REGRESSION METHODS FOR CATEGORICAL DEPENDENT VARIABLES:

EFFECTS ON A MODEL OF STUDENT COLLEGE CHOICE

Kelly E. Rapp

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Counseling and Educational Psychology,
Indiana University
March, 2012

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Ginette Delandshere, Ph.D.

Jonathan Plucker, Ph.D.

Chao-Ying Joanne Peng, Ph.D.

Robert Toutkoushian, Ph.D.

January 30, 2012

ACKNOWLEDGMENTS

To my dissertation committee: thank you for your patience as I struggled to balance the pursuit of my degree with my personal and family commitments. I appreciate you all sticking with me through time and distance. Thank you, Dr. Delandshere, for your perspective on the measurement and inquiry-based matters of my dissertation and for agreeing to chair my research committee. Dr. Peng, your expertise in logistical regression techniques helped immensely to guide my analysis. Dr. Toutkoushian, your dedication to the practical and applied aspects of my work has proven invaluable. It was in your higher education research methods (econometrics) course that I came up with the idea for my dissertation research. I have appreciated your continued feedback, even after you left the university.

To Jonathan Plucker, my professional mentor: without your support, I would not be graduating with my Ph.D. You gave me a great opportunity when you hired me as a research associate in 2003 and shortly after encouraged me to pursue my doctorate. At times I was reluctant to continue and even took some time off, but you never gave up on me. Your persistence on my behalf gave me the strength to finish what I started.

I would also like to thank Dr. Suzanne Eckes, associate professor in Educational Leadership and Policy Studies, who referred me to Dr. Plucker when she only knew me as the wife of one of her research assistants. My husband and I both are grateful for Suzanne – she has been a colleague, friend, and "cheerleader" throughout the years.

To my good friend Kelli Paul, 2006 graduate of the educational psychology inquiry program: I relied on you tremendously to help me navigate the logistical and primarily the psychological aspects of being a graduate student. Thank you for lending an ear time and time again on all matters degree-related.

Thank you to Deborah Harris and Peggy Loveless at ACT®, Inc., for your help in acquiring the data for my dissertation. Thank you Deborah for granting me permission to use the data, and thank you Peggy for your extensive legwork in compiling the database (a couple of times). I am grateful to you and the others at ACT® as well for providing me the opportunity to intern with you in 2006.

To Linda Caldwell, chair of the education and psychology departments at Ozarks Technical Community College in Springfield, MO: thank you for encouraging me to finish my degree and for providing me the flexibility in teaching schedule which made it possible.

Thank you Rob Weislocher and Brian King, proprietors of the Mudhouse, the Springfield, MO, coffee house at which the entirety of this dissertation was composed. I am grateful to you for envisioning and operating a comfortable space where ideas can be generated and examined (and where delicious coffee, sandwiches, and pastries can be consumed!).

Finally, I must express my gratitude to my family. Without you, I would not have been able to complete my degree. Thank you to my husband, Kevin, for always supporting me (in fact, pushing me at times) in this endeavor no matter the sacrifice on your part. You have been patient with me as I worked through the requirements of the degree, even when parenting responsibilities added to our time constraints. Most importantly, you have always believed in and trusted me. My sister Kimberly has offered moral support and friendship through the years, and my father, Allan Thompson, has been a great listener when the process got me down – he also kindly volunteered to proofread my entire dissertation!

For my mother, Betty Thompson, I am eternally grateful. She has always been my greatest supporter, as only a mother can be. She has never doubted my ability to earn a doctorate (even years before I started!). In order to make this possible, she gave her time and energy to my children and household when I needed to work uninterrupted. Words cannot express the extent or the degree of help my mother has provided as I completed this degree (not to mention all other accomplishments in my life).

To my beautiful children, Bennett and Avery: thank you for giving me inspiration and a reason to finish my dissertation. I look forward to spending more time with you!

Kelly E. Rapp

REGRESSION METHODS FOR CATEGORICAL DEPENDENT VARIABLES:

EFFECTS ON A MODEL OF STUDENT COLLEGE CHOICE

The use of categorical dependent variables with the classical linear regression model (CLRM) violates many of the model's assumptions and may result in biased estimates (Long, 1997; O'Connell, Goldstein, Rogers, & Peng, 2008). Many dependent variables of interest to educational researchers (e.g., professorial rank, educational attainment) are categorical in nature but are analyzed using the CLRM (Harwell & Gatti, 2001) even though alternate regression techniques for categorical dependent variables are recommended (Agresti, 1996; Long, 1997). Data obtained from ACT®, Inc., on 5,200 high school seniors in Illinois and Colorado were used to analyze effects of regression method on a model of ascriptive and academic influences on selectivity of postsecondary institution attended. The dependent variable was measured in rank-ordered categories based on self-reported institutional admissions policies and analyzed with classical linear, multinomial logistic, and ordered logistic regressions. Choice of regression method did not affect overall model performance as evidenced by significant F and Likelihood Ratio χ^2 tests. The full CLRM was fit moderately-well to the data (R^2 = .391), surpassing some previous findings (Hearn, 1988, 1991; Davies & Guppy, 1997). McFadden's R_L^2 measure of strength of association was larger in the multinomial regression than in the ordered regression (R_L^2 = .191 vs. R_L^2 = .158). The multinomial logistic method also correctly predicted dependent variable category with the greatest accuracy (46.3% correct), but Somers' D_{yx} measure of association was smallest for the multinomial model. Direction and significance of relationship between predictors and the dependent variable was

substantively consistent across the CLRM and logistic methods. In all regressions, ACT® score had the most impact on selectivity of institution attended. Threshold values were significant, supporting the assumption of an ordered dependent variable. Due to the CLRM's theoretical and predictive shortcomings and the multinomial model's complexity in interpretation, ordered logistic regression was determined to be the most appropriate for explaining influences on selectivity of postsecondary institution attended.

Table of Contents

CHAPTER 1 – INTRODUCTION 1
Analysis of a Categorical Dependent Variable: Student College Choice Research as an Illustrative Case.. 2
CHAPTER 2 – LITERATURE REVIEW 5
Levels of Measurement 5
Regression Models 14
The General Linear Model 14
Regression for Categorical Dependent Variables 21
Student College Choice 30
Theoretical Frameworks for Study 39
Influences on College Attendance 43
Influences on Attendance at Selective Colleges 54
Methodologies 59
CHAPTER 3 – METHODOLOGY 68
Data Sources 68
Sample 70
Variables 75
Data Screening 91
Analysis 113
CHAPTER 4 – RESULTS 130
Assumptions 130
Classical Linear Regression 144
Multinomial Logistic Regression 152
Ordered Logistic Regression 167
Effects of Model Specification 179
Final Model 191
CHAPTER 5 – DISCUSSION 192
Model Comparisons 192
Conclusions 207
Student College Choice Findings 213
Limitations 218
REFERENCES 220

List of Tables

Table 1. Properties of Regression Models 31
Table 2. Sample and Complete-Data Population Description........................ 74
Table 3. References for Predictor Variables 90
Table 4. Missing Data Percentages 91
Table 5. Continuous Variable Means for Missing and Non-Missing SPS Data... 97
Table 6. Categorical Variable Percentages for Missing and
Non-Missing SPS Data 98
Table 7. Continuous Variable Means for Missing and
Non-Missing Income Data 101
Table 8. Categorical Variable Percentages for Missing and
Non-Missing Income Data 101
Table 9. Continuous Variable Means for Missing and
Non-Missing Race Data 103
Table 10. Categorical Variable Percentages for Missing and
Non-Missing Race Data 103
Table 11. Continuous Variable Means for Missing and
Non-Missing High School Transcript Data 105
Table 12. Categorical Variable Percentages for Missing and
Non-Missing High School Transcript Data 106
Table 13. Description of Cases with Complete Data on
All Variables and of Cases Excluded from Analysis 108
Table 14. Assumptions Required for Three Regression Models 113
Table 15. Definitions of Variables 123
Table 16. Classical Linear Regression Results: Equation C1 145
Table 17. Classical Linear Regression Results: Equation C2 146
Table 18. Classical Linear Regression Results: Equation C3 147
Table 19: Summary of Classical Linear Regression Models 150
Table 20. Behavior of Predictors across Classical Linear Regression Models..... 151
Table 21. Multinomial Logistic Regression Results: Equation M1 154
Table 22. Multinomial Logistic Regression Results: Equation M2 157
Table 23. Multinomial Logistic Regression Results: Equation M3 161
Table 24: Summary of Multinomial Logistic Regression Models 165
Table 25. Behavior of Predictors across Multinomial Logistic Regression
Models for Selective Admissions Policy 166
Table 26. Ordered Logistic Regression Results: Equation O1 168
Table 27. Ordered Logistic Regression Results: Equation O2 171
Table 28. Ordered Logistic Regression Results: Equation O3 173
Table 29: Summary of Ordered Logistic Regression Models 177
Table 30. Behavior of Predictors across Ordered Logistic Regression Models.... 178
Table 31. Model Selection for Describing Influences on
Selectivity of College Attended 180
Table 32. Model Significance and Effect Size for Classical Linear,
Multinomial Logistic, and Ordered Logistic Regression Models 195

Table 33. Classification Tables for Classical Linear, Multinomial Logistic, and Ordered Logistic Regression Models 196
Table 34. Predictive Accuracy of Models by Admissions Policy Category 199
Table 35. Serial Position of Admissions Policy Frequency for Observed and Predicted Data 200
Table 36. Behavior of Predictors across Classical Linear, Multinomial Logistic, and Ordered Logistic Regression Models 204

Chapter 1: Introduction

Ordinal scales are commonly used in the behavioral and social sciences and in educational research (Agresti, 1996; Harwell & Gatti, 2001; Kirk, 1999; Long, 1997). In fact, as opposed to being continuous and interval level, many response variables in education have discrete categories (O'Connell, Goldstein, Rogers, & Peng, 2008) which are often ordered (Long, 1997; Zumbo & Ochieng, 2002). Specific to educational research, professorial rank, level of educational attainment, and ordered categorical responses obtained from questionnaires or surveys are examples of ranked categories that are often used (Kirk, 1999; Long, 1997; Zumbo & Ochieng, 2002).

Another example is the commonly-used variable for measuring attitudes in survey research, the Likert scale. This scale asks participants to indicate whether they strongly agree, agree, have no opinion, disagree, or strongly disagree with a statement (Long, 1997). Harwell and Gatti (2001) examined the prevalence of ordinal-scaled dependent variables in educational research by studying three prominent educational research journals. The researchers defined an ordinal scale as one that did not have equal differences between numbers but could be rank ordered. In 1997, 73% of the articles published in the journals in Harwell and Gatti's (2001) study used dependent variables that were measured on an ordinal scale – specifically, all of the ordered dependent variables found were Likert scales.

Researchers often treat such ordinal dependent variables as if they were interval level and use the parametric statistical procedure of classical linear regression (Harwell & Gatti, 2001; Long, 1997). Harwell and Gatti (2001), in their analysis of educational research studies published in 1997, found that nearly all studies used statistical analyses usually described as requiring interval-scaled data even though the majority of the dependent variables were measured on an ordinal scale. However, treating a scale as interval level involves the assumption that the distances between categories, such as between agreeing and strongly agreeing and between agreeing and having no opinion, is the same (Long, 1997), but this contradicts Stevens' (1946) definition of the ordinal scale of measurement. In fact, many of the assumptions of the classical linear regression model (CLRM) are violated for ordinal outcomes (O'Connell et al., 2008). The use of the CLRM with discrete ordered dependent variables may introduce bias and inefficiency and provide misleading results (Long, 1997). Therefore, alternate regression techniques especially designed to analyze categorical dependent variables have been developed and are recommended (Agresti, 1996; Long, 1997).

Analysis of a Categorical Dependent Variable: Student College Choice Research as an Illustrative Case

The current study uses the field of student college choice research as a backdrop for investigating the impacts of different statistical methods when the dependent variable is categorical in nature. Studies of student college choice primarily define the dependent variable of college attendance either as dichotomous, multi-category, or continuous and use logistic, multinomial, and classical linear regression models to analyze data. Data exist on the selectivity, or prestige, of postsecondary institutions, although this designation may be defined in various ways. Regardless, there is a presumed ordering of selectivity level among colleges, making this an appropriate dependent variable for an examination of alternate regression techniques for categorical and continuous variables. Furthermore, the economic and social benefits of attending a highly selective college are great (Ehrenberg, 2003; Hearn, 1991), and accurate information on the factors that influence attendance at such institutions may help students maximize their chances of attendance.

