_26I.SOC.STUDIES.HISTORY [o]
BYT3_26J.SOC.STUDIES.HISTORY [o] BYT3_26K.SOC.STUDIES.HISTORY [o]
BYT3_27.SOC.STUDIES.HISTORY [n] BYT3_28.SOC.STUDIES.HISTORY [o]
BYT3_29.SOC.STUDIES.HISTORY [o] BYT3_30A.SOC.STUDIES.HISTORY [o]
BYT3_30B.SOC.STUDIES.HISTORY [o] BYT3_30C.SOC.STUDIES.HISTORY
[o] BYT3_30D.SOC.STUDIES.HISTORY [o]
BYT3_30E.SOC.STUDIES.HISTORY [o] BYT3_30F.SOC.STUDIES.HISTORY
[o] BYT3_30G.SOC.STUDIES.HISTORY [o]
BYT3_30H.SOC.STUDIES.HISTORY [o] BYT3_31.SOC.STUDIES.HISTORY [o]
BYT3_32.SOC.STUDIES.HISTORY [o] BYT1_2.MATH [n] BYT1_3.MATH [n]
BYT1_4.MATH [n] BYT1_6.MATH [n] BYT1_7.MATH [n] BYT1_8.MATH [n]
BYT2_2.MATH [n] BYT2_3.MATH [s] BYT2_7H.MATH [s] BYT2_7M.MATH [s]
BYT2_8A.MATH [o] BYT2_8B.MATH [o] BYT2_8C.MATH [o] BYT2_9B.MATH
[o] BYT2_9C.MATH [o] BYT2_9D.MATH [o] BYT2_11.MATH [o]
BYT2_12A.MATH [o] BYT2_12B.MATH [o] BYT2_12C.MATH [o]
BYT2_12D.MATH [o] BYT2_12E.MATH [o] BYT2_12F.MATH [o]
BYT2_13A.MATH [n] BYT2_13B.MATH [n] BYT2_13C.MATH [n]
BYT2_13D.MATH [n] BYT2_13E.MATH [n] BYT2_13F.MATH [n]
BYT2_13G.MATH [n] BYT2_14.MATH [o] BYT2_15.MATH [s]
BYT2_16A.MATH [o] BYT2_16B.MATH [o] BYT2_16C.MATH [o]
BYT2_16D.MATH [o] BYT2_16E.MATH [o] BYT2_16F.MATH [o]
BYT2_16G.MATH [o] BYT3_1.MATH [n] BYT3_2.MATH [n] BYT3_3Y.MATH
[o] BYT3_4.MATH [o] BYT3_5.MATH [o] BYT3_7A.MATH [n]
BYT3_7B.MATH [n] BYT3_7C.MATH [n] BYT3_7D.MATH [n] BYT3_8.MATH
[n] BYT3_9A1.MATH [n] BYT3_9A2.MATH [n] BYT3_9B2.MATH [n]

BYT3_9C1.MATH [n] BYT3_9C2.MATH [n] BYT3_9D1.MATH [n]
BYT3_9D2.MATH [n] BYT3_9E1.MATH [n] BYT3_9E2.MATH [n]
BYT3_9G1.MATH [n] BYT3_9G2.MATH [n] BYT3_10A.MATH [n]
BYT3_11B.MATH [o] BYT3_12C.MATH [o] BYT3_13.MATH [n]
BYT3_19.MATH [o] BYT3_20A.MATH [n] BYT3_20B.MATH [n]
BYT3_20C.MATH [n] BYT3_20D.MATH [n] BYT3_20E.MATH [n]
BYT3_21.MATH [n] BYT3_26A.MATH [o] BYT3_26B.MATH [o]
BYT3_26C.MATH [o] BYT3_26D.MATH [o] BYT3_26E.MATH [o]
BYT3_26F.MATH [o] BYT3_26G.MATH [o] BYT3_26H.MATH [o]
BYT3_26I.MATH [o] BYT3_26J.MATH [o] BYT3_26K.MATH [o]
BYT3_27.MATH [n] BYT3_28.MATH [o] BYT3_29.MATH [o] BYT3_30A.MATH
[o] BYT3_30B.MATH [o] BYT3_30C.MATH [o] BYT3_30D.MATH [o]
BYT3_30E.MATH [o] BYT3_30F.MATH [o] BYT3_30G.MATH [o]
BYT3_30H.MATH [o] BYT3_31.MATH [o] BYT3_32.MATH [o] BYT2_3.ALL [s]
BYT2_6.ALL [s] BYT2_7H.ALL [s] BYT2_7M.ALL [s] BYT2_15.ALL [s]
BYT2_2.ALL [n] BYT3_2.ALL [n] BYT3_8.ALL [n] BYT3_27.ALL [n]
BYT1_2.ALL [n] BYT1_3.ALL [n] BYT1_6.ALL [n] BYT1_8.ALL [n]
BYT2_13A.ALL [n] BYT2_13B.ALL [n] BYT2_13C.ALL [n] BYT2_13D.ALL [n]
BYT2_13E.ALL [n] BYT2_13F.ALL [n] BYT2_21.ALL [n] BYT2_27A.ALL [n]
BYT2_27B.ALL [n] BYT3_1.ALL [n] BYT3_7A.ALL [n] BYT3_7B.ALL [n]
BYT3_7C.ALL [n] BYT3_7D.ALL [n] BYT3_9A1.ALL [n] BYT3_9C1.ALL [n]
BYT3_9G1.ALL [n] BYT3_10A.ALL [n] BYT310A1.ALL [n] BYT310A2.ALL [n]
BYT310C1.ALL [n] BYT310E1.ALL [n] BYT310G1.ALL [n] BYT3_20A.ALL [n]
BYT3_20B.ALL [n] BYT3_20C.ALL [n] BYT3_20E.ALL [n] BYT3_21.ALL [n]
BYT2_8A.ALL [o] BYT2_8B.ALL [o] BYT2_8C.ALL [o] BYT2_9B.ALL [o]
BYT2_9C.ALL [o] BYT2_9D.ALL [o] BYT2_11.ALL [o] BYT2_12A.ALL [o]
BYT2_12B.ALL [o] BYT2_12C.ALL [o] BYT2_12D.ALL [o] BYT2_12E.ALL [o]
BYT2_12F.ALL [o] BYT2_14.ALL [o] BYT2_16A.ALL [o] BYT2_16B.ALL [o]
BYT2_16C.ALL [o] BYT2_16D.ALL [o] BYT2_16E.ALL [o] BYT2_16F.ALL [o]
BYT2_16G.ALL [o] BYT2_17A.ALL [o] BYT2_17B.ALL [o] BYT2_17C.ALL [o]
BYT2_17D.ALL [o] BYT2_17E.ALL [o] BYT2_17F.ALL [o] BYT2_18A.ALL [o]
BYT2_18B.ALL [o] BYT2_18C.ALL [o] BYT2_18D.ALL [o] BYT2_18E.ALL [o]
BYT2_18F.ALL [o] BYT2_18G.ALL [o] BYT2_19.ALL [o] BYT2_20A.ALL [o]
BYT2_20B.ALL [o] BYT2_20C.ALL [o] BYT2_20D.ALL [o] BYT2_20E.ALL [o]
BYT2_20F.ALL [o] BYT2_20G.ALL [o] BYT2_20H.ALL [o] BYT2_20I.ALL [o]
BYT2_20J.ALL [o] BYT2_22.ALL [o] BYT2_23A.ALL [o] BYT2_23B.ALL [o]
BYT2_23C.ALL [o] BYT2_23D.ALL [o] BYT2_23E.ALL [o] BYT2_23F.ALL [o]
BYT2_23G.ALL [o] BYT2_23H.ALL [o] BYT2_24A.ALL [o] BYT2_24B.ALL [o]
BYT2_24C.ALL [o] BYT2_24D.ALL [o] BYT2_24E.ALL [o] BYT2_24F.ALL [o]
BYT2_24G.ALL [o] BYT2_24H.ALL [o] BYT2_24I.ALL [o] BYT2_24J.ALL [o]
BYT2_24K.ALL [o] BYT2_24L.ALL [o] BYT2_24M.ALL [o] BYT2_24N.ALL [o]
BYT2_24O.ALL [o] BYT2_24P.ALL [o] BYT2_24Q.ALL [o] BYT2_25.ALL [o]
BYT2_26.ALL [o] BYT2_28.ALL [o] BYT2_29.ALL [o] BYT3_3Y.ALL [o]