The present study adds to the existing research on student college choice by providing a methodological focus. Previous studies of the selectivity-level of institution have primarily conceptualized this dependent variable as dichotomous (Alexander et al., 1987; Hu & Hossler, 1998; Kaufman & Gabler, 2004; Owings et al., 1998) or as approximately continuous (average SAT score of incoming freshman; Davies & Guppy, 1997; Hearn, 1991; McDonough & Antonio, 1996). One study of influences on selectivity of college attended paired an ordered, discrete selectivity dependent variable with a recommended ordered regression technique (Hilmer, 2001). In Hilmer's study, multinomial logistic regression, ordered probit regression, and bivariate probit with sample selection were compared; however, Hilmer (2001) did not examine the use of classical linear regression with his dependent variable. Therefore, an analysis of possible differential results from the use of the classical linear regression model and regression techniques designed for categorical dependent variables is needed.

Using student- and institution-level data from ACT®, Inc., and a dependent variable representing five categories of selectivity of college (ranging from non-college attendance to attendance at a highly selective institution), the present study compares three statistical methods on a sample of 5,200 high school seniors in Illinois and Colorado: classical linear regression with ordinary least squares estimation and multinomial and ordered logistic regression with maximum likelihood estimation. Specifically, results of overall model significance tests (analysis of variance *F*-tests and likelihood ratio χ^2 tests) and the Pearson χ^2 and the deviance-based inferential tests of goodness-of-fit are assessed for each regression. Effect sizes (R^2), measures of association (R_L^2, Somers' D_{yx}), and predictive ability are also used as bases for comparison. Behavior of predictor variables in terms of significance, direction, and standardized impact is also examined for any differences across method and model. Evaluations of regression assumptions are presented, and diagnostic analyses are also performed. Finally, practical implications of each method for understanding the student college choice process are discussed.

Because of different modeling procedures used in the regressions, direct comparisons across method are not possible on the above-mentioned criterion recommended for selecting a superior model (Peng & So, 2002). However, model performance is analyzed in terms of these factors in order to provide a methodological recommendation for student college choice studies of selectivity of college attended.

Chapter 2: Literature Review

The concern over the pairing of appropriate statistical methods with categorical dependent variables which frames the current study is rooted in a long-standing debate over levels of measurement, sometimes referred to as the measurement-statistics debate (Zumbo & Zimmerman, 1993). To understand the methodological issues surrounding regression models for categorical dependent variables such as the ordered variable used in the current study, a discussion of levels of measurement is needed and provided below. Following from this discussion, regression models for continuous and categorical dependent variables are then explored.

In addition to a review of methodological considerations, a review of the literature on student college choice is beneficial for understanding the context of the current study. Specifically, models of student college choice are presented, followed by an examination of existing research on the college choice process.

Methodological Considerations

Levels of Measurement

Stevens' (1946) seminal paper, "On the Theory of Scales of Measurement," attempted to make explicit the rules for the assignment of numbers to objects or events, the mathematical properties of the measurement scales, and the statistical operations appropriate for each type of scale. Stevens (1946) defined four scales of measurement which are widely used by researchers to this date (Harwell & Gatti, 2001): nominal, ordinal, interval, and ratio. In the nominal scale, numbers are only used as labels, and their assignment is arbitrary (Stevens, 1946). The relevant rule for assigning numbers within a nominal scale is that of distinctness: the same number must not be assigned to different classes (Kirk, 1999; Stevens, 1946).

The ordinal measurement scale contains more information than the nominal scale in that number assignments are both distinct and rank-ordered (Kirk, 1999). Importantly, the assignment of numbers by the ordinal rule of measurement only indicates relative rank and not magnitude between classes. Any transformation of an ordinal scale that leaves the original ordering intact, or any monotonic transformation such as addition, subtraction, multiplication, or log transformation, is permissible. However, statistics such as means and standard deviations should not be used with ordinal scales, according to Stevens, as they imply a knowledge of more than simply rank order (such as equal intervals between ranks; Kirk, 1999; Stevens, 1946).

Moving from qualitative, or categorical, variables commonly measured with nominal and ordinal scales, quantitative variables can be measured on interval or ratio scales. The important distinction between these two types of scales is the absence or presence of a true, or absolute, zero. Interval scales assume equal distances between numbers; therefore, more meaningful transformations and statistical procedures can be used. Linear transformations are acceptable because they preserve the relative differences; however, due to the fact that there is no value for the absence of a characteristic, it is not acceptable to say that some value is twice as great as another, for example. These kinds of comparisons can only be made on a ratio scale, in which an absolute zero is always implied (Kirk, 1999; Stevens, 1946).

Stevens' (1946) work has been influential, but his coupling of measurement scales and statistical procedures has not been widely accepted without debate (Harwell & Gatti, 2001). Stevens' (1946) definitions of the rules for measurement scales are viewed as a benefit for measurement theory (Gaito, 1980), appearing in introductory statistical texts and journal articles (Harwell & Gatti, 2001). Furthermore, his prescriptions for data analysis techniques for the measurement scales have been supported by many researchers, specifically for use in descriptive statistics (Harwell & Gatti, 2001; Marcus-Roberts & Roberts, 1987). However, some opponents of Stevens' strict pairing of permissible statistical procedures with levels of measurement argue that his prescriptions are unrealistic (Harwell & Gatti, 2001), sparking one of the longest-standing debates in behavioral science methodology (Zumbo & Zimmerman, 1993).

Measurement-Statistics Debate

The controversy over Stevens' (1946) position centers on the nature of the relationship between ordinal and interval scales and the associated use with nonparametric and parametric statistics, respectively (Gardner, 1975; Zumbo & Zimmerman, 1993). Essentially, Stevens (1946) stated that ordinal measurements place objects in a rank order but that no meaning can be given to the size of the interval between objects or to the shape of the frequency distribution of the measurements. On the other hand, interval scales have equal distributions between units across the entire scale (Gardner, 1975). This distinction led to the prescription of permissible mathematical transformations – monotonic, nonlinear transformations for ordinal scales and linear transformations for interval scales – such that the original scales were left invariant. For Stevens, this logically led to the use of different statistical procedures with each scale. Nonparametric statistics do not make assumptions about equivalence of units or shape of frequency distribution in the population and were therefore deemed appropriate for use with ordinal-level data, and parametric statistics, which assume that the population from which samples are drawn is normally distributed, were therefore appropriate for use with interval scales (Gardner, 1975). Generally, permissible statistics for ordinal scales included mode, median, percentiles, and ordinal correlations, and interval data allowed for means, standard deviations, and product-moment correlations and all related tests of inference (Velleman & Wilkinson, 1993).

Although initially accepted by some researchers and textbooks (most notably Siegel, 1956, and Senders, 1958, and later Sharp, 1979, and Blalock, 1979, as cited in Gaito, 1980), controversy over Stevens' ideas sprang up almost immediately. Gaito (1960) describes the debate as being generated by a confusion between measurement and statistical theories, and Gardner (1975) similarly classifies the debate as belonging to two distinct domains, psychometrics and statistics. Indeed, there is disagreement about whether a distinction between ordinal and interval scales can easily be made (Gaito, 1960; Gardner, 1975; Kirk, 1999). In the matter of statistical theory, the question is over the necessity of meeting all the requirements for the use of parametric statistics prescribed by Stevens (1946) and Siegel (1956).

The two strands of debate are actually somewhat linked: the prescription of specific statistical procedures for use only with certain measurement scales fails to recognize that measurement of many variables in the behavioral and social sciences lies somewhere between the ordinal and interval levels (Gaito, 1960; Kirk, 1999; Long, 1997). It is typical that more than just rank-order is known about data, such as the distance between at least some of the objects (Gardner, 1975). Gardner (1975) uses the term "summated scales" to describe the large number of instruments that fall in the ambiguous region between ordinal- and interval-level data. These scales are not strictly interval strength, but they are only mildly distorted interval scales. Achievement tests in which the total score is the sum of a set of nominal (right/wrong) measures and summated attitude scales in which the total score is a sum of a set of ordinal ratings are examples (Gardner, 1975).

This summated scale category includes a large portion of instruments used in educational and psychological research. A common example is the IQ (intelligence quotient) scale (Kirk, 1999; Labovitz, 1967). The difference between an IQ of 100 and 110 is agreed to represent a slightly smaller intellectual difference than the same 10-point difference between 130 and 140. The 10-point intervals are not identical across the scale, as is required by interval-level measurement, but they are believed to be very similar, yielding more information than ordinal-level measurement (Kirk, 1999). Stevens (1946), although not explicitly discussing an ambiguous area between the two levels of measurement, did acknowledge that most psychological measurement aspires to create interval scales but runs into problems when trying to equalize the units of scale, and he used intelligence as an example of an ordinal scale which tries to approximate an interval scale.

Labovitz (1967) provides a good example of an ambiguously-classified scale from the field of educational research. He summarizes a previous argument that number of years of formal education, while seemingly an interval scale with equal unit differences, is actually an ordinal scale because quality of education may be different among individuals within those levels of education. Labovitz (1967) concludes that although this argument has some merit, researchers cannot conclude that because a scale is not interval it must be ordinal; rather, some scales necessarily lie in between. Similarly, Kirk (1999) illustrates that although differences between rankings on an attitude scale are likely not identical, they are probably similar and therefore do contain some information about magnitude of differences above what an ordinal scale would assume. Partially because of this uncertainty in the ability to unambiguously categorize data into one level of measurement or another, researchers suggest that Stevens' (1946) position that only nonparametric statistics should be used for ordinal data should be looked at merely as a guide (Gaito, 1960; Kirk, 1999).

When data are treated as ordinal rather than interval, any known information regarding the differences between objects is compromised. Labovitz (1967) argues that some knowledge of at least an approximation to equal distances between scores is more useful than merely rank-ordered information. If scales that fall just short of interval level are treated as interval scales, much more powerful and interpretable statistics may be used. Especially in the social sciences, where measurement techniques are crude, the benefits of assuming interval-level data offset the disadvantages of not knowing for sure that the differences between scores is equal, Labovitz (1967) argues. In fact, logical and empirical arguments made since Stevens' 1946 work largely support this position (Gardner, 1975; Marcus-Roberts & Roberts, 1987; Velleman & Wilkinson, 1993; Zumbo & Zimmerman, 1993).

Gaito (1960) points out that the requirement of an interval scale in order to perform parametric tests cannot be found in the mathematical assumptions underlying, for example, the analysis of variance procedure. Anderson (1961) argues against Stevens' assertion of invariance of results under a permissible transformation of scale as the necessary criterion for appropriateness of statistics and concludes that psychological meaning is not a statistical matter and that type of scale has no relevance to the use of parametric or nonparametric procedures. Furthermore, numerous empirical studies have demonstrated that altering the metric properties of scales does not significantly alter the conclusions reached (Gardner, 1975; Labovitz, 1967; Zumbo & Zimmerman, 1993). Specifically, the validity of Siegel's (1956) assumption that variables must be measured on at least an interval scale in order to perform parametric tests has been successfully challenged such that most researchers argue that scale type is irrelevant to choice of nonparametric or parametric statistics (Gardner, 1975).

Kirk (1999) suggests that rather than find the solution from the work of mathematicians and statisticians, the choice of appropriate statistics must be made by the researchers themselves – those most acquainted with the use of numerical answers for meaningful application to real-world problems within a discipline. Additionally, Long (1997) notes the importance of theoretical context and substantive purpose of the analysis when determining the level of measurement of a variable. A review of literature by Kirk (1999) revealed that, in fact, experts in the behavioral sciences and education do use parametric statistics even though the data typically fall somewhere between ordinal and interval level, and these experts interpret the results to maximize the utility of whatever magnitude information they have about the numbers. Even Stevens (1946) conceded that the use of means and standard deviations with ordinal scales has pragmatic applications in that it often yields useful results. His later work reflected an emphasis on the analysis of the degree to which an incorrect statistic would impact the conclusions drawn from the research rather than a strict adherence to the pairing of one type of statistic to a level of measurement (Gardner, 1975). Lord's (1953) quote at the outset of the controversy can be used to summarize the debate as well: "The numbers do not know where they came from" (p. 751).

Therefore, for the purposes of choice of nonparametric or parametric statistical procedures, it does not seem to matter much whether data are classified as interval or ordinal. Researchers agree that the benefits of using parametric statistics outweigh the costs, and decades of debate have effectively resolved this issue to determine that, in fact, use of parametric statistics with ordinal-level data is acceptable[1] (Borgatta & Bohrnstedt, 1980; Gaito, 1980; Gardner, 1975; Labovitz, 1967; Marcus-Roberts & Roberts, 1987; Velleman & Wilkinson, 1993; Zumbo & Zimmerman, 1993). However, even within the

[1] An interesting strand of critique spanning nearly two decades, however, is the work of Joel Michell. Michell argues that psychometrics is a pathological science in that it has accepted the hypothesis that psychological attributes are quantitative without any serious attempts to test this hypothesis or even to recognize that its blind acceptance is a problem. Key in promoting this pathology, according to Michell, is Stevens' ubiquitous definition of measurement. Michell contends that psychometricians' desire to construct procedures that can be labeled as measurement outweighs the desire to uncover the true properties of an attribute. Psychometricians have adopted an operationist interpretation of Stevens' work through which the issue of scale type has centered on admissible transformations rather than a consideration of the quantitative structure of an attribute. Further, Michell argues that the determination of psychological scales such as test scores as interval-level is made to justify the widespread use of linear transformation with such scales and to avoid raising the hypothesis that psychological attributes are quantitative (see generally Michell, 2008a, 2008b). Based on Michell's arguments, the resolution of the debate between the use of parametric or nonparametric statistics with ordinal data might be considered meaningless because it occurred within a pathological science which has not examined underlying, relevant assumptions about quantitative attributes. However, because in practice most researchers do use parametric statistics with ordinal or quasi-interval data, I will continue to operate under this premise for the purposes of this study.

domain of parametric statistics, there is great variation among available procedures, and again the distinction between ordinal and interval data becomes important.

Parametric Statistics for Ordinal Scales

The level of measurement of a variable is often not unambiguous, and researchers must consider the context in which it is to be used. However, once the level of the dependent variable is determined, it is important to match the statistical model employed to the level of measurement (Long, 1997). As noted, researchers in the behavioral and social sciences will likely choose a parametric procedure for ordinal- or interval- level data, but many researchers treat ordered dependent variables as if they were measured on an interval scale by numbering the categories sequentially and using the linear regression model (Harwell & Gatti, 2001; Long, 1997).

However, in the instance where serious inequality of units along an ordinal scale is suspected, data transformation is recommended to make the distribution nearly symmetric (Gardner, 1975). Additionally, transformations can make variability consistent across groups and can make relationships more linear. Common transformation techniques include logarithmic transformation and powers and roots. These are monotonic but nonlinear and are therefore appropriate for use with ordinal scales (Gardner, 1975; Velleman &Wilkinson, 1987). Data transformations are the basis for the use of parametric regression procedures with non-linear, non-interval scale data and will be discussed further in the next section.

Depending on the types of problems to be addressed and the types of data encountered, each discipline of research has its own statistical idiosyncrasies and therefore a preferred set of techniques for analysis (Anderson, 1961; R. Toutkoushian, personal communication, Fall 2005). Econometrics is the application of mathematical statistics and the tools of statistical inference to the quantitative relationships evident in economic theory (Greene, 2000). Because of the nature of economic relationships and the lack of controlled experiments, the standard assumptions of the classical linear regression model are seldom met, and econometrics attempts to circumvent problems caused by violations of these assumptions (Kennedy, 1998). Econometric techniques are parametric in nature and are regression-based (R. Toutkoushian, personal communication, Fall 2005), although most models are nonlinear (Long, 1997). Many econometric techniques are used in fields other than economics, such as in education research where dependent variables of interest may be categorical or restricted in nature.

Regression Models

The current study, from the field of education research, makes use of econometric techniques due to the ordered, categorical nature of the dependent variable of college selectivity. Following an overview of the general regression model, a description of the parametric regression methods recommended for categorical dependent variables is given.

The General Linear Model

A statistical model controls for the effects of confounding variables and can incorporate several explanatory variables at once (Agresti, 1996). In non-experimental studies, when researchers do not have enough information on variables for which there are inadequate controls, statistical techniques such as regression can filter out their effects (Becker, 1983). All models are special cases of the general (or generalized) linear model (GLM), including ordinary regression for continuous dependent variables and models for categorical dependent variables. GLMs have three components: the random component, the systematic (or linear) component, and the link. The random component, or dependent variable, assumes a probability distribution; the systematic component specifies explanatory variables used as predictors; and the link describes the functional relationship between the systematic component and the expected value of the random component. The mean of the dependent variable can be modeled directly or through a monotonic function of the mean, and the GLM relates this function of the mean to the explanatory variables through a linear prediction equation. In ordinary regression with continuous data, the link function is simply the mean, but other links permit the mean to be nonlinearly related to the predictors. For example, the logit link function is the log of an odds ratio and is useful when the mean is between zero and one (Agresti, 1996; O'Connell et al., 2008).

Each probability distribution for the random component has one special function of the mean, called the natural parameter or canonical link. If the random component is assumed to have a normal distribution, the canonical link is the mean; if the random component is assumed to be binomial, the canonical link is the logit of success probability. The GLM allows a random component to have a probability distribution other than the normal and a function other than the mean, which is useful for modeling categorical response data. Additionally, in GLMs the choice of link is separate from the choice of random component distribution (Agresti, 1996).

Estimators. Statistical modeling involves estimating parameter values. Before data are gathered, or observed, the parameter value is unknown (Agresti, 1996). This parameter value for the observed data can be estimated, resulting in an "estimate," but it can never be known for sure. Therefore, the estimate cannot be defended with certainty, but the process for finding the estimate – the estimation method – can be justified (Kennedy, 1998). The formula, or recipe, by which the observed data are transformed into an estimate is called the "estimator," and the method for generating a good estimator varies depending on the situation. An estimator does not have the same characteristics in all estimating situations and therefore may be preferred in some situations and not in others. According to Kennedy, much of the study of econometrics revolves around how to generate a preferred estimator in any given situation. Therefore, researchers must evaluate the characteristics of an estimator in a situation according to some standards, or criteria (Kennedy, 1998).

One such criterion is that of least squares. The estimated values of the dependent variable can be calculated from observed data and subtracted from the actual values to get residuals. A good estimator should make these residuals small. Some researchers believe all residuals should be weighted equally and advocate choosing the estimator which minimizes the sum of absolute values of residuals. Other researchers believe large residuals should be avoided and argue for heavily weighting large residuals by choosing an estimator that minimizes the sum of squared residual values. The latter method is referred to as ordinary least squares (OLS) and is the most popular among researchers doing empirical work (Kennedy, 1998).

The OLS estimator will always minimize the sum of squared residuals, by definition, but in certain estimating situations it might not possess other properties researchers deem important such as unbiasedness or efficiency. Researchers typically check the OLS estimator first and make an overall judgment about whether it meets the criteria deemed important by the researcher. The best unbiased criterion is usually given the most importance, but when all unbiased estimators have variances that are too large, the mean square error criteria may be used or asymptotic properties evaluated (Kennedy, 1998).

A common alternative to the OLS estimator is the maximum likelihood (ML) estimator. The ML estimator of the parameter is the pair of mean and variance values that creates the greatest probability of obtaining the observed data (Kennedy, 1998). These values are found by substituting observed sample data into the probability function and viewing it as a function of the unknown parameter value (a likelihood function). The outcome and the sample size, or number of trials, is known, but the population parameter is unknown and must be estimated from sample information (Long, 1997). For a binomial distribution, for example, this parameter value would be the probability of obtaining observed data, and it would range from zero to one. Different probability values, or parameters, are chosen and plugged into this likelihood function. The parameter value for which the probability of getting the observed data takes its greatest value is the ML estimate of the parameter in the original regression equation (Agresti, 1996). In other words, the ML estimate is the value of the parameter that makes the observed data most likely (Long, 1997).

The ML estimator has several desirable asymptotic, or large sample, properties (Agresti, 1996; Kennedy, 1998). ML estimators are the most precise, meaning they have the smallest standard errors, in large samples (Agresti, 1996). They are also asymptotically unbiased, consistent, and normally distributed (Kennedy, 1998). In order to calculate the ML estimator, a specific distribution for the error term must be assumed. The ML estimator has a high computational cost in that it requires several algebraic manipulations, or iterations, but most software packages incorporate these, and the fast speed of modern computers virtually eliminates the problem of computational cost (Kennedy, 1998).

Classical Linear Regression Model. In choosing an appropriate, preferred estimator, the mechanism generating the observations must be known (Kennedy, 1998). An estimator's sampling distribution characteristics are the estimator criteria such as unbiasedness and efficiency, taken from repeated samples. Because an estimator does not have the same sampling distribution characteristics for all ways in which data are generated and will perform poorly in some situations, it is important to know the estimating situation for the statistical problem at hand (Kennedy, 1998).

The classical linear regression model (CLRM) makes assumptions about the way data are generated and might be considered to be the standard estimating situation in the social sciences (Kennedy, 1998; Long, 1997). When all of the assumptions of the CLRM hold, OLS is the best linear unbiased estimator (Kennedy, 1998; Long, 1997). If the errors are also assumed to be normally distributed in CLRM, the OLS estimator also produces the maximum likelihood estimate (Kennedy, 1998; Long, 1997). Because of all these desirable properties of the OLS estimator in the CLRM, it is often the standard to which all other estimators are compared. However, the OLS only has these properties when the estimating situation can be accurately characterized by the CLRM. If even one of the assumptions is violated, a different estimating situation is created, and the OLS estimator may not be preferred (Kennedy, 1998). The OLS estimator often takes the CLRM assumptions for granted even though they are seldom true (Becker, 1983): many economic situations occur because of a violation of one or more of the assumptions

(Kennedy, 1998). For example, when categorical dependent variables are used, the assumptions of the CLRM no longer hold (O'Connell et al., 2008).

The first assumption that the CLRM makes is that the dependent variable is a linear function of a specific set of explanatory variables, related through the parameters, and an error term (Greene, 2000; Kennedy, 1998; Kirk, 1999; Long, 1997). In other words, for each unit increase in an explanatory variable, there is a fixed change in the number of units (given by the estimated parameter) in the dependent variable across the entire range of values of the explanatory variable (Dey & Astin, 1993; Long, 1997; Winship & Mare, 1984). However, oftentimes this linear relationship is unrealistic in many applications – some explanatory variables may have different effects on a dependent variable as their values change (Long, 1997). For example, in exponential relationships, as an explanatory variable increases, a dependent variable increases at an increasingly faster rate. In logarithmic relationships, the opposite is true. The logistical functional form describes a relationship in which the probability that an observed dependent variable takes on successively higher values rises (or falls) slowly at small values of an explanatory variable, more rapidly for mid-range values, and slowly again at large values (Winship & Mare, 1984).

When a nonlinear relationship is found to exist between the explanatory variables and predicted values, one or more variables can be transformed and then included in the regression model, resulting in nonlinear-linear regression models (Long, 1997). As long as there is a linear relationship between the transformed variables, regression analysis is appropriate. As an example, if a logistic functional form is discovered, the logit transformation (the dependent variable becomes the natural log of the odds) will result in a linear relationship (Greene, 2000; Long, 1997; O'Connell et al., 2008). If there is an exponential relationship between explanatory and predicted variables, the log of both sides of the equation can be taken. This exponential transformation results in an equation that is linear in the log of the dependent variable even though it is not linear in the dependent variable (Long, 1997). When dealing with a binary outcome, the assumption of a linear relationship between untransformed variables does not hold and is discussed further in the next section.

Levels of measurement revisited. Although the assumptions of the CLRM make no specific requirements on the level of measurement of the dependent variable, the use of ordinal-level data can present problems (McKelvey & Zavoina, 1975). The relationship between explanatory variables and an ordinal dependent variable is not linear because it is not a truly continuous variable, at least in its observed form (McKelvey & Zavoina, 1975). The CLRM, because of its nature as a parametric model, assumes that the dependent variable is a continuous, interval-level measure with an infinite range of real number values (Dey & Astin, 1993; Frone, 1997; Gardner, 1975; Long, 1997). In contrast, an ordinal scale has discrete and bounded values (Becker, 1983). The CLRM assumes that data are distributed around a line with mean error equal to zero and a constant variance (Kennedy, 1998; McKelvey & Zavoina, 1975). When the relationship is not linear, these assumptions do not hold and the OLS estimator may not be appropriate (Becker, 1983).