BYT3_4.ALL [o] BYT3_5.ALL [o] BYT3_11B.ALL [o] BYT3_12C.ALL [o] BYT3_19.ALL [o] BYT3_26A.ALL [o] BYT3_26B.ALL [o] BYT3_26C.ALL [o] BYT3_26D.ALL [o] BYT3_26E.ALL [o] BYT3_26F.ALL [o] BYT3_26G.ALL [o] BYT3_26H.ALL [o] BYT3_26I.ALL [o] BYT3_26J.ALL [o] BYT3_26K.ALL [o] BYT3_28.ALL [o] BYT3_29.ALL [o] BYT3_30A.ALL [o] BYT3_30B.ALL [o] BYT3_30C.ALL [o] BYT3_30D.ALL [o] BYT3_30E.ALL [o] BYT3_30F.ALL [o] BYT3_30G.ALL [o] BYT3_30H.ALL [o] BYT3_31.ALL [o] BYT3_32.ALL [o] BYT3_33.ALL [o] FAMCOMP [n] G8CTRL1 [n] G8CTRL2 [n] G10REGON [n] G12COHRT [n] G12CTRL1 [n] G12URBN3 [n] G12REGON [n] TRNURBN3 [n] TRNREGON [n] F2PNLWT [s] F1S20 [n] F1HSPROG [n] F1S15A [n] F1S15B [n] F1S15D [n] F1S16A [n] F1S16E [n] F1S16F [n] F1S26A [o] F1S26B [o] F1S26C [o] F1S26D [o] F1S28A [o] F1S28B [o] F1S28C [o] F1S28D [o] F1S29A [o] F1S29B [o] F1S29C [o] F1S29D [o] F1S29E [o] F1S29F [o] F1S29G [o] F1S29H [o] F1S29L [o] F1S29M [o] F1S29N [o] F1S30A [o] F1S30B [o] F1S30C [o] F1S30D [o] F1S30E [o] F1S31A [o] F1S31B [o] F1S31C [o] F1S31D [o] F1S31E [o] F1S32A [o] F1S32B [o] F1S32C [o] F1S32D [o] F1S32E [o] F1S32F [o] F1S32G [o] F1S32H [o] F1S32I [o] F1S7A [o] F1S7B [o] F1S7C [o] F1S7D [o] F1S7F [o] F1S7G [o] F1S7H [o] F1S7O [o] F1C91A [o] F1C91B [o] F1C91C [o] F1C91D [o] F1C91E [o] F1C91F [o] F1C91G [o] F1C91H [o] F1C93A [o] F1C93B [o] F1C93C [o] F1C93D [o] F1C93E [o] F1C93F [o] F1C93G [o] F1C93H [o] F1C93I [o] F1C93J [o] F1C93K [o] F1C93L [o] F1C93M [o] F1C97A [o] F1C97B [o] F1C97C [o] F1C97D [o] F1C97E [o] F1C97F [o] F1C97G [o] F1C97H [o] F1C97I [o] F1C97J [o] F1C97K [o] F1C97L [o] F1C97M [o] F1C98A [o] F1C98B [o] F1C98C [o] F1C98D [o] F1C98E [o] F1C98F [o] F1C98G [o] F1C98H [o] F1C98I [o] F1C98J [o] F1C98K [o] F1C98L [o] F1C103A [o] F1C103B [o] F1C103C [o] F1C103D [o] F1SCH_ID [s] F1C6 [n] F1C7 [o] F1C8 [o] F1C9 [o] F1C11A [s] F1C11B [s] F1C11C2 [s] F1C11C3 [s] F1C11C4 [s] F1C11C5 [s] F1C11C6 [s] F1C11C7 [s] F1C11C8 [s] F1C11C9 [s] F1C12A [o] F1C12B [o] F1C12C [o] F1C12D [o] F1C12E [o] F1C12F [o] F1C12G [o] F1C12H [o] F1C12I [o] F1C12J [o] F1C12M [o] F1C13A [o] F1C13B [o] F1C13C [o] F1C13D [o] F1C13E [o] F1C13F [o] F1C13G [o] F1C13H [o] F1C13I [o] F1C13J [o] F1C14 [n] F1C17A [o] F1C17B [o] F1C17C [o] F1C17D [o] F1C17E [o] F1C17F [o] F1C18A [n] F1C18B [n] F1C18C [n] F1C18D [n] F1C18E [n] F1C18F [n] F1C18G [n] F1C18H [n] F1C18I [n] F1C18J [n] F1C18K [n] F1C18M [n] F1C19 [o] F1C20 [o] F1C21 [o] F1C22A [n] F1C22B [n] F1C22C [n] F1C22D [n] F1C22E [n] F1C23 [n] F1C24 [s] F1C25 [s] F1C26 [s] F1C27F [o] F1C28 [o] F1C29 [o] F1C30A [o] F1C30B [s] F1C30C [s] F1C30D [s] F1C30E [s] F1C30F [s] F1C30G [s] F1C30H [s] F1C30I [s] F1C30J [s] F1C30K [s] F1C32 [s] F1C33 [s] F1C34 [s] F1C35 [o] F1C36 [o] F1C37 [n] F1C37A [s] F1C38 [n] F1C39 [n] F1C40A [n] F1C40B [n] F1C40C [n] F1C41A [o] F1C41B [o] F1C41C [o] F1C41D [o] F1C41E [o] F1C41F [o] F1C41G [o] F1C41H [o] F1C41I [o] F1C41J [o] F1C41L [o] F1C42A [o] F1C42B [o] F1C43A [o] F1C43B [o] F1C43C [o] F1C43D [o] F1C43E [o] F1C44A [o] F1C44B [o] F1C44C [o] F1C44D [o] F1C44E [o] F1C45 [s] F1C46 [s] F1C47A [o] F1C47B [o] F1C47C [o]