If the statistical model chosen assumes the wrong level of measurement for the dependent variable, the estimator could be biased, inefficient, or inappropriate (Long, 1997). As noted, if there are no violations of the assumptions of the CLRM, the estimates of parameters by OLS are unbiased (Harwell & Gatti, 2001). Interval-scaled, continuous data assume equal distances between units, and because of this property, meaning can be assigned to the shape of the distribution of a variable (Gardner, 1975). Parametric statistics such as the CLRM make the assumption of continuous data with a common distance between units (Agresti, 1996; Porter, 1999). Due to the varying or unknown distances between rankings in ordinal data, Stevens (1946) prescribed different permissible transformations for ordinal versus interval data. However, the debate on whether to use parametric or nonparametric statistics for ordinal-level data has effectively been resolved in favor of parametric statistics (Gardner, 1975; Labovitz, 1967). Therefore, parametric regression techniques especially designed to analyze qualitative, or categorical, dependent variables – ordered or un-ordered – are recommended (Agresti, 1996; Becker, 1983; Harwell & Gatti, 2001; Long, 1997). Such techniques include binary logistic regression, multinomial logit models, and ordered logit (or probit) regression.

Regression for Categorical Dependent Variables

Binary logistic/probit regression. When the dependent variable is dichotomous, possessing only two response categories, one approach is to analyze it with the linear probability model (Agresti, 1996; Kennedy, 1998; Long, 1997). This model assumes a linear relationship between the explanatory and dependent variables, a binomial distribution for the random component (dependent variable), and the identity link function, which models the mean directly (Agresti, 1996). However, linear functions take values over the entire real number line, so it is possible that predicted values may fall outside the possible values of zero or one, especially for very large or small values of explanatory variables (Agresti, 1996; Kennedy, 1998). Additionally, even though this is a linear regression model, OLS estimators are not good in this situation. The variance of the outcome is not constant for all values of the explanatory variable but depends on the predicted values (O'Connell et al., 2008). In this situation, ML estimators have smaller standard errors (Agresti, 1996). Furthermore, a linear relationship may not be appropriate when dealing with dichotomous data. Specifically, the relationship between explanatory variables and a probability (a value between zero and one) is usually nonlinear: changes in explanatory variables are likely to have more impact in the middle of the range of values (Dey & Astin, 1993). Also, a fixed change in an explanatory variable may have less of an impact when the probability is near zero or one (Agresti, 1996).

These type of nonlinear relationships are usually monotonic, taking an S-shape curve (Agresti, 1996; Kennedy, 1998). The most important of these is the logistic regression function, which is a special case of the GLM for binary data with a binomial random component and a logit link (Agresti, 1996). By limiting the predicted values to be between zero and one, the logit models do not encounter the problem of predicting outside the range of possible values that the linear probability model does (Agresti, 1996; Long, 1997). Additionally, the logistic distribution is useful for modeling the non-linear relationship between explanatory variables and a probability because of its ease of implementation and straightforward interpretation (O'Connell et al., 2008).

Logistic regression attempts to model the odds of an event's occurrence by comparing the probability that an event occurs to the probability that it does not occur (O'Connell et al., 2008). The estimated parameter values determine the rate of increase of the logistic curve – as the parameters increase, the curve (slope) becomes steeper (Agresti, 1996). Additionally, an odds ratio (OR) can be constructed to examine an explanatory variable's impact on the odds of an event's occurrence. The OR provides information on how odds for the dependent variable change as the explanatory variables increase or decrease; however, ORs range between zero and infinity, resulting in a skewed distribution. Therefore, logistic regression models the natural log of the odds, or logits, of a distribution because these logits can range from negative infinity to positive infinity. This results in a linear model in which a null hypothesis value of zero can be tested. The infinite range of the logits and the ability to conduct a hypothesis test are familiar properties that are easy to work with and parallel the CLRM (O'Connell et al., 2008).

Another way of conceptualizing the model for dichotomous response variables is to suppose an unobserved, latent variable that generates the observed values when the propensity crosses a threshold that results in the observed decision (Long, 1997). This latent variable is continuous, but since it is unobserved, the OLS estimator is inappropriate. What is observed is a discrete, nominal-level variable. Because of the resulting nonlinear relationship between the explanatory variables and the dichotomous dependent variable, an exact algebraic method for estimation does not exist; therefore, the iterative estimation method of ML is more appropriate (Long, 1997; Winship & Mare, 1984).

The logit models are almost always estimated by the ML estimator (Kennedy, 1998; Long, 1997). As previously noted, the use of the ML estimator requires an assumption about the distribution of the error (Kennedy, 1998; Long, 1997). By assuming a specific form for the distribution of the error, it is possible to compute the probability of the observed data for a given value of the explanatory variable (Long, 1997). The probability of the dependent variable given the explanatory variables can be described by a cumulative density function. In the probit model, the cumulative density function is the cumulative normal distribution, and in the logistic model, the cumulative density function is the cumulative logistic distribution. In other words, if the errors are assumed to be normally distributed, the result is the probit model, and if logistically distributed errors are assumed, the result is the logit model. The actual variance of the errors cannot be estimated, so a variance is assumed – it is either one in the probit model or $\pi^2/3$ in the logit model (Long, 1997).

The choice between the use of a logit or probit curve is largely one of convenience and convention since the curves are nearly identical, and the substantive results are largely indistinguishable for practical purposes (Agresti, 1996; Kennedy, 1998; Long, 1997). The difference between the curves is in the tails, so extremely large sample sizes are required to distinguish whether observations came from a logit or probit model, and it is rare even then to find data for which a logit or probit model fit differently (Agresti, 1996; Long, 1997; Maddala, 1983). The estimate of the parameter is not comparable between the models, though, since they are on different scales (Agresti, 1996; Maddala, 1983), but some texts offer equivalence equations (Long, 1997). The logit model was introduced after the probit model and is preferred by some possibly due to the relatively simple interpretation of logit coefficients as odds ratios (Agresti, 1996; Long, 1997).

Multinomial logit models. When dependent variable data are nominal (with more than two categories) and ordinal level, a multinomial (as opposed to binomial or normal) distribution is assumed for the random component of the GLM (Agresti, 1996). The multinomial probability distribution specifies the probability for each possible way of allocating *N* observations into *J* categories (Agresti, 1996). Multinomial logit models can be used when the dependent variable is ordinal, but there is a loss of efficiency since information about the ordering is being ignored. However, if there is any question about the ordering of the data, the multinomial model should be used because results of a model for ordered data would not make sense or would be biased (Long, 1997).

There are various models for analyzing nominal data. The conditional logit model is appropriate when there is information on the choices, and characteristics of the choices are used to predict the outcome (Long, 1997; Porter, 1999). It is not common in educational data to have such information, so a multinomial logit model is often used (Porter, 1999). In this model, data over the individual are analyzed, and the effects of the explanatory variables are allowed to differ for each outcome (Long, 1997; Porter, 1999). A multinomial probit, as opposed to logit, model is impractical because of computational difficulties (Long, 1997).

The multinomial logit model can be thought of as an extension of the binary logit model, like a linked set of binary logits. However, the binary logit equations comparing *J-1* pairs of categories are estimated simultaneously because doing so separately would be inefficient and would result in parameter estimates with large errors (Agresti, 1996; Long, 1997). Because more than one set of parameters can generate the same probabilities of the observed outcomes, the multinomial logit model is not identified, and a constraint must be imposed on a parameter: usually one of the parameter vectors is set to zero for one of the choices (Long, 1997).

There are two general classes of multinomial logit models: an odds model or a probability model (Long, 1997). These models differ in regards to what the logits represent. In the odds model, baseline category logits are calculated which represent the log odds that a response is category A vs. category B (neither one of which being the baseline category) given the explanatory variables (Agresti, 1996). Alternately, the multinomial logit model can be expressed directly in terms of response probabilities of observing a particular outcome (no baseline category is used) (Agresti, 1996; Long, 1997). The odds model is simpler than the probability model because the effect of a unit change in explanatory variable on the logit does not depend on the level of explanatory variable (Long, 1997; Peng & Naegle Nichols, 2003), but it is hard to interpret and convey the meaning of a change in the log of the odds (Long, 1997). However, interpretation of the parameters (slope values) can be aided through the use of odds ratios, or the change in the odds of the dependent variable occurring given a unit change in the explanatory variable (Peng & Naegle Nichols, 2003).

An alternate conceptualization of the multinomial logit model is as a discrete choice, or random utility, model in which an individual chooses the outcome that maximizes the utility derived from that choice. In this model, the utility to a consumer of an alternate is specified as a linear function of characteristics of the consumer and attributes of the alternative, plus an error term. The resulting logit is the probability that the utility of an alternative to a particular consumer is greater than that of all other alternatives (Long, 1997; Kennedy, 1998). However, this model runs into a problem because it is characterized by the independence of irrelevant alternatives (IIA) property which presumes that the introduction of an alternative that is nearly identical to an existing alternative would cut the probability of choosing the existing alternative in half and leave the probabilities of choosing the other alternatives unaffected (Kennedy, 1998). However, in practice, the relative probability of existing outcomes is often not unaffected by the addition of another alternative, making the random utility conceptualization of the multinomial logit model inappropriate when two or more alternatives are close substitutes (Hilmer, 2001; Kennedy, 1998).

Ordered logit/probit regression. The latent variable conceptualization of a dependent variable is useful for understanding models for ordinal data. The observed ordering in a dependent variable can be thought of as resulting from a continuous, unobserved measure which is a linear function of explanatory variables (Kennedy, 1998). When the latent continuous variable increases beyond a threshold value, the observed variable takes on a higher value, or score (Zumbo & Ochieng, 2002). The common view of ordered variables is that they are nonstrict monotonic transformations of unobserved interval-level variables. That is, one or more values of the latent, continuous variable are mapped into the same value of a transformed ordinal variable: the observed variables are discrete realizations of the unmeasured continuous variables (Harwell & Gatti, 2001; Winship & Mare, 1984).

The latent variable is assumed to be interval level, but because of inadequate measurement techniques, it cannot be measured directly (Harwell & Gatti, 2001; McKelvey & Zavoina, 1975). The manifest, or observed, variable is measured on an ordinal scale, resulting in incomplete data: the observed categories are ordered but separated by unknown distances (Harwell & Gatti, 2001; McKelvey & Zavoina, 1975). Because the coding of such a variable reflects only the ranking and not the difference between units, the OLS estimator is not appropriate when an ordinal variable is used as the dependent variable in a statistical model due to the violation of the linear relationship assumption of the CLRM (Kennedy, 1998). Therefore, estimation of the parameters is by ML. In ordered logit or probit regression, the threshold parameters, or the unknown boundary values which cause the latent variable to take on a certain observed value, are also estimated, providing additional information about the distribution of the dependent variables such as the distance between categories (Kennedy, 1998; Long, 1997; Winship & Mare, 1984).

Statistical methods for ordered categorical data are an improvement over the multinomial logit model because they incorporate the information about ordering of the dependent variable, resulting in simpler interpretations and greater power (Agresti, 1996; Zumbo & Ochieng, 2002) and the avoidance of the IIA problem encountered in discrete choice multinomial logit models (Hilmer, 2001). However, the assumed natural ordering must be realistic in order to make the use of the ordered logit or probit model appropriate. Tests for significance of the thresholds in the ordered models can provide a check for this assumption (Hilmer, 2001).

The ordered logit/probit regression models produce cumulative probabilities, using the cumulative logit extension of the logit link, that reflect the ordering of categories. These cumulative logits represent the probability that the observed dependent variable falls into a certain category or below relative to those ranked above it (Long, 1997; O’Connell et al., 2008). The cumulative odds for the highest category, then, is necessarily equal to one (O’Connell et al., 2008).

In the proportional odds model for ordered responses, there is a common effect for the parameter on all categories. In other words, the probability curves all have the same slope because the model has invariance to the choice of response category (Agresti, 1996). Long (1997) refers to this as the proportional odds, or parallel regression, assumption. This parsimonious assumption is useful when the research interest is limited to the likelihood of a response being at or below a category, as it implies that the effect of any explanatory variable is constant regardless of response value. Proportionality is restrictive because equal log-odds for each explanatory variable are assumed across all cumulative logits, but in many research situations this is a reasonable assumption (O'Connell et al., 2008) which can be tested (Long, 1997; O'Connell et al., 2008).

As with the models for binary outcomes, the unobserved mean and variance are unknown and must be assumed. If the normal probability distribution is used for the error term, the model is called ordered probit; if the logistic probability distribution is chosen, the model is called ordered logit (Winship & Mare, 1984). The choice is largely one of convenience, but if interpreting parameters in terms of odds is desired, the ordered logit should be used (Long, 1997). In terms of the mean of the dependent variable, which cannot be estimated with certainty because the variable is unobserved, a constraint must be imposed on one of the parameters (a process called parameterization) in order for the equation to be identified. The choice of constraint is arbitrary and does not affect the slopes or significance tests, but different software uses different parameterizations (Long, 1997).

Advantages of these models for ordered response data, besides the fact that they account for the ordered nature of the data, are that they are straightforward to estimate and they allow for analysis of the data within a familiar parametric statistical framework (Long, 1997; Winship & Mare, 1984). Furthermore, they take into account the ceiling and floor restrictions of the outcomes whereas the CLRM would not. This is advantageous especially when the observed dependent variable is highly skewed (Winship & Mare, 1984). However, the nonlinear relationship between explanatory variables and predicted probabilities makes interpretation difficult, and researchers must be sure that the dependent variable is indeed ordinal and that the assumption of parallel regression holds. Therefore, it is recommended to analyze data with both ordinal and multinomial models (Long, 1997).

Table 1 summarizes key properties of all four above-mentioned regression models.

Student College Choice

Sociologists, economists, and educational researchers have been studying the postsecondary educational decisions of students since the 1970s (Litten, 1982; Toutkoushian, 2001). These studies of student demand for higher education investigate the impacts of individual and institutional attributes on students' decision processes regarding where to attend college (Litten, 1982; Toutkoushian, 2001). Two strains of research on college-going behavior exist: generally, student demand models estimate equations to explain enrollments as a function of characteristics of the population of potential enrollees and of a set of existing schools. Student choice models predict student behavior regarding postsecondary decisions on an individual level (DesJardins, Dundar, & Hendel, 1999; Fuller, Manski, & Wise, 1982).

Table 1

Properties of Regression Models

Model	Results	Associated Dependent Variable	Probability Distribution	Link	Estimator	Assumptions
CLRM	Models the amount of change in the dependent variable for a one-unit change in a predictor variable	Continuous	Normal	Identity (mean)	OLS	1. Zero mean error. $E(\varepsilon_i) = 0$ 2. Non-stochasticity. $Cov(\varepsilon_i, x_i) = 0$ 3. Normality of errors. $\varepsilon_i \sim N(0, \sigma^2)$ 4. Homoskedasticity. $Var(\varepsilon_i) = \sigma^2 < \infty$ 5. Non-autocorrelation. $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ 6. Linear relationship between the predictors and the dependent variable. 7. Absence of high partial multicollinearity. (Garson, 2012b; Gujarati, 2003; Kreiberg, n.d.; Tabachnick & Fidell, 2001)
Binary Logistic/Probit	Models the natural log of the odds (logits) of success probability	Dichotomous	Binomial	Logit	ML	1. Non-autocorrelation. $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ 2. Linear relationship between the predictors and the log odds (logit) of the dependent variable. 3. Absence of high partial multicollinearity. 4. Large samples. 5. Adequate expected cell frequencies. (Bewick, Cheek, & Ball, 2005;

						Brant, 2004; Garson, 2012a; Tabachnick & Fidell, 2001)
Multinomial Logistic	Models the log odds of being in a category *m* versus a reference category	Polytomous (> 2 categories)	Multinomial	Multinomial extension of Logit	ML	1. Non-autocorrelation. $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ 2. Linear relationship between the predictors and the log odds (logit) of the dependent variable. 3. Absence of high partial multicollinearity. 4. Large samples. 5. Adequate expected cell frequencies. (Garson, 2012a; Tabachnick & Fidell, 2001)
Ordered Logit/Probit	Models the log odds that an observation falls at or below a given category	Polytomous with a presumed natural ordering	Multinomial	Cumulative logit extension of Logit	ML	1. Non-autocorrelation. $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ 2. Linear relationship between the predictors and the log odds (logit) of the dependent variable. 3. Absence of high partial multicollinearity. 4. Large samples. 5. Adequate expected cell frequencies. 6. Proportional odds (parallel regression). (Garson, 2011a, 2012a; Long, 1997; Tabachnick & Fidell, 2001)

The student college choice decision-making process can be described in various stages, or phases. On the most basic scale, a student must first decide whether to continue his or her education beyond high school. Then the student decides at which specific postsecondary institution to enroll (Hossler & Stage, 1992). Litten (1982) summarizes early literature on the stages of the student college choice process and describes five steps: college aspirations, decision to start the process, information gathering, application, and enrollment. Variations of this model expand on some of the stages to include, for example, admissions decisions by colleges and selection of a specific institution (Kotler, 1981). Kotler and Fox (1985), as cited in Paulsen (1990), describe the process in general decision-making terms which can be applied to the college selection process: need arousal, information gathering, decision evaluation, and decision execution. Specifically, Kotler and Fox (1985) describe a seven-stage model where students consider 1) generic alternatives (work, military, college); 2) alternatives among colleges (i.e. public vs. private); 3) a total college set; 4) an awareness set; 5) a consideration set; 6) a choice set; and 7) decision (Paulsen, 1990).

For ease of use in research in terms of focusing efforts and classifying findings, a three-stage model is generally employed (Clarke, 2007; DesJardins et al., 1999; Paulsen, 1990; Southerland, 2006). Clearly, each stage can be broken down into sub-phases and described more fully, especially in terms of characteristics that influence each stage, but the model developed by Hossler & Gallagher (1987) encompasses most aspects in its three stages: 1) predisposition (aspirations of students regarding postsecondary plans); 2) search (consideration of various institutions and other options, potentially resulting in a choice set), and 3) choice (application to the choice set and enrollment).

The majority of the research efforts in developing models of student demand for higher education have focused on the outcome of the final stage: enrollment (Hossler, Braxton, & Coopersmith, 1989). This is primarily due to the readily-available cross-section of post-secondary student enrollment numbers at any given point in time that can be examined through complex statistical models (Hossler & Stage, 1992; Toutkoushian, 2001). However, restriction to only this final-stage outcome limits the accuracy and usefulness of student demand equations found in such research. Only a small sub-sample of all eligible college students is studied in such cross-sectional research, resulting in a situation in which only part of the relationship between the dependent and predictor variables is observed. Even if all college applicants, not just college matriculants, are studied, the resulting equations cannot fully explain all the stages and sub-stages of the student college choice process (Toutkoushian, 2001). Additionally, a lack of accurate data regarding the number of applications a student submits complicates the study of the complete third stage (Hossler & Gallager, 1987). Therefore, in order to get a more complete picture of the entire college choice process, studies are now attempting to explain the predisposition phase and consideration sub-phase (Hossler & Stage, 1992; Hu & Hossler, 1998; Toutkoushian, 2001).

Although studies of student college choice attempt to explain factors that influence student decisions at each stage, the current study is limited to the final outcome, enrollment. Knowledge of factors that impact a student’s matriculation at various types of institutions is important to high school students when preparing to attend college. Therefore, the scope of the following literature review is limited to research on the final stage – enrollment in an institution of higher education. Following an argument for the importance of attending selective colleges, a description of the common frameworks for the study of student college choice research is given. In order to inform the building of a model for the current study, factors found to influence college attendance in general are examined. Because the current study investigates influences on selectivity of college attended, a review of the literature on factors specifically influencing attendance at selective institutions is also presented. The section concludes with a discussion of the methods used in student college choice research.

Importance of College Selectivity

The need for a college degree is becoming increasingly important for ensuring economic security, as education level is the single most important factor influencing income (Council for Aid, 1997). According to the latest Current Population Survey by the U.S. Census Bureau, average earnings in 2002 increased with each education level: the average worker with a high school diploma earned $27,280 annually, whereas the average annual income for a bachelor's degree holder was $51,194 (Stoops, 2004). This wage differential between education levels has been increasing over the last two decades (Scurry, 2003), and the Commission on National Investment in Higher Education predicts that the salaries of the highest paid workers will continue to rise with inflation while the salaries of lower-paid workers will decrease drastically by the year 2015 (Council for Aid, 1997). In fact, between 1972 and 1999, the median household income for someone with only a high school diploma decreased by 13.1% while the median income for a college graduate increased by 9.9% (Scurry, 2003). Further, as of 2003, 70% of jobs in growth industries required some post-secondary education (Scurry, 2003), and there is no reason to believe that this percentage is decreasing. Clearly, a college degree is necessary for economic survival.

Within the subset of adults who hold a college degree, however, there are additional benefits for those who have graduated from a college or university considered to be selective (Davies & Guppy, 1997; Ehrenberg, 2003). For example, attending a more selective institution has been associated with positive impacts on educational, income, and status attainments as well as on socially valued individual characteristics (Hearn, 1991). The majority of the research done in this area has focused on the impacts of college quality on post-graduation income levels, but recent research has also explored the relationship between selectivity of college or university attended and wage growth and adult leadership (Arnold, 2002; Thomas & Zhang, 2005).

One of the first studies of the economic effects of quality of baccalaureate education was a survey of World War II veterans regarding their income after college (Solmon, 1972). Solmon found a significant effect on lifetime earnings for college selectivity even after controlling for individual factors such as student ability. Decades of research since Solmon (1972) has largely confirmed the positive impact of college quality on earnings (see Brewer, Eide, & Ehrenberg, 1999; Fitzgerald, 2000; James, Alsalam, Conaty, & To, 1989; Monks, 2000; Mueller, 1988; Rumberger & Thomas, 1993; Thomas, 2003; and Trusheim & Crouse, 1981). For example, Mueller (1988) controlled for a variety of background variables including family income, academic ability, and degree aspirations and found that selectivity had a significant, albeit small, impact on post-college earnings. Brewer et al. (1999) used the National Longitudinal Study of the High School Class of 1972 and the High School and Beyond datasets and found a significant effect of college selectivity on wages and earnings for three high school graduating classes six, ten, and fourteen years after high school (Brewer et al., 1999).

Similarly, Monks (2000) examined a longitudinal survey spanning 17 years in order to obtain data on students from the time they attended college through their early labor market years and found that students who attended selective colleges earned significantly more than their peers who attended less-selective colleges. Both Brewer et al. (1999) and Monks (2000) controlled for sample selection effects, or individual characteristics which influence the enrollment process. Dale and Krueger (2002) used a matched sample to control for the bias that may result due to selective colleges' acceptance of students with characteristics that are positively related to future earnings and found that although there was no significant difference in earnings for students with similar ability levels, low-income students earned more if they attended selective colleges.

Rather than studying earnings at discrete points in time, Thomas and Zhang (2005) examined the impact of college selectivity on wage growth. A previous study by one of the authors (Thomas, 2003) found a small yet significant effect of college selectivity on initial earnings of college graduates, and Thomas and Zhang (2005) extended this work and looked at the changes in earnings of graduates from one year after college graduation to four years after graduation. One year after graduation, graduates from high-quality schools had a 9% earnings advantage over graduates from low-quality institutions; however, four years after graduation, this advantage increased to 20%. Therefore, this research suggests a changing pattern of influence whereby stronger effects of college selectivity on earnings emerge over time (Thomas & Zhang, 2005).

Finally, there is some evidence that attending a prestigious university leads to leadership in the top levels of occupations (Arnold, 2002). Data from a longitudinal study of academically talented students, the Illinois Valedictorian Project, suggest that top career achievers are set apart from other valedictorians by their attendance at the most prestigious colleges. Furthermore, a study of college graduates who received Rhodes scholarships – an already very select group – indicated that attending a highly prestigious undergraduate institution is the strongest predictor of adult professional prominence (Arnold, 2002).

The precise mechanisms by which graduation from a selective college or university yields economic benefits are unknown (Arnold, 2002; Monks, 2000; Thomas & Zhang, 2005). Some researchers posit that institutional characteristics result in a higher-quality education at selective versus non-selective colleges, thereby increasing human capital accumulation. Additionally, the name recognition and credentials of prestigious institutions may signal to employers that a graduate has superior capabilities (Arnold, 2002; Monks, 2000; Thomas & Zhang, 2005). Whatever the case, students who are interested in maximizing their lifetime earnings likely attempt to enroll in selective colleges and universities (Ehrenberg, 2003; Monks, 2000). Furthermore, Arnold (2002) argues that prestige of undergraduate institution is even more important in competitive environments. With the rate of adults earning bachelors' degrees at an all-time high (Stoops, 2004), it is important for a college graduate to stand out among his peers by having attended a selective institution.