F1C48 [s] F1C49 [o] F1C50 [s] F1C51 [n] F1C52 [n] F1C53A [n] F1C53B [n] F1C53C [n] F1C53D [n] F1C53G [n] F1C53H [n] F1C53I [n] F1C53J [n] F1C54A [o] F1C54B [o] F1C54C [o] F1C54D [o] F1C55 [n] F1C58A [n] F1C61A [n] F1C61B [n] F1C61C [n] F1C61D [n] F1C62A [o] F1C62B [o] F1C62C [o] F1C62D [o] F1C62E [o] F1C62F [o] F1C62G [o] F1C62H [o] F1C63A [o] F1C63B [o] F1C64 [o] F1C65 [n] F1C66B [n] F1C66C [n] F1C66D [n] F1C66E [n] F1C67 [o] F1C68 [n] F1C69 [n] F1C69AA1 [s] F1C69AB1 [s] F1C69AB2 [s] F1C69AC1 [s] F1C69AC2 [s] F1C69AD1 [s] F1C69AD2 [s] F1C69AF2 [s] F1C69B [o] F1C70A [o] F1C70B [o] F1C70C [o] F1C70D [o] F1C70E [o] F1C70F [o] F1C70G [o] F1C70H [o] F1C70I [o] F1C70J [o] F1C70K [o] F1C70L [o] F1C70M [o] F1C70N [o] F1C70O [o] F1C71A [n] F1C71C [n] F1C71D [n] F1C71E [n] F1C71F [n] F1C71G [n] F1C71H [n] F1C71I [n] F1C71J [n] F1C71K [n] F1C71L [n] F1C71N [n] F1C71O [n] F1C71P [n] F1C71Q [n] F1C71R [n] F1C71T [n] F1C71U [n] F1C71W [n] F1C72 [n] F1C73A1 [n] F1C73A2 [n] F1C73A3 [n] F1C73A4 [n] F1C73B1 [n] F1C73B2 [n] F1C73B3 [n] F1C73B4 [n] F1C73C1 [n] F1C73C2 [n] F1C73C3 [n] F1C73C4 [n] F1C73D1 [n] F1C73D2 [n] F1C73D3 [n] F1C73D4 [n] F1C73E1 [n] F1C73E2 [n] F1C73E3 [n] F1C73E4 [n] F1C73F1 [n] F1C73F2 [n] F1C73F3 [n] F1C73F4 [n] F1C73G1 [n] F1C73G2 [n] F1C73G3 [n] F1C73G4 [n] F1C73H1 [n] F1C73H2 [n] F1C73H3 [n] F1C73H4 [n] F1C73I1 [n] F1C73I2 [n] F1C73I3 [n] F1C73I4 [n] F1C73J1 [n] F1C73J3 [n] F1C73J4 [n] F1C73K1 [n] F1C73K2 [n] F1C73K3 [n] F1C73K4 [n] F1C73L1 [n] F1C73L2 [n] F1C73L3 [n] F1C73L4 [n] F1C73M1 [n] F1C73M2 [n] F1C73M3 [n] F1C73M4 [n] F1C73N1 [n] F1C73N2 [n] F1C73N3 [n] F1C73N4 [n] F1C73O1 [n] F1C73O2 [n] F1C73O3 [n] F1C73O4 [n] F1C73P1 [n] F1C73P2 [n] F1C73P3 [n] F1C73P4 [n] F1C73Q1 [n] F1C73Q3 [n] F1C73Q4 [n] F1C73R1 [n] F1C73R2 [n] F1C73R3 [n] F1C73R4 [n] F1C73S1 [n] F1C73S2 [n] F1C73S3 [n] F1C73S4 [n] F1C74A1 [n] F1C74A2 [n] F1C74A3 [n] F1C74A4 [n] F1C74A5 [n] F1C74C1 [n] F1C74C2 [n] F1C74C3 [n] F1C74C4 [n] F1C74C5 [n] F1C75A1 [n] F1C75A2 [n] F1C75A3 [n] F1C75B1 [n] F1C75B2 [n] F1C75C1 [n] F1C75C2 [n] F1C75C3 [n] F1C75D1 [n] F1C75D2 [n] F1C75E1 [n] F1C75E2 [n] F1C75F3 [n] F1C75F4 [n] F1C75G1 [n] F1C75G2 [n] F1C75G3 [n] F1C75G4 [n] F1C75H2 [n] F1C75H3 [n] F1C75H4 [n] F1C75I1 [n] F1C75I2 [n] F1C75I3 [n] F1C75I4 [n] F1C75J2 [n] F1C75J3 [n] F1C75J4 [n] F1C75K1 [n] F1C75K2 [n] F1C75K3 [n] F1C75L1 [n] F1C75L2 [n] F1C75M1 [n] F1C75M2 [n] F1C75N1 [n] F1C75N2 [n] F1C75O1 [n] F1C75O2 [n] F1C75P1 [n] F1C75P2 [n] F1C75Q1 [n] F1C75Q2 [n] F1C75R1 [n] F1C75R2 [n] F1C75S1 [n] F1C75S2 [n] F1C75T1 [n] F1C75T2 [n] F1C75U1 [n] F1C75U2 [n] F1C75V1 [n] F1C75V2 [n] F1C75W1 [n] F1C75W2 [n] F1C75Y2 [n] F1C75Y3 [n] F1C75Y4 [n] F1C75Z4 [n] F1C75AA1 [n] F1C75AA2 [n] F1C75AA3 [n] F1C75AA4 [n] F1C75BB1 [n] F1C75BB2 [n] F1C75CC1 [n] F1C75CC2 [n] F1C75CC3 [n] F1C75DD1 [n] F1C75DD2 [n] F1C75DD3 [n] F1C75DD4 [n] F1C75EE1 [n] F1C75EE2 [n] F1C75EE3 [n] F1C75EE4 [n] F1C75FF1 [n] F1C75FF2 [n] F1C75FF3 [n] F1C75GG1 [n] F1C75GG2 [n] F1C75HH1 [n] F1C75HH2 [n] F1C75HH3 [n] F1C76 [s] F1C78 [s] F1C79B [n] F1C79C [n] F1C79E [n] F1C79F