Therefore, it is to a high school student's benefit to know how to best prepare for admission to a highly selective university. Research on student college choice can offer insight into the factors that influence matriculation to a selective college in order to determine what, if anything, students can do to enhance their chances. Additionally, it is important for higher education administrators to understand factors that influence student matriculation to their university or university type so that they can accurately plan enrollment management efforts.

The competitive nature of the student recruitment process in which students are seen as consumers, or shoppers among academic institutions, demands market-oriented strategies on the part of colleges (Paulsen, 1990). An understanding of the factors that shape the decision to attend a particular institution helps college administrators plan and forecast enrollment (Paulsen, 1990) and manage their institution's involvement in this decision process (Chapman, 1981; Litten, 1982; Paulsen, 1990). It is important for colleges and universities to develop appropriate strategies to attract students (DesJardins et al., 1999; Hossler & Stage, 1992) and to influence the college decision process (Paulsen, 1990), not only to fund their budgets but potentially to enhance their reputations: the enrollment of high quality students leads to a better student profile, or reputation, which in turn attracts more high quality students (Toutkoushian, 2001). Additionally, knowledge of the factors that influence matriculation to particular institutions can be helpful to policymakers interested in providing equitable access to higher education for underserved populations (Hearn, 1991; Hossler & Stage, 1992; Toutkoushian, 2001).

Theoretical Frameworks for Study

The final, "choice" stage of the student postsecondary decision process begins when a student applies to one or more institutions and subsequently is admitted to and then enrolls in one of these schools in the "choice set" (Hossler & Gallagher, 1987). Included in the literature reviewed below are studies of factors that affect the dichotomous decision of whether or not to attend college as well as studies of influences on attendance at different types of colleges. Studies of students' aspirations, predispositions, or search processes are not examined.

The impacts on college attendance of many different individual and institutional characteristics have been studied. The choice of variables to include in a model depends on the theoretical framework used to describe the college decision process. McDonough and Antonio (1996) posit that there are three basic approaches to the study of college choice influences: social psychological studies, economic studies, and sociological status attainment studies. Southerland (2006) describes four distinct approaches: economic models, psychological models, sociological and environmental models, and organizational and interactional models. Related to economic models, Hossler et al. (1989) and Hu and Hossler (1998) add another perspective: the consumer, or marketing approach. In addition to the commonly-used economic and sociological frameworks, researchers acknowledge that many studies, especially those conducted by educational researchers (Hu & Hossler, 1998), adopt a combined model (Hossler et al., 1989; Southerland, 2006).

Strictly economic models of postsecondary participation and choice of college are based on human capital theory and explain student decisions in terms of cost/benefit analysis (DesJardins et al., 1999; Hu & Hossler, 1998; Southerland, 2006). This type of study employs a theoretical framework that views students as selecting optimal educational decisions by which to maximize benefits and minimize costs. Through this mechanism, returns to high school and college capital accumulation can be converted to occupational payoffs (McDonough, Antonio, & Horvat, 1997). Students will invest in education until the benefits, largely in terms of increased future earnings, are offset by the costs of more education (Venti & Wise, 1982). In making the decision, students compare benefits with costs for all alternatives and select the greatest net benefit given their individual preferences and characteristics (Perna & Titus, 2004). Some short-term benefits of college attendance are enjoyment of learning and enhancement of social status, whereas long-term benefits may include increased lifetime earnings, a more fulfilling work environment, and a lower probability of unemployment (Perna & Titus, 2005). Costs of college attendance include tuition (offset by any financial aid), foregone earnings and free time, and any perceived discomfort of expending the effort needed to take courses and/or travel between home and the institution (Perna & Titus, 2005; Venti & Wise, 1982). Factored in to these decisions are a student's academic ability and financial resources and the local labor market opportunities (Perna & Titus, 2004; Venti & Wise, 2004). McDonough and Antonio (1996) argue further that economic models of student college choice assume that a student has perfect information and is engaged in a process of rational choice.

In contrast to economic studies, research from sociological, psychological, and/or environmental perspectives emphasizes the role of personal characteristics in the college choice process. The role of student background characteristics such as gender, race, parental income and education level, and student academic ability and achievement are typically examined in these types of studies (Hu & Hossler, 1998). Southerland (2006) makes the distinction among psychological, sociological, and organizational models. In his conception of theoretical frameworks, psychological models account for individual psychological characteristics such as personality, motivation, self-image, adaptability, and coping mechanisms, whereas sociological models include the influence of the environment, in terms of social pressures and norms, on the college choice process. The focus in these studies is larger societal issues above and beyond individuals and institutions. This is similar to McDonough and Antonio's (1996) category of sociological status attainment studies which examine inequalities in college access due to individuals' social status and the impact of social status on the development of educational aspirations and expectations. Finally, Southerland (2006) describes organizational studies as focusing on the impact of an institution's characteristics, as well as an individual's interaction with the institution, on the college decision-making process. This orientation is very similar to what McDonough and Antonio (1996) describe as social psychological studies through which institutional characteristics such as academic program offerings, location, and campus social climate as well as a student's assessment of his or her "fit" with the college are examined.

Most frequently, however, theoretical frameworks of this clearly complex college choice process incorporate multiple perspectives to form a combined model. For example, Perna and Titus (2005) employed a conceptual model of influences on college enrollment that included the student-level characteristics of race, gender, economic capital (i.e., family income, importance of tuition costs and financial aid), cultural capital (i.e., parents' education level, participation in cultural classes), human capital (academic achievement and preparation), and social capital (parental involvement). Drawing from research done through 1980, Chapman (1981) developed a conceptual model of student college choice that takes into account the impact of student characteristics such as socioeconomic status, academic ability, and educational aspirations and external influences such as tuition and other institutional characteristics and the influence of peers and adults.

As noted, Hossler & Gallagher (1987) proposed a model that is widely used to explain the student decision process because it incorporates many of the most useful aspects of other frameworks (Hu & Hossler, 1998; Southerland, 2006). Hossler and Gallagher's (1987) model is based on previous work and is interactive in nature, emphasizing both the roles of individual and organizational factors in each of the three stages. The utility for educational researchers and policymakers of a combined model is that it approaches the decision process from a policy analysis perspective in order to find opportunities for intervention within the process (Hossler et al., 1989).

The literature reviewed below encompasses studies primarily from a combined theoretical framework, as this approach is of the most use to educational policy makers (Hu & Hossler, 2006). Although many studies of influences on college attendance include some economic variables, studies utilizing strictly economic models of cost/benefit analyses are not included.

Influences on College Attendance

Depending on the theoretical perspective of the researchers, certain variables thought to affect college attendance decisions are selected for study. Some studies examine the impact of primarily ascriptive and/or other student-level characteristics (e.g., DesJardins et al., 1999; Hearn, 1991), whereas others focus more on testing the explanatory power of hypotheses within cultural or social capital theories (e.g., Kaufman

& Gabler, 2004; Perna & Titus, 2005; Teachman, 1987). Christensen, Melder, and Weisbrod in 1975 noted that the principal explanatory variables examined at that time were academic ability and socioeconomic status (SES). Litten (1982) listed the variables that had received substantial research attention as race, income, parent education level, academic ability, class rank, courses taken, and price of college. Research on college enrollment typically includes two broad categories of student-level characteristics that may influence students' postsecondary decisions: ascriptive, or assigned, background characteristics such as socioeconomic measures, race/ethnicity, and gender, and academic characteristics such as ability, achievement, and preparation. Often student cognitive characteristics such as expectations and preferences are examined as well. The institutional characteristics of tuition, prestige, location, control (public versus private), and, to a lesser extent, curricular or program offerings, are also commonly examined (DesJardins et al., 1999; Do, 2004; Hearn, 1991; Heller, 2003; Hu & Hossler, 2006; Kinzie et al., 2004; McDonough et al., 1997; Monks, 2000; Southerland, 2006; Toutkoushian, 2001; Wolniak & Engberg, 2007). Additionally, three studies were found that specifically examined the impact of participation in high school extracurricular activities on college attendance (Kaufman & Gabler, 2004; Mahoney, Cairns, & Farmer, 2003; Marsh, 1992).

The primary determinants of whether and where a student attends college are academic ability and/or achievement and ability to pay, or family income level (Hearn, 1991; McDonough et al., 1997; Monks, 2000). Additionally, race, ethnicity, and gender play an indirect role through their effects on academic outcomes in secondary school (Hearn, 1991). Specific to attendance at prestigious schools, Hearn (1991) reported that the literature revealed that ascriptive factors as well as test scores and grades play a large role: students from lower-socioeconomic status families, minorities, and females were disproportionately located in less selective colleges through the 1970s, even after controlling for academic characteristics.

A more detailed description of common variables examined in student choice research is given below, and the specific impact of each on college attendance is reviewed in detail. The most commonly studied variables, ascriptive student background characteristics, are first discussed, followed by student academic characteristics and finally the institution characteristic of tuition, including the impact of financial aid offers. Following this review of the impacts of various factors on attendance at college, a separate section describes the subset of selective colleges, including a description of the students who attend selective colleges and a review of studies that have examined the impact of the common variables on selectivity of college attended.

Socioeconomic status (SES). The ascriptive characteristic of family background, most commonly referred to as socioeconomic status (SES), is traditionally measured by family income, parent education level, and family size (Hearn, 1988; Hearn, 1991; McDonough & Antonio, 1996; Teachman, 1987). Occasionally parent occupation is included (Alexander, Holupka, & Pallas, 1987; Christensen et al., 1975; Mahoney et al., 2003; Perna & Titus, 2004), and one study (Mahoney et al., 2003) used an already-configured socioeconomic index. Perna & Titus (2004) and Alexander et al. (1987) used a composite based on the above factors as well as number of household possessions such as newspapers and computers in order to approximate family background. Perna & Titus (2004) argue that a composite is a better measure of family background than individual factors such as family income because income should be adjusted for family size and does not account for “habitus,” or preferences for college enrollment.

Previous research has shown a hierarchy of the impact of student background effects with the strongest being SES (McDonough et al., 1997). Chapman (1981) summarizes findings that students of different SES levels enter postsecondary education at different rates and distribute differently across type of institution, with higher SES students more likely attending four-year colleges and universities than lower SES students. Hossler et al.’s (1989) review of the literature revealed that SES is related to the quality of institution attended – high-SES students are more likely to attend selective institutions – but not to the cost of institution attended. Heller (2003) reports that previous research has found a relationship between family income and parent education level and college participation: lower income students are less likely to attend college and when they do, they are less likely to enroll in four-year institutions. Similarly, students whose parents have higher levels of education are more likely to enroll in college (Heller, 2003) and when they do, they are more likely to enroll in selective institutions (Hossler et al., 1989).

According to Christensen et al. (1975), SES is thought to positively influence attendance in college in two ways: income can be thought of as a determinant of resources available to finance education, and the education and occupation of parents is likely related to amount of parent encouragement to attend college. Additionally, Chapman (1981) posits that family socioeconomic status’ impact is seen both through income as it interacts with the institutional characteristics of financial aid and tuition and through SES’s influence on other attitudes related to college choice such as aspirations and expectations. McDonough et al. (1997) contend that SES operates indirectly through aptitude, mediated by parents' knowledge. On the other hand, Teachman (1987) argues that family background does not explain much variance in educational attainment due to the imperfection of demographic measures and the fact that they do not account for home environment.

Christensen et al. (1975), examining the impact of SES using four separate variables (family income, occupation of father, education level of father and of mother) found that all SES measures significantly influenced attendance at college. Studying students from Wisconsin in 1963 and 1967, the researchers used probit analysis to find that family income has the least impact on college attendance and that father's education level had the greatest impact.

McPherson & Shapiro (1994) looked at changes in enrollment over time and found that from 1980-1993, low income students were increasingly represented at two-year colleges whereas middle & high income students were decreasingly represented. As for four-year universities, enrollment of middle income students was declining slightly at private institutions but increasing at public four-year institutions. Upper income student enrollment was also declining at private, four-year institutions but increasing at public universities. Similarly, Akerhielm, Berger, Hooker, and Wise (1998) used National Education Longitudinal Study of 1988 (NELS: 88) data and found that postsecondary education increases with family income: lower income students are less likely than higher income students to attend four-year schools, even among high ability students.