[n] F1C79G [n] F1C79H [n] F1C80 [s] F1C82 [n] F1C83 [s] F1C84B [n] F1C84C [n] F1C84D [n] F1C84E [n] F1C84F [n] F1C84G [n] F1C85A [n] F1C85B [n] F1C85C [n] F1C85D [n] F1C86 [n] F1C88A [n] F1C88B [n] F1C88C [n] F1C88D [n] F1C88E [n] F1C88F [n] F1C88G [n] F1C88H [n] F1C88I [n] F1C88J [n] F1C89 [n] F1C92A [s] F1C92B [s] F1C92C [s] F1C92D [s] F1C94B [n] F1C94C [n] F1C94D [n] F1C94E [n] F1C94F [n] F1C94G [n] F1C94H [n] F1C94I [n] F1C95A [o] F1C95B [o] F1C95C [o] F1C95D [o] F1C95E [o] F1C95F [o] F1C95G [o] F1C95H [o] F1C95I [o] F1C95J [o] F1C95K [o] F1C95L [o] F1C95M [o] F1C96A1 [n] F1C96A2 [n] F1C96A3 [n] F1C96B1 [n] F1C96B2 [n] F1C96B3 [n] F1C96C1 [n] F1C96C2 [n] F1C96C3 [n] F1C96D1 [n] F1C96D2 [n] F1C96D3 [n] F1C96D5 [n] F1C96E2 [n] F1C96E3 [n] F1C96E4 [n] F1C96E5 [n] F1C96F2 [n] F1C96F3 [n] F1C96F4 [n] F1C96F5 [n] F1C96G2 [n] F1C96G3 [n] F1C96G4 [n] F1C96G5 [n] F1C96H3 [n] F1C96H4 [n] F1C96H5 [n] F1C96I2 [n] F1C96I3 [n] F1C96I4 [n] F1C96I5 [n] F1C96J2 [n] F1C96J3 [n] F1C96J4 [n] F1C96J5 [n] F1C96K2 [n] F1C96K3 [n] F1C96K4 [n] F1C96K5 [n] F1C96L1 [n] F1C96L2 [n] F1C96L3 [n] F1C96M1 [n] F1C96M2 [n] F1C96M3 [n] F1C96M5 [n] F1C96N3 [n] F1C96N4 [n] F1C96N5 [n] F1C96O2 [n] F1C96O3 [n] F1C96O4 [n] F1C96O5 [n] F1C96P1 [n] F1C96P2 [n] F1C96P3 [n] F1C96Q1 [n] F1C96Q2 [n] F1C96Q3 [n] F1C96AA1 [n] F1C96AA2 [n] F1C96AA3 [n] F1C96AA5 [n] F1C96BB1 [n] F1C96BB2 [n] F1C96BB3 [n] F1C96BB4 [n] F1C96BB5 [n] F1C96CC1 [n] F1C96CC2 [n] F1C96CC3 [n] F1C96CC4 [n] F1C96CC5 [n] F1C96DD1 [n] F1C96DD2 [n] F1C96DD3 [n] F1C96DD4 [n] F1C96DD5 [n] F1C96EE2 [n] F1C96EE3 [n] F1C96EE4 [n] F1C96EE5 [n] F1C96FF2 [n] F1C96FF3 [n] F1C96FF4 [n] F1C96FF5 [n] F1C96GG3 [n] F1C96GG4 [n] F1C96GG5 [n] F1C96HH3 [n] F1C96HH4 [n] F1C96HH5 [n] F1C96II3 [n] F1C96II4 [n] F1C96II5 [n] F1C96JJ3 [n] F1C96JJ4 [n] F1C96JJ5 [n] F1C96KK3 [n] F1C96KK4 [n] F1C96KK5 [n] F1C96LL2 [n] F1C96LL3 [n] F1C96LL5 [n] F1C96MM2 [n] F1C96MM3 [n] F1C96MM4 [n] F1C96MM5 [n] F1C96NN3 [n] F1C96NN4 [n] F1C96NN5 [n] F1C96OO2 [n] F1C96OO3 [n] F1C96OO4 [n] F1C96OO5 [n] F1C96PP1 [n] F1C96PP2 [n] F1C96PP3 [n] F1C96PP4 [n] F1C96PP5 [n] F1C96QQ1 [n] F1C96QQ2 [n] F1C96QQ3 [n] F1C96QQ4 [n] F1C96QQ5 [n] F1C99A [o] F1C99B [o] F1C99C [o] F1C99D [o] F1C99E [o] F1C100 [s] F1C101 [s] F1C102A [o] F1C102B [o] F1C102C [o] F1C102D [o] F1C102E [o] F1C102F [o] F1C102G [o] G10CTRL1.1 [n] G10URBAN.1 [n] F1SCENRL.1 [n] G10ENROL.1 [n] F1SGSPAN [n] F1T2_13C.ALL [n] F1T2_13D.ALL [n] F1T2_13E.ALL [n] F2S7A [o] F2S7B [o] F2S7C [o] F2S7D [o] F2S7G [o] F2S7H [o] F2S7I [o] F2S7J [o] F2S7K [o] F2S7L [o] F2HSPROG [n] F2P42A [o] F2P42G [o] F2P42H [o] F2P42I [o] F2P42J [o] F2P42K [o] F2P42L [o] F2P42M [o] F2P42N [o] F2P42O [o] F2P42P [o] F2P42Q [o] F2P42R [o] F2P42S [o] F2P42T [o] F2P42U [o] F2P43A [o] F2P43B [o] F2P43C [o] F2P43D [o] F2T3_4 [o] F2T3_5A [o] F2T3_5F [o] F2T3_5G [o] F2T3_5H [o] F2T3_5I [o] F2T3_5J [o] F2T3_5K [o] F2T3_5L [o] F2T3_5M [o] F2T3_5N [o] F2T3_6A [o] F2T3_6B [o] F2T3_6C [o] F2T3_6D [o] F2T3_6E [o] F2T3_6F [o] F2T3_7A [o] F2T3_7B [o] F2T3_7C [o] F2T3_7D [o] F2T3_7E [o]

F2T3_7F [o] F2T3_7G [o] F2T3_7H [o] F2T3_7I [o] F2T3_15A [o] F2T3_15B [o] F2T3_15C [o] F2T3_16A [o] F2T3_16B [o] F2T3_16C [o] F2T3_16D [o] F2T3_16E [o] F2T3_16F [o] F2T3_16G [o] F2T3_16H [o] F2T3_16I [o] F2T3_16J [o] F2T3_16K [o] F2T3_16L [o] F2T3_16M [o] F2T3_16N [o] F2T3_16O [o] F2T3_16P [o] F2T4_19A [n] F2T4_19B [n] F2T4_19C [n] F2T4_19D [n] F2T4_19E [n] F2T4_19F [n] F2T4_19G [n] F2C60G [o] F2C60H [o] F2C61A [o] F2C61B [o] F2C61C [o] F2C61D [o] F2C61E [o] F2CRDRQ1 [s] F2CRDRQ2 [s] /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO
/DEPCATEGORIES USEVALUES=[VALID]
/PRINT MODELSUMMARY CLASSIFICATION RISK
/SAVE PREDVAL PREDPROB
/GAIN CATEGORYTABLE=YES TYPE=[NODE] SORT=DESCENDING CUMULATIVE=NO
/RULES NODES=TERMINAL SYNTAX=GENERIC LABELS=YES OUTFILE='C:\Users\Emi\Desktop\rules.txt'
/METHOD TYPE=QUEST MAXSURROGATES=AUTO PRUNE=SE(1)
/GROWTHLIMIT MAXDEPTH=AUTO MINPARENTSIZE=2000 MINCHILDSIZE=1232
/VALIDATION TYPE=SPLITSAMPLE(training) OUTPUT=BOTHSAMPLES
/QUEST ALPHASPLIT=0.05
/COSTS EQUAL
/PRIORS FROMDATA ADJUST=NO
/MISSING NOMINALMISSING=MISSING.

23. Need to follow up with each ruleset, to calculate relevant accuracy measures for each rule.

Association Rule Mining

1. Open large dataset created above, before expansion (nels_bla_NOTexpanded_v1.Rda).
2. Reduce cases to only those who got lower than 16.649 in 8th grade math score
3. Shuffle the data
4. Calculate residuals according to CART
5. Drop cases without residuals, and cases that have residuals between 7.6 and 10. Create dichotomous outcome variable of the residuals (those who scored at least 10 points above predicted value vs those who are 7.5 points or below predicted value)
6. Get rid of all commas in the dataset
7. Convert variables in dataset to character (or non-numeric), if they are not to be used as part of the data mining.
8. Recode all numeric variables in dataset into ordered factors (split into 4 groups)
9. Stratified random sampling into training and test sets (70-30 split, stratified by outcome)
10. Stratified random sampling of training set into generation and screening sets (50-50 split, stratified by outcome)

11. Expand datasets according to weights (make sure weights close to 0 get rounded up to 1).
12. Convert STU_ID into factor, and delete all the other variables that are in the right format (i.e., delete characters).
13. Split each of the datasets (generation, screening, training, test) by outcome (high achievers vs not)
14. Convert datasets (that are dataframes) into transactional form
15. Remove unneeded datasets to save memory
16. Set minimum support at 25% (since many questions are asked in a 4 point-scale). Confidence is set at 1 (since the dataset only contains high-achievers).
17. Run association rules using the generation set (for high achievers)

```
rules1 <- apriori(gen_in_t,
        parameter=list(support= sup, confidence=conf, minlen=3, maxlen=3,
maxtime = 50),
        appearance = list(rhs=c("gr12math=high"), default="lhs"),
        control=list(memopt=TRUE, load=FALSE)
        )
print("Summary of generation rules")
summary(rules1)
```

18. Calculate support in screening set, and only retain subset with support >=.25.

```
quality(rules1)<-cbind(quality(rules1),
        support_s=interestMeasure(rules1,
                measure="support",
                transactions=scr_in_t,
                reuse=FALSE
        ))

rules1<- subset(rules1, support_s >=sup)

print("Summary of rules that have required stats in screening set as well")
summary(rules1)
```

19. Calculate total training set coverage

```
quality(rules1)<-cbind(quality(rules1),
        coverage_tr_in=interestMeasure(rules1,
                measure="coverage",
                transactions=train_in_t,
                reuse=FALSE
        ))

quality(rules1)<-cbind(quality(rules1),
        coverage_tr_out=interestMeasure(rules1,
                measure="coverage",
                transactions=train_out_t,
                reuse=FALSE
```

```
                    ))
```

20. Calculate total test set coverage

```
quality(rules1)<-cbind(quality(rules1),
                    coverage_ts_in=interestMeasure(rules1,
                                                   measure="coverage",
                                                   transactions=test_in_t,
                                                   reuse=FALSE
                    ))
quality(rules1)<-cbind(quality(rules1),
                    coverage_ts_out=interestMeasure(rules1,
                                                    measure="coverage",
                                                    transactions=test_out_t,
                                                    reuse=FALSE
                    ))

print("Summary of rules")
summary(rules1)
```

21. Save rules as a dataframe and write as CSV

```
rules1.df <- as(rules1, "data.frame")
saveRDS(rules1.df, file="BL_ARgr8mLT17hi_len3.Rda")
write.csv(rules1.df, file = "BL_ARgr8mLT17hi_len3.csv")
```

22. Need to conduct additional analyses using the output to identify which rules are most relevant.

Run rCBA

Note: This only worked when the dataset was small.