The conceptualization of the SES variable and the methods of analysis used seem to have an impact on the findings, however. Unlike McPherson and Shapiro (1994) and

Akerhielm et al. (1998) who only used family income as the measure of SES, Alexander et al. (1987) treated SES as a linear composite of father's occupation, father and mother education level, family income, and household possessions. An earlier study of Alexander et al.'s found that SES had a strong impact on college attendance, but their 1987 analysis showed that SES does not play much of a role in attending a two-year versus a four-year institution; therefore, SES plays a role in limiting opportunities to go to college but not for sorting into a particular type of institution. However, the SES effects were reduced once academic credentials were added.

Heller (2003) examined in detail four NCES reports on college participation: two descriptive analyses and two multivariate analyses. The tabular, or descriptive, studies confirmed great differences in college participation among students from different incomes and parent education levels, but the multivariate analyses did not. Due to the fact that the multivariate analyses can control for other factors, they are more causal in nature. Results of these NCES causal studies conflicted with existing research in that they found that the differences in college attendance among income levels were minimized or eliminated.

Race. Past research has shown that when academic achievement is held constant, student ascriptive characteristics exert influence over college attendance in the following order of importance: social class, followed at a distance by race and then gender (McDonough & Antonio, 1996). Akerhielm et al. (1998) analyzed NELS:88 data and found that Native American, Black, and Hispanic students are the least likely to go to college, whereas Asian/Pacific Islanders are the most likely. Alexander et al. (1987) found a modest effect of race on two-year versus four-year college attendance, with

Hispanic students overrepresented in two-year colleges. When academic characteristics were taken into account, Black students actually had a greater likelihood than White students of attending a four-year college or university (Alexander et al., 1987).

Results can differ depending on the variables included in a model. For example, DesJardins et al. (1999) found that African American students were the most likely to apply to the four-year institution in their study, but only after student educational characteristics were added to demographic factors in the model. Additionally, Perna (2000) found that when only race and gender were in their model, African American and Hispanic students were less likely than White students to enroll in four-year institutions. However, when controls for sex, costs, benefits, financial resources, ability, and cultural and social capital were added, African American students were the most likely to enroll in four-year institutions, followed by White and then Hispanic students. Depending on the combination of variables added to the model, the results changed. Perna (2000) also found many interactions and ran analyses separately for each racial group and found important differences in attendance influences. These results indicate that the process of deciding to invest in college is different for different groups; therefore, when trying to understand ethnic differences in student choice behavior, it is important to examine differential impacts among racial/ethnic groups of the variables that influence college enrollment decisions.

Sex. Studies of student college choice using various dependent variables (attendance in college or not; attendance at two- vs. four-year colleges; selectivity of institution attended) have found no significant sex differences (Alexander et al., 1987; Davies & Guppy, 1997; Hearn, 1988; McDonough & Antonio, 1996). For example, Alexander et al. (1987) reported that sex differences for attendance at two-year versus four-year colleges were "virtually nonexistent" (p. 67). However, there seems to be a differential impact of the variables that influence the student college choice process on males and females (Christensen et al., 1975; Hearn, 1991). Christensen et al. (1975) found, for example, that females are more sensitive to the influences of mother's education level and cost of college than are males but that both sexes have the same response to the influence of ability and the SES measures of income, father's education level, and father's occupation.

Academic characteristics. Student academic ability or achievement is most often measured by standardized test scores such as the SAT or ACT® and high school performance as evidenced by grade point average (GPA) or class rank (Christensen et al., 1975; Hand & Prather, 1987; McDonough & Antonio, 1996; Venti & Wise, 1982). Frequently, high school course-taking patterns are incorporated, typically by categorizing a student as having taken an academic, or college preparatory, track or not (Akerhielm et al., 1998; Hearn, 1991). Perna and Titus (2004) propose, however, that this track measure is unreliable and that the highest level course taken in a particular subject area should be measured instead. Additionally, some studies incorporate extracurricular activity participation as a measure of academic background or preparation (Hearn, 1991; Mahoney et al., 2003; Marsh, 1992).

Previous research has shown that student achievement measures affect the overall demand for higher education and the type of institution attended; specifically, greater academic success in high school is associated with greater likelihood of attendance at college and a wider variety of options for where to attend (Toutkoushian, 2001; Wolniak

& Engberg, 2007). Hossler et al.'s (1989) review of the research concluded overwhelmingly that academic ability is positively associated with selectivity of postsecondary institution applied to and attended. Perna and Titus (2004) posit that this effect is because an individual's academic achievement and quality of academic preparation are indicative of his or her available human capital to invest in the educational process. Chapman (1981) explains that students with good academic records possibly receive more encouragement to continue their education and are more likely to receive college advising from guidance counselors and scholarships from colleges.

Venti and Wise in 1982 found that SAT scores were more related to student application and choice of college quality than were the actual college admissions decisions. In terms of the impact of academic predictors on subsets of students, Hand and Prather (1987) found in separate analyses that SAT scores and GPA exhibited similar patterns of influence across minority group and gender.

Two studies found contradictory results regarding the relative impact of academic characteristics and SES variables. Christensen et al. (1975) tested the hypothesis that ability, measured by high school class rank and a standardized aptitude test, should be positively related to attendance at college because ability is a factor in determining the return on investment of education. Their hypothesis was supported, but the measures of ability were not as important an influence on college attendance as the SES variables in their study. In contrast, Alexander et al. (1987) analyzed the influence of a composite standardized test score, a dichotomous curriculum measure (academic or non-academic track), and self-reported GPA. The researchers found that these academic measures were not only all significantly related to type of college attended (two- vs. four-year), but they had a much bigger influence than the ascriptive characteristics of SES, race, and gender.

In two studies of the impact of extracurricular participation on college attendance, this particular academic variable was found to have a positive influence. Marsh (1992) measured extracurricular participation in sophomore and senior years and controlled for background variables (student ascriptive characteristics, school-level variables, geographic location, and college expectations). He found that total participation was significantly positively related to college attendance, although results differed based on the particular activity. Similarly, Mahoney et al. (2003) found that extracurricular participation in middle adolescence had a significant effect on educational status at age 20 when controlling for gender, family economic status, interpersonal competence, and educational aspirations. Further, a path analysis revealed that extracurricular activities had positive, significant links to interpersonal competence in middle adolescence which in turn influenced educational aspirations and ultimately educational status (Mahoney et al., 2003).

Tuition and financial aid. Research findings present mixed results on the effect of cost on college attendance, and there is reason to believe that the impact should not be studied separately from the influence of financial aid (Chapman, 1981). According to Chapman (1981), early research suggested that cost exerted more of an influence over whether or not a student goes to college than on the particular college attended, but other research suggested that cost did make a difference in college selection. Specifically, in studies of students' reasons for choosing a particular college, the students frequently reported cost as being important in their decision. However, Toutkoushian (2001) summarizes that student demand is “relatively insensitive to changes in the cost of attendance” (p. 248).

Again, these differing conclusions could be due to the impact of financial aid. Chapman (1981) writes that previous research suggests that without financial assistance, a high percent of students would be severely restricted in their college options. Additionally, Hu & Hossler (1998) note that “Previous studies suggest that increases on tuition costs will reduce the demands for postsecondary education and the offers of financial aid will offset this decrease” (p. 5). Hu & Hossler (1998) found that student sensitivity to tuition and financial aid varied among income levels and type of institution preferred, further indicating that the two factors (tuition and financial aid) should not be studied separately.

As financial aid is theoretically supposed to increase students’ college choices by reducing or eliminating the obstacle of cost (Chapman, 1981), many studies of the impact of financial aid have examined issues of equity in access to this aid. McDonough and Antonio (1996) found that financial aid and college cost are more important to African American students than to White students. Similarly, Wetzel, O’Toole, and Peterson (1998) examined students’ sensitivity to changes in the cost of college using enrollment numbers as a dependent variable. The researchers found that Black students’ sensitivity to changes in cost is two-thirds higher than White students’ sensitivity, suggesting that minority students have been responding positively to financial aid.

Hossler and Gallagher (1987) summarize the research on the impact of financial aid in the third, or choice, stage of the college choice process and conclude that the impact is limited: a large amount of aid is needed to move a second-choice to a first-choice school. However, the researchers note that the relationship between choice and price is sensitive to a number of variables and does not lend itself easily to generalizations. Although financial aid and net cost do have an impact on matriculation, this impact varies among students and institutions.

Influences on Attendance at Selective Colleges

Owings, Madigan, and Daniel (1998) used NELS:88 data to describe the characteristics of students who attend highly ranked national universities. The researchers found that Asian/Pacific Islanders were more likely than other ethnic groups to attend *U.S. News and World Report*-ranked Tier 1 National universities, as were students from high SES backgrounds. Additionally, students with SAT scores of greater than or equal to 1100 or with GPAs of 3.5 or higher were also more likely to attend these highly-ranked schools. Rigorous course-taking, participation in extracurricular activities, approval of teachers, and in-home computer access also described students at the Tier 1 national universities (Owings et al., 1998).

Using similar characteristics to approximate criteria of importance to admissions officers at highly selective colleges, Owings, McMillen, Burkett, and Daniel (1995) described the type of high school students who met these criteria. More males than females (6.9% vs. 4.7%) met the five criteria in the study (GPA of 3.5 or above, SAT of 1100 or above, college-preparatory course-taking pattern, positive teacher comments, and participation in two or more extracurricular activities). Asian students met the criteria at the highest rate (8.8%), followed by White students (6.5%), Hispanic students (2.5%), Black students (0.4%), and American Indian students (0%). Furthermore, students from high SES backgrounds met the criteria at the greatest rate, 8.8%, compared to 4.1% for middle SES and 1.5% for low SES students.

Findings from primary studies using regression, rather than descriptive, methods to analyze the impact of student characteristics on attendance at selective institutions are summarized below.

Socioeconomic status (SES). Previous research has indicated a relationship between income and college selectivity, whereby upper income youth are especially likely to enter elite colleges (McDonough et al., 1997). However, once other student characteristics – especially academic achievement and/or ability – are controlled for, this relationship changes.

Hearn (1988) examined influences of student characteristics on attendance at selective institutions using cost of institution as the continuous[2] dependent variable. Examining four different predictor variables related to SES, Hearn found that when only the ascriptive characteristics of race, gender, and SES were in the model, father's and mother's education level and family income were significantly related to attendance at a selective institution but that family size was not. However, when academic characteristics

---

[2] Davies and Guppy (1997), Hearn (1988, 1991), and McDonough and Antonio (1996) all defined average SAT scores as a continuous variable and therefore used the classical linear regression model in their analyses. However, the status of average SAT score as a continuous measure can be challenged as it does not take on an infinite range of real values and in fact may be considered to be a limited dependent variable – a continuous variable whose observed values do not cover its entire range (Frone, 1997). Limited outcomes pose problems for the linear regression model, and nonlinear models for such data exist (Frone, 1997). Although average SAT score may not truly be a continuous variable, I will continue to refer to it as continuous in this paper because the student college choice researchers conceptualized it as such.

(standardized test scores, high school grades, academic track, and educational aspirations) were included, the SES effects all but went away (only mother's education level was still related to selective institution attendance). Therefore, Hearn (1988) concluded that the effects of these ascriptive background characteristics were mediated through academics. A similar study using selectivity as a continuous dependent variable measured by average SAT score of all students at an institution essentially replicated these findings (Hearn, 1991).

Davies and Guppy (1997) extended the work of Hearn (1991) to examine the impact of ascriptive characteristics on the selectivity of college attended in order to determine if students from varying backgrounds enjoy equal opportunity and to explore how such inequalities, if present, occur indirectly through academic factors. Like Hearn (1991), Davies and Guppy (1997) used average SAT scores as a continuous dependent variable to measure institution selectivity; however, Davies et al. used a different data set and included interaction terms as predictor variables as well. Hearn's (1991) findings were essentially replicated, but there was an interaction with SES and ability.

Additionally, Owings et al. (1998) employed logistic regression to discover that family SES, measured as a composite of both parents' occupations and education levels and family income, had no significant relationship to enrollment in Tier 1 national universities after controlling for student academic characteristics. However, having a computer in the home in eighth grade was significantly related to subsequent enrollment in a highly-ranked university (Owings et al., 1998).

McDonough and Antonio (1996) found that parent education level and family income were positively associated with more selective college attendance for White students only once a full model including various cultural capital variables was used. These SES effects disappeared for Black, Asian, and Hispanic students once test scores and grades were added to the restricted model. Clearly, SES does not work in isolation to influence students' post-high school enrollments – many factors must be taken into consideration when determining its effect.