1. Open small dataset.
2. Set mathach to factor, take out STU_ID and numeric version of math achievement
3. Shuffle the dataset
4. Recode all numeric variables in dataset into ordered factors (split into 4 groups)
5. Get rid of all the commas in the dataset
6. Stratified random sampling into training and test sets (70-30 split, stratified by outcome)
7. Split the training set into 50-50 (generate & screen)
8. Convert into transactional dataset, look at general stats
9. Remove unneeded data
10. Conduct Apriori (see steps above)
11. Convert rules to data frame

```
training.df <- as(rules1b, "data.frame")
```

12. Conduct CBA

```
# (First assign 40gigs of heap space to java environment per
http://www.bramschoenmakers.nl/en/node/726)
print(nrow(training.df))
options( java.parameters = "-Xmx40g")
```

```
library(rCBA)
prunedRulesFrame <- pruning(training.data, training.df, method="m2cba")
print(nrow(prunedRulesFrame))

#Save output
out2 <- capture.output(print(nrow(prunedRulesFrame)))
cat("n_pruned_rules", out2, file="bs_v3b_prunedRulesFrame.txt", sep="\n",
append=FALSE)
```

13. Evaluate how good the ruleset is on the test set

```
##Note, with the following program, no commas allowed in the dataset (or rules)
##Name the CBA output data frame as "prunedRulesFrame" (as I've done above)
##Name the to-be-evaluated-dataframe "eval.df" (as below)
##Results in dataset called "Output" with whether each of the rules applied and what the
final prediction was

eval.df <- test.data  #Designate data frame to be evaluated here!

prunedRulesFrame$ant<-gsub("} => .*|[{]", "", prunedRulesFrame$rules)
prunedRulesFrame$consq<-gsub(".*=> |[{}]", "", prunedRulesFrame$rules)
varname1 <- paste("Rule", row.names(prunedRulesFrame), sep = "_")
results <- data.frame(setNames(replicate(length(varname1), numeric(0), simplify = F),
varname1))

for (j in 1:((NROW(prunedRulesFrame))-1)) {
x0<-prunedRulesFrame[j, "consq"]
x1<-strsplit(prunedRulesFrame[j, "ant"], ",")[[1]]
x2<-unname(mapply(sub, "=", "zzz", x1))
value<-unname(mapply(sub, "zzz.*", "", x2))
attribute<-unname(mapply(sub, ".*zzz", "", x2))
df1<-cbind(value, attribute)
eval.df$pred<-NROW(df1)
myList<-list()
for (i in 1:NROW(df1)) myList[[paste('A', i)]] <-
  ifelse(eval.df[,print(match(df1[i,1], colnames(eval.df)))]==df1[i,2], 1, 0)
df2 <- data.frame(matrix(unlist(myList), nrow=NROW(eval.df), byrow=FALSE))
df3 <- merge(df2, eval.df, by=0)
if (NROW(df1)>1) df3$sumX <-rowSums(df3[,2:(NROW(df1)+1)])
   else (df3$sumX <- df3[,2])
results[1:NROW(df3),j] <- ifelse(df3$sumX==NROW(df1), x0, 0)
}
results[is.na(results)] <-0 #convert missing into 0 for the results
#find CBA prediction
results$prediction<-0
for (j in 1:NROW(results)) {
```

```
  for (i in 1:(NROW(prunedRulesFrame)-1)) {
    if (results[j, i] != 0) {results[j,"prediction"]<-results[j,i]; break}
  }
}
#Create final dataset
output <- merge(df3, results, by=0)
########################
table(output$mathach, output$prediction, exclude=NULL)

#Save CM
out3 <- capture.output(table(output$mathach, output$prediction, exclude=NULL))
cat("bs_rCBA_CM", out3, file="bs_rCBA_v3b_CM.txt", sep="\n", append=FALSE)

#Save
write.csv(output, file = "bs_testset_rCBA_v3b_pred.csv")
write.csv(prunedRulesFrame, file = "bs_rCBA_v3b_rules.csv")
######################
```

BIBLIOGRAPHY

Agrawal, R., Imieliński, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases.* Paper presented at the ACM SIGMOD Record.

AlShammari, I. A., Aldhafiri, M. D., & Al-Shammari, Z. (2013). A meta-analysis of educational data mining on improvements in learning outcomes. *College Student Journal, 47*(2), 326-333.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, D.C: National Research Council and National Academy of Education.

Azevedo, A. & Santos M. F. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. *IADIS*

Bache, S. M., & Wickham, H. (2014). Magrittr: A forward-pipe operator for R. Retrieved from https://CRAN.R-project.org/package=magrittr

Bailey, B. L. (2006). Let the data talk: Developing models to explain IPEDS graduation rates. *New Directions for Institutional Research, 2006*(131), 101-115.

Barnes, T. J. (2013). Big data, little history. *Dialogues in Human Geography, 3*(3), 297-302.

Bay, S. D., & Pazzani, M. J. (2000). *Discovering and describing category differences: What makes a discovered difference insightful?* Paper presented at the Proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. Washington DC: U.S. Department of Education, Office of Educational Technology.

Boström, H. (1995). *Covering vs. Divide-and-conquer for top-down induction of logic programs.* Paper presented at the Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.

Bourassa, M. A. J. (2011). *Interestingness: Guiding the search for significant information.* (Dissertation/Thesis), ProQuest, UMI Dissertations Publishing. Retrieved from http://pitt.idm.oclc.org/login?url=http://search.proquest.com/docview/926962494?accountid=14709

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics, 26*(3), 801-849.

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

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

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The x-random case. *International Statistial Review*, 291-319.

Brin, S., Motwani, R., Ullman, J., & Tsur, S. (1997, 1997). *Dynamic itemset counting and implication rules for market basket data.* Paper presented at the 1997 ACM SIGMOD International Conference on Management of Data, Tuscon, AZ.

Bringmann, B., Nijssen, S., & Zimmermann, A. (2009). *Pattern-based classification: A unifying perspective.* Paper presented at the The ECML/PKDD-09 Workshop (LeGo-09), Bled, Slovenia.

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics, 30*(4), 927-961.

Byrnes, J. P., & Miller, D. C. (2007). The relative importance of predictors of math and science achievement: An opportunity–propensity analysis. *Contemporary Educational Psychology, 32*(4), 599-629.

Carvalho, D. R., & Freitas, A. A. (2002). A genetic-algorithm for discovering small-disjunct rules in data mining. *Applied Soft Computing, 2*(2), 75-88.

Carvalho, D. R., & Freitas, A. A. (2004). A hybrid decision tree/genetic algorithm method for data mining. *Information Sciences, 163*(1–3), 13-35.

Carvalho, D. R., Freitas, A. A., & Ebecken, N. (2003). *A critical review of rule surprisingness measures.* Paper presented at the Proc. Data Mining IV-Int. Conf. on Data Mining.

Cendrowska, J. (1987). Prism: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies, 27*(4), 349-370.

Chandrahasan, R. K., Y, A. C., Sridhar, U. R., & L, A. (2011). An empirical comparison of boosting and bagging algorithms. *International Journal of Computer Science and Information Security, 9*(11), 147-152.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Crisp-dm 1.0: Step-by-step data mining guide*. Retrieved from https://the-modeling-agency.com/crisp-dm.pdf

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview *Data mining and knowledge discovery handbook* (pp. 853-867): Springer.

Clark, P., & Boswell, R. (1991). *Rule induction with cn2: Some recent improvements.* Paper presented at the Proceedings of the European working session on learning on Machine learning.

Clark, P., & Niblett, T. (1989). The cn2 induction algorithm. *Machine Learning, 3*(4), 261-283.

Cohen, W. W. (1995). *Fast effective rule induction.* Paper presented at the Proceedings of the twelfth international conference on machine learning.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*(2), 394-409.

Corrin, L., & de Barba, P. (2014). *Exploring students' interpretation of feedback delivered through learning analytics dashboards.* Paper presented at the Australasian Society for Computers in Learning in Tertiary Education, Dunedin, NZ.

Cota-Robles, E., & Gordan, E. (1999). Reaching the top: A report of the national task force on minority high achievement. *New York: The College Board*.

Curram, S. P., & Mingers, J. (1994). Neural networks, decision tree induction and discriminant analysis: An empirical comparison. *Journal of the Operational Research Society, 45*(4), 440-450.

Delen, D. (2006). Let the data talk: Developing models to explain IPEDS graduation rates. *New Directions for Institutional Research*(131), 101-115.

Delen, D. (2012). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice, 13*(1), 17-35.

Dietterich, T. G. (2000). Ensemble methods in machine learning *Multiple classifier systems* (Vol. 1857, pp. 1-15): Springer Berlin Heidelberg.

Dong, G., & Li, J. (1998). Interestingness of discovered association rules in terms of neighborhood-based unexpectedness *Research and development in knowledge discovery and data mining* (pp. 72-86): Springer.

ElAtia, S., Ipperciel, D., & Hammad, A. (2012). Implications and challenges to using data mining in educational research in the canadian context. *Canadian Journal of Education, 35*(2), 101-119.

Elkan, C. (2012). *Evaluating classifiers*. Retrieved from https://pdfs.semanticscholar.org/2bdc/61752a02783aa0e69e92fe6f9b449916a095.pdf

Eykamp, P. W. (2006). Using data mining to explore which students use advanced placement to reduce time to degree. *New Directions for Institutional Research*(131), 83-99.

Faulkner, R., Davidson, J. W., & McPherson, G. E. (2010). The value of data mining in music education research and some findings from its application to a study of instrumental learning during childhood. *International Journal of Music Education, 28*(3), 212-230.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). *Knowledge discovery and data mining: Towards a unifying framework.* Paper presented at the KDD.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining*. Menlo Park, CA: AAAI Press.

Finch, W. H., & Schneider, M. K. (2006). Misclassification rates for four methods of group classification: Impact of predictor distribution, covariance inequality, effect size, sample size, and group size ratio. *Educational and Psychological Measurement, 66*(2), 240-257.

Flores, R., Inan, F., & Lin, Z. (2013). How do the different types of computer use affect math achievement? *Journal of Computers in Mathematics and Science Teaching, 32*(1), 67-87.

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter, 12*(1), 49-57.

Frank, E., & Witten, I. H. (1998). *Generating accurate rule sets without global optimization.* Paper presented at the Proceedings of the Fifteenth International Conference on Machine Learning.

Freitas, A. A. (1998). On objective measures of rule surprisingness. In J. Żytkow & M. Quafafou (Eds.), *Principles of data mining and knowledge discovery* (Vol. 1510, pp. 1-9): Springer Berlin Heidelberg.

Freund, Y., & Schapire, R. E. (1995). *A desicion-theoretic generalization of on-line learning and an application to boosting.* Paper presented at the European conference on computational learning theory.

Freund, Y., & Schapire, R. E. (1996). *Experiments with a new boosting algorithm.* Paper presented at the Icml.

Fürnkranz, J. (1999). Separate-and-conquer rule learning. *Artificial Intelligence Review, 13*(1), 3-54.

Fürnkranz, J., & Flach, P. A. (2005). Roc 'n' rule learning—towards a better understanding of covering algorithms. *Machine Learning, 58*(1), 39-77.

Fürnkranz, J., Gamberger, D., & Lavrač, N. (2012). *Foundations of rule learning*. Berlin, Heidelberg: Springer.

Fürnkranz, J., & Widmer, G. (1994). *Incremental reduced error pruning.* Paper presented at the The 11th International Conference on Machine Learning (ML-94).

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(4), 463-484.

Gaševic, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends: Linking Research and Practice to Improve Learning, 59*(1), 64-71.

Geng, L., & Hamilton, H. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR), 38*(3), 9-es.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: Wiley.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22*(4), 521-563.

Goethals, B. (2002). *Efficient frequent pattern mining.* (Doctor in de Wetenscheppen, richting Informatica), LUC/UM. (D/2002/2541/46)

Grover, L. K., & Mehra, R. (2008). The lure of statistics in data mining. *Journal of Statistics Education, 16*(1).

Grzymala-Busse, J. W. (2005). Rule induction (pp. 277-294). Boston, MA: Springer US.

Guruler, H., Istanbullu, A., & Karahasan, M. (2010). A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education, 55*(1), 247-254.

Hahsler, M., Buchta, C., Gruen, B., & Hornik, K. (2016). Arules: Mining association rules and frequent itemsets. Retrieved from https://CRAN.R-project.org/package=arules

Hahsler, M., Chelluboina, S., Hornik, K., & Buchta, C. (2011). The arules R-package ecosystem: Analyzing interesting patterns from large transaction datasets. *Journal of Machine Learning Research, 12*, 1977-1981.

Hahsler, M., Gruen, B., & Hornik, K. (2005). Arules -- a computational environment for mining association rules and frequent item sets. *Journal of Statistical Software, 14*(15), 1-25.

Hahsler, M., Grün, B., Hornik, K., & Buchta, C. (2009). Introduction to arules–a computational environment for mining association rules and frequent item sets. *The Comprehensive R Archive Network.*

Hand, D. J. (1997). *Construction and assessment of classification rules*. Chichester; New York: Wiley.

Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician, 52*(2), 112-118.

Hand, D. J. (2000). Data mining: New challenges for statisticians. *Social Science Computer Review, 18*(4), 442-449.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-a-vis regression. *New Directions for Institutional Research*(131), 17-33.

Hidber, C. (1999). *Online association rule mining.* Paper presented at the 1999 ACM SIGMOD international conference on Management of data, New York, NY.

Hilderman, R. J., & Hamilton, H. J. (1999). *Knowledge discovery and interestingness measures: A survey*: Department of Computer Science, University of Regina.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832-844.

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two-group classification methods. *Educational and Psychological Measurement, 71*(5), 870-901.

Hong, J., Mozetic, I., & Michalski, R. S. (1986). *Aq15: Incremental learning of attribute-based descriptions from examples, the method and user's guide*. Retrieved from http://digilib.gmu.edu/jspui/bitstream/handle/1920/1605/86-05.pdf?sequence=1&isAllowed=y

Horner, S. B., Fireman, G. D., & Wang, E. W. (2010). The relation of student behavior, peer status, race, and gender to decisions about school discipline using chaid decision trees and regression modeling. *Journal of School Psychology, 48*(2), 135-161.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets weka. *Computational Statistics, 24*(2), 225-232.

Hussain, F., Liu, H., Suzuki, E., & Lu, H. (2000). Exception rule mining with a relative interestingness measure. *Knowledge Discovery and Data Mining. Current Issues and New Applications*, 86-97.

Iwatani, E., Stone, C. A., & Shealy, C. (forthcoming). How classification trees can be helpful to educational research: A case study on study abroad learning satisfaction.

Janssen, F., & Fürnkranz, J. (2009). *A re-evaluation of the over-searching phenomenon in inductive rule learning.* Paper presented at the SIAM International Conference on Data Mining (SDM-09), Sparks, NV.

Japkowicz, N. (2013). Assessment metrics for imbalanced learning. In H. He & Y. Ma (Eds.), *Imbalanced learning: Foundations, algorithms, and applications* (pp. 187-206). Hoboken, New Jersey: John Wiley & Sons, Inc.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education & Praeger.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society.Series C (Applied Statistics), 29*(2), 119-127.

Kayri, M., & Gunuc, S. (2010). An analysis of some variables affecting the internet dependency level of turkish adolescents by using decision tree methods. *Educational Sciences: Theory and Practice, 10*(4), 2487-2500.

Klösgen, W. (1996). Explora: A multipattern and multistrategy discovery assistant. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 249-271). Menlo Park, CA: AAAI Press.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the International Joint Conference on Artificial Intelligence.

Kopiez, R., Weihs, C., Ligges, U., & Lee, J. I. (2006). Classification of high and low achievers in a music sight-reading task. *Psychology of Music, 34*(1), 5-26.

Kotsiantis, S. (2011). Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review, 35*(3), 223-240.

Kotsiantis, S. (2014). Bagging and boosting variants for handling classifications problems: A survey. *The Knowledge Engineering Review, 29*(1), 78.

Kuchar, J. (2015). rCBA: CBA classifier for R. Retrieved from https://CRAN.R-project.org/package=rCBA

Kuhn, M., Weston, S., Coulter, N., Quinlan, R., & Culp., M. C. (2015). C50: C5.0 decision trees and rule-based models. Retrieved from https://CRAN.R-project.org/package=C50

Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*, 70.

Lang, S., & Baehr, C. (2012). Data mining: A hybrid methodology for complex and dynamic research. *College Composition and Communication, 64*(1), 172-194.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News, 2*(3), 18-22.

Lim, T., Loh, W., & Shih, Y. (2000). Comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning, 40*(3), 203-228.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14-16.

Liu, B., Hsu, W., & Chen, S. (1997). *Using general impressions to analyze discovered classification rules.* Paper presented at the KDD.

Liu, B., Hsu, W., Chen, S., & Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *Intelligent Systems and their Applications, IEEE, 15*(5), 47-55.

Liu, B., Hsu, W., & Ma, Y. (1998). *Integrating classification and association rule mining.* Paper presented at the The 4th International Conference on Knowledge Discovery and Data Mining (KDD-98), Menlo Park, CA.

Liu, B., Hsu, W., & Ma, Y. (1999). *Pruning and summarizing the discovered associations*. Paper presented at the Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA.

Liu, B., Ma, Y., & Wong, C.-K. (2000). *Improving an exhaustive search based rule learner.* Paper presented at the The 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000), Lyon, France.

Liu, X., & Ruiz, M. E. (2008). Using data mining to predict K-12 students' performance on large-scale assessment items related to energy. *Journal of Research in Science Teaching, 45*(5), 554-573.

Liu, X., & Whitford, M. (2011). Opportunities-to-learn at home: Profiles of students with and without reaching science proficiency. *Journal of Science Education and Technology, 20*(4), 375-387.

Liu, Y. C., & Hsu, Y. C. (2013). Predicting adolescent deviant behaviors through data mining techniques. *Educational Technology & Society, 16*(1), 295-308.

Loh, W., & Shih, Y. (1997). Split selection methods for classification trees. *Statistica sinica, 7*(4), 815-840.

Luan, J., & Zhao, C.-M. (2006). Practicing data mining for enrollment management and beyond. *New Directions for Institutional Research*(131), 117-122.

Lupton, D. (2013). *Digital sociology: Beyond the digital to the sociological.* Paper presented at the The Australian Sociological Association 2013 Conference, Melbourne.

Maimon, O., & Rokach, L. (2005). Introduction to supervised methods (pp. 149-164). Boston, MA: Springer US.

Manovich, L. (2012). Trending: The promises and challenges of big social data. In G. K. Matthew (Ed.), *Debates in the digital humanities* (pp. 460-475). Minneapolis, MN: University of Minnesota Press.

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence, 38*(3), 315-330.

Martin, T., & Sherin, B. (2013). Learning analytics and computational techniques for detecting and evaluating patterns in learning: An introduction to the special issue. *Journal of the Learning Sciences, 22*(4), 511-520.

Masunaga, H., & Lewis, T. (2011). Self-perceived dispositions that predict challenges during student teaching: A data mining analysis. *Issues in Teacher Education, 20*(1), 35-49.

McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review, 20*(1), 39-61.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-966.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

Michalski, R. S. (1969). *On the quasi-minimal solution of the general covering problem*. Paper presented at the Proceedings of the 5th International Symposium on Information Processing, Bled, Yugoslavia. http://www.mli.gmu.edu/papers/69-78/69-2.pdf

Michalski, R. S., & Kaufman, K. (2001). *The AQ19 system for machine learning and pattern discovery: A general description and user's guide*. Retrieved from Fairfax, VA: http://www.mli.gmu.edu/papers/2001-2002/AQ19ug.pdf

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). *Machine learning, neural and statistical classification*. New York: E. Horwood.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery, 2*(4), 345-389.

Mutter, S., Hall, M., & Frank, E. (2004). Using classification to evaluate the output of confidence-based association rule mining *Ai 2004: Advances in artificial intelligence* (pp. 538-549). Berlin: Springer.

Nystrom, N. A., Levine, M. J., Roskies, R. Z., & Scott, J. (2015). *Bridges: A uniquely flexible hpc resource for new communities and data analytics.* Paper presented at the Proceedings of the 2015 Annual Conference on Extreme Science and Engineering Discovery Environment (St. Louis, MO, July 26-30, 2015).

Ordonez, C. (2006, 2006). *Comparing association rules and decision trees for disease prediction.* Paper presented at the The International Workshop on Healthcare Information and Knowledge Management (HIKM '06).

Padmanabhan, B., & Tuzhilin, A. (1998). *A belief-driven method for discovering unexpected patterns.* Paper presented at the KDD.

Padmanabhan, B., & Tuzhilin, A. (1999). Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems, 27*(3), 303-318.

Padmanabhan, B., & Tuzhilin, A. (2000). *Small is beautiful: Discovering the minimal set of unexpected patterns.* Paper presented at the Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining.

Pai, P.-F., Lyu, Y.-J., & Wang, Y.-M. (2010). Analyzing academic achievement of junior high school students by an improved rough set model. *Computers & Education, 54*(4), 889-900.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society, 17*(4), 49-64.

Peters, A., & Hothorn, T. (2015). Ipred: Improved predictors. Retrieved from https://CRAN.R-project.org/package=ipred

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules *Knowledge discovery in databases* (pp. 229-248): AAAI/MIT Press.

Polikar, R. (2006). Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE, 6*(3), 21-45.

Powers, D. M. W. (n.d.). *What the f-measure doesn't measure: Features, flaws, fallacies and fixes*. Retrieved from http://arxiv.org/ftp/arxiv/papers/1503/1503.06410.pdf

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning, 42*(3), 203-231.

Provost, F., & Fawcett, T. (2013). *Data science for business : What you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O'Reilly Media.

Quinlan, J. R. (1990). Learning logical definitions from relations. *Machine Learning, 5*, 239-266.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, Calif: Morgan Kaufmann Publishers.

Quinlan, J. R. (2013). See5/c5.0 (Version 2.10): Rulequest Research. Retrieved from http://www.rulequest.com/see5-info.html

Quinlan, J. R., & Cameron-Jones, R. M. (1995). *Oversearching and layered search in empirical learning.* Paper presented at the 14th International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, QC.

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

Reimann, P., Markauskaite, L., & Bannert, M. (2014). E-research and learning theory: What do sequence and process mining methods contribute? *British Journal of Educational Technology, 45*(3), 528-540.

Rogers, E. M. (2003). *Diffusion of innovations* (Vol. 5th). New York: Free Press.

Rokach, L., & Maimon, O. (2005). Decision trees (pp. 165-192). Boston, MA: Springer US.

Rokach, L., & Maimon, O. (2014). *Data mining with decision trees : Theory and applications (2nd edition)*. Singapore: World Scientific Publishing Company.

RStudio Team. (2015). Rstudio: Integrated development for R.

Ruppert, E. (2012). The governmental topologies of database devices. *Theory, Culture & Society, 29*(4-5), 116-136.

Ruppert, E. (2013). Rethinking empirical social sciences. *Dialogues in Human Geography, 3*(3), 268-273.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery, 1*(3), 317-328.

SAS Institute, I. (1998) SAS institute white paper, data mining and the case for sampling: Solving business problems using SAS enterprise miner software. Carey, NC: SAS Institute.

Sasaki, Y. (2007). The truth of the f-measure. *Teach Tutor mater*, 1-5.

Savage, M. (2009). Contemporary sociology and the challenge of descriptive assemblage. *European Journal of Social Theory, 12*(1), 155-174.

Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and algorithms*: MIT press.

Schumacher, P., Olinsky, A., Quinn, J., & Smith, R. (2010). A comparison of logistic regression, neural networks, and classification trees predicting success of actuarial students. *Journal of Education for Business, 85*(5), 258-263.

Selwyn, N. (2015). Data entry: Towards the critical study of digital data and education. *Learning, Media and Technology, 40*(1), 64-82.

Shotwell, M. (2014). Sas7bdat: SAS database reader (experimental). Retrieved from https://CRAN.R-project.org/package=sas7bdat

Silberschatz, A., & Tuzhilin, A. (1995). *On subjective measures of interestingness in knowledge discovery.* Paper presented at the KDD.

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *Knowledge and Data Engineering, IEEE Transactions on, 8*(6), 970-974.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3-22.

Streifer, P. A., & Schumann, J. A. (2005). Using data mining to identify actionable information: Breaking new ground in data-driven decision making. *Journal of Education for Students Placed at Risk, 10*(3), 281-293.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323-348.

Suzuki, E., & Kodratoff, Y. (1998). Discovery of surprising exception rules based on intensity of implication. In J. M. Żytkow & M. Quafafou (Eds.), *Principles of data mining and knowledge discovery: Second european symposium, pkdd '98 nantes, france, september 23–26, 1998 proceedings* (pp. 10-18). Berlin, Heidelberg: Springer Berlin Heidelberg.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.

Tan, P.-N., Kumar, V., & Srivastava, J. (2002). *Selecting the right interestingness measure for association patterns.* Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.

Therneau, T., Atkinson, B., & Ripley, B. (2015). Rpart: Recursive partitioning and regression trees. Retrieved from https://CRAN.R-project.org/package=rpart

Thomas, L. L. (2006). *Pathways to success or failure: Factors affecting academic achievement among black students.* (3203974 Ph.D.), State University of New York at Buffalo, Ann Arbor. Retrieved from http://pitt.idm.oclc.org/login?url=http://search.proquest.com/docview/304938982?accountid=14709 ProQuest Dissertations & Theses Full Text; ProQuest Dissertations & Theses Global database.

Thuneberg, H., & Hotulainen, R. (2006). Contributions of data mining for psycho-educational research: What self-organizing maps tell us about the well-being of gifted learners. *High Ability Studies, 17*(1), 87-100.

Toivonen, H. (1996). *Sampling large databases for association rules.* Paper presented at the The 22nd VLDB Conference, Mumbai, India.

Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., & Mannila, H. (1995). *Pruning and grouping discovered association rules.*

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., . . . Wilkens-Diehr, N. (2014). Xsede: Accelerating scientific discovery. *Computing in Science & Engineering, 16*(5), 62-74.

Tufféry, S. (2011). *Data mining and statistics for decision making.* Chichester, West Sussex: Wiley.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics. 33* (1), 1-67. doi:10.1214/aoms/1177704711.

Tung, A. K. H. (2009). Rule-based classification. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 2459-2462): Springer US.

United States Department of Education National Center for Education Statistics. (1995). *National Education Longitudinal Study, 1988: Second follow-up (1992)*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research (ICPSR) [distributor]. Retrieved from: http://doi.org/10.3886/ICPSR06448.v1

United States Department of Education National Center for Education Statistics. (1999). *National Education Longitudinal Study, 1988: First follow-up (1990)*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research (ICPSR) [distributor]. Retrieved from: http://doi.org/10.3886/ICPSR09859.v1

United States Department of Education National Center for Education Statistics. (2006a). *National Education Longitudinal Study, 1988*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research (ICPSR) [distributor]. Retrieved from: http://doi.org/10.3886/ICPSR09389.v1

United States Department of Education National Center for Education Statistics. (2006b). *National Education Longitudinal Study: Base year through fourth follow-up, 1988-2000*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research (ICPSR) [distributor]. Retrieved from: http://doi.org/10.3886/ICPSR03955.v1

University of Pittsburgh University Library System. (2015, Jul 13, 2015). Education databases. Retrieved from http://pitt.libguides.com/educationdatabases

Valsamidis, S., Kontogiannis, S., Kazanidis, I., Theodosiou, T., & Karakos, A. (2012). A clustering methodology of web log data for learning management systems. *Educational Technology & Society*, *15*(2), 154-167.

Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting academic performance by data mining methods. *Education Economics, 15*(4), 405-419.

Vaughn, B. K., & Wang, Q. (2008). Classification based on tree-structured allocation rules. *The Journal of Experimental Education, 76*(3), 315-340.

Webb, G. I. (1995). OPUS: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research*, 431-465.

Weerts, D. J., & Ronca, J. M. (2009). Using classification trees to predict alumni giving for higher education. *Education Economics, 17*(1), 95-122.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software, 40*(1), 1-29.

Wickham, H., & Francois, R. (2016). Dplyr: A grammar of data manipulation. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Miller, E. (2016). Haven: Import and export 'SPSS', 'stata' and 'SAS' files. Retrieved from https://CRAN.R-project.org/package=haven

Willett, T., & Hom, W. (2007). Student flow analysis for a community college. *Journal of Applied Research in the Community College, 15*(1), 17-27.

Williams, G. J. (2011). *Data mining with rattle and R: The art of excavating data for knowledge discovery*. New York: Springer

Wing, J., Kuhn, M., Weston, S., Williams, A., Keefer, C., Engelhardt, A., . . . Hunt., T. (2016). Caret: Classification and regression training. Retrieved from https://CRAN.R-project.org/package=caret

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining : Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.

Wolpert, D. H. (2012). *What the no free lunch theorems really mean: How to improve search algorithms*. Retrieved from http://www.santafe.edu/media/workingpapers/12-10-017.pdf

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67-82.

Yao, Y. Y., & Zhong, N. (1999). An analysis of quantitative measures associated with rules *Methodologies for knowledge discovery and data mining* (pp. 479-488): Springer.

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering, 12*(3), 372-390.

Zaman, M. F., & Hirose, H. (2011). Classification performance of bagging and boosting type ensemble methods with small training sets. *New Generation Computing, 29*(3), 277-292.

Zhao, C.-M., & Luan, J. (2006). Data mining: Going beyond traditional statistics. *New Directions for Institutional Research*(131), 7-16.