Race. Although Hearn in 1988 concluded that race was not important in determining enrollment at a selective institution, as measured by cost of attending the institution, he later found different results (Hearn, 1991). When only the effects of student ascriptive background characteristics (race, gender, and four SES variables) on attendance at selective institutions (a continuous dependent variable of average SAT score) were tested, being Black or Hispanic had a significant effect. However, when academic characteristics were added, the significance of being Hispanic disappeared, leaving the only relationship between race and attendance at a selective institution a negative one for Black students (i.e., Black students are less likely to attend selective institutions even if their academic ability is high). Davies and Guppy's (1997) extension of Hearn's 1991 study mirrored the findings of Hearn's (1988) earlier work in that race was not found to have a significant impact on selectivity of institution attended, either when only ascriptive characteristics were included or when academic characteristics were added.

Sex. Similar to the findings regarding college attendance in general, sex does not seem to have an effect on attendance at selective postsecondary institutions. Hearn (1988), McDonough and Antonio (1996), and Davies and Guppy (1997) all concluded that sex is not an important determinant of selectivity of college attended, especially when other variables are controlled for; however, Hearn (1991) found that being female was negatively related to selectivity when all other ascriptive background and academic characteristics were in the model (i.e., females are less likely to attend a selective institution even if they do have high academic ability). Kaufman and Gabler (2004) examined the impact of extracurricular activities on attendance at elite institutions and found the same results for males and for females, with the only exception being that elite colleges seemed to prefer students who pursued activities stereotypically associated with the opposite gender (such as boys who served on the yearbook or newspaper staff).

Academic characteristics. Hearn (1991) concluded that the primary direct influences on selectivity of college destination are academically based. Specifically, test scores, grades, academic track, extracurricular involvement (in student government), and educational expectations are all positively related to selectivity of college attended. Although there were also significant direct effects for ascriptive factors, when both ascriptive and academic variables are considered, the most powerful are academically based – primarily test scores. As previously noted, Hearn (1991) explains that the SES variables have an indirect effect by helping to shape students' academic credentials. McDonough and Antonio (1996) and McDonough et al. (1997) found that receiving high grades and SAT scores and taking foreign language courses in high school are positively associated with attendance at selective versus non-selective colleges for all students (in the aggregate and broken down by ethnic group), even when controlling for student background characteristics and multiple measures of cultural capital. Additionally, visiting an art gallery or museum while in high school was significantly positively associated with attendance at a selective college for the aggregate student population studied as well as for the ethnic groups of White, Black, and Asian students but not for Hispanic students (McDonough & Antonio, 1996; McDonough et al., 1997).

Kaufman and Gabler (2004), in a study of hypotheses couched within cultural capital theory, examined the effects of various extracurricular activities on the college-going behavior of White students. Using NELS:88 data, the researchers analyzed the impacts of student characteristics on attendance at a four-year institution as well as on attendance at an elite or non-elite institution. Kaufman and Gabler controlled for background variables in order to account for any spurious correlations between cultural capital and educational achievement and found that participation in extracurricular cultural activities has different causal effects on college-bound and elite-college bound students. Specifically, participation in tangible, hands-on activities such as music or art lessons improved the odds of attending college but not of attending elite colleges. Interestingly, student direct exposure to the arts did not significantly impact college or elite college attendance; however, museum going by parents – whether the students go or not – was significantly positively related to attendance at elite colleges, perhaps because of increased family cultural capital (Kaufman & Gabler, 2004).

Methodologies

The dependent variable in student college choice research can be conceptualized in various ways, and this definition impacts the choice of statistical model used. Typically, the dependent variable is dichotomous, but occasionally the student decision is represented by multiple categories. In the case of studies where the impact of explanatory variables on selectivity of college attended is examined, the dependent variable is sometimes represented by an approximately continuous measure. Logistic regression and classical linear regression models are the most often-used in studies of college student behavior, but some researchers have employed multinomial logit regression techniques, and one study of college selectivity employed ordered probit analysis (Hilmer, 2001). The following section details the various types of dependent variables used and the associated statistical models, first in terms of general college attendance and then with regards to selectivity of institution attended.

Dichotomous Dependent Variables

A dichotomous dependent variable in studies of general student college choice can be operationalized as either going to college versus not going to college or as going to a particular institution or institution type versus not going to that institution/institution type. The choice of dependent variable varies based on the researchers' interests and available data.

Christensen et al. (1975) and Kodde and Ritzen (1988) examined the impact of student background and academic characteristics and either cost or expectations for earnings/employment on students' decisions to enroll or not enroll in any type of higher education. Kodde and Ritzen (1988) used logit analysis, whereas Christensen et al. (1975) used probit analysis, possibly because it was developed earlier than logit analysis (Agresti, 1996; Long, 1997).

Perna (2000) used an expanded econometric model of college enrollment which included social and cultural capital variables to study influences on enrollment in a four-year university versus no enrollment in higher education. Perna (2000) explained her decision to ignore enrollment at two-year colleges because students consider different criteria when applying to two- versus four-year institutions; additionally, her focus was on minority under-representation in higher education, and this effect is worse in four-year universities. Kaufman and Gabler (2004) similarly examined impacts on attendance at four-year institutions versus no attendance in any higher education institution but did not give an explanation of why they chose to operationalize the dependent variable in this way. Perna (2000) used logistic regression analysis, and Kaufman & Gabler (2004) used probit analysis, possibly because theirs was a two-part study in which probit analysis was needed in the second stage.

DesJardins et al. (1999) and Wolniak & Engberg (2007) employed logistic regression analysis to study impacts on student attendance at particular universities. DesJardins et al. (1999) examined how a variety of student characteristics and institutional factors influenced the decision to attend a large, land-grant university. Wolniak & Enberg (2007) were interested in the effects of high school feeder networks on attendance at eight different private, four-year universities.

Dependent Variables with Multiple Categories

Two studies from an economic framework have used multinomial logit models to explain student college choice. Specifically, Fuller et al. (1982) and Venti and Wise (1982) adopted the random utility conceptualization of the multinomial logit model, which presumes that a student, weighing the benefits and costs of attendance at colleges in his or her choice set, chooses the option with the maximum utility value. Because these researchers incorporated information on the choices (colleges applied to), the conditional logit model was used. For example, Venti and Wise (1982) estimated a system of equations with three possible outcomes: did not apply to college; applied to a college of a certain quality and was admitted; applied to a college of a certain quality and was not admitted.

Educational researchers Perna and Titus (2004; 2005) studied influences on college choice using multinomial logit models as well. Perna and Titus (2004) examined the impact of state public policies, after accounting for student background and academic characteristics, on five outcomes: not enrolled in college; enrolled in an in-state, public two-year college; enrolled in an in-state, public four-year university; enrolled in an in-state, private four-year institution; and enrolled in an out-of-state institution. Perna and Titus's 2005 study investigated how social capital in terms of parent involvement impacted students' choices to not enroll in college, to enroll in a two-year college, or to enroll in a four-year institution. The researchers in both studies did not assume or test for any ranking, or ordering, of quality of outcome categories (e.g., that attending a four-year or private institution is more desirable than attending a two-year or public institution). In contrast to other studies of student college choice, however, Perna and Titus (2004; 2005) took into account the nested nature of educational data and used multilevel models, specifically the multinomial extension of the Hierarchical Linear Model.

Finally, Mahoney et al. (2003), in their examination of the impacts of extracurricular activities on post-secondary educational attainment, used a dependent variable with four categories, coded one through four: dropped out in high school; earned a high school equivalency degree; earned a high school diploma; and enrolled in higher education. Interestingly, the researchers treated this dependent variable as continuous and used ordinary regression analysis, which implies an underlying natural ordering of the categories, although this was never discussed.

Selectivity of Institution as a Dependent Variable

The selectivity level of an institution is depicted in student college choice studies as either discrete or approximately continuous (see footnote one). Discrete conceptualizations are almost always dichotomous and employ logistic regression methods. For example, Alexander et al. (1987) rationalized that because prospects for attending college were improving at the time of their study, more research was warranted on the diversity of college type attended by students. Furthermore, the researchers argued that categorizing college attendance as either in two- or four-year institutions is a proxy for selectivity because prospects for degree completion are greater in four-year institutions and career prospects are enhanced by four-year college attendance (Alexander et al., 1987). Alexander et al. (1987) used logistic regression analysis to determine how access to college type (two- versus four-year) was affected by background characteristics and academic resources.

Similarly, Hu and Hossler (1998) examined how student sensitivity to tuition and financial aid and background and academic characteristics impacted preference to attend a public or private postsecondary institution, arguing that attendance at a private institution could lead to improved educational attainment and career outcomes. The students' self-reported importance of attending a public versus private institution was used to construct the dichotomous dependent variable: if the importance given to attending a private school was greater than the importance given to attending a public school, the dependent variable was coded with a one; otherwise the dependent variable was coded with a zero. Logistic regression analysis was then used (Hu & Hossler, 1998).

Hilmer (2001) extended this rationale to define the selectivity dependent variable in his study as having three rank-ordered levels: attendance at a four-year college; attendance at a two-year college; and non-attendance in college. He reasoned that the natural ordering of this variable is realistic because of the lower costs at two-year colleges and the transfer costs associated with moving from a two-year to a four-year college. Therefore, to avoid the monetary and time costs associated with transferring, a student who thinks he or she is likely to graduate will start at the most desired institution: a four-year college or university. If a student is unsure of his or her abilities, he or she will start at a two-year college, and a student who is not likely to graduate will choose to work (Hilmer, 2001). To examine the appropriate specification of this presumably ordered dependent variable, Hilmer conducted his analysis using a multinomial logit, an ordered probit, and a bivariate probit model. Although the options used in Hilmer's study might be thought of as ordered, his rationale is couched within the economic theoretical framework of opportunity-cost analysis and maximization of utility value and is similar to the arguments underlying use of discrete choice multinomial models (Kennedy, 1998).

Some studies use the *U.S. News and World Report* (USNWR) ranking of colleges and universities as measures of selectivity. The USNWR rankings provide summary data on average ACT® and/or SAT scores, costs, and acceptance rates at postsecondary institutions and are therefore commonly thought of as a gauge for selectivity (Owings et al., 1998). Owings et al. (1998) used the four categories of the USNWR rankings to define his dependent variable. Specifically, colleges and universities can be classified as either national universities, national liberal arts colleges, regional universities, or regional liberal arts colleges, and the top 50 schools in each category are referred to as "Tier 1."

Owings et al. (1998) used logistic regression analysis on a sample of college enrollees to determine the influences on attending a Tier 1 national university versus attending any other four year university. Similarly, Kaufman and Gabler (2004) examined the influence of participation in various extracurricular activities on attendance at elite institutions, defined as the top 25 USNWR-ranked national universities and the top 25 national liberal arts colleges. This analysis was the second stage in their study, so they used probit analysis with sample selection to adjust the elite-college results for selection into the overall college-going sample (Kaufman & Gabler, 2004).

The most common measure of selectivity as an approximately continuous dependent variable is average SAT scores of either incoming students or of the total student body of the institution. Average SAT score is used as an indicator of institutional prestige partly because this type of selectivity is highly correlated with measures of tuition and with other indicators of prestige such as high levels of faculty training and of intellectual quality in campus life and graduates' attainments (Davies & Guppy, 1997; Hearn 1991). Additionally, an institution's prestige level is often inversely related to its level of openness, a proxy of which is average SAT score (Davies & Guppy, 1997; Hearn, 1991). When selectivity is presented as a continuous variable, the classical linear regression model is used. For example, Hearn (1991) examined academic and nonacademic influences on the selectivity of college attended, as measured by average SAT scores (or ACT® equivalents) for all students at the institution, and Davies and Guppy (1997) virtually replicated Hearn's 1991 study but defined selectivity as average SAT score for incoming students only.

Interestingly, McDonough and Antonio (1996), using a similar conceptualization of difficulty of admission as a proxy for prestige, also chose the average SAT score of the freshman class as the dependent variable but divided the variable into five categories corresponding to ranges of average SAT (i.e., 1100-1199). The researchers treated the dependent variable as continuous anyway and used linear regression analysis.

Hearn (1988) used cost of institution as another proxy for openness, or prestige, reasoning that colleges with more selective admission criteria tend to be those with higher costs. Hearn (1988) defined the dependent variable, undergraduate tuition and fees rates of institutions attended by high school graduates in his sample, as continuous[3] and therefore used linear regression with ordinary least squares estimation. Additionally, Hearn (1991) argued that institutions' per-student spending on educational programs is an indicator of school quality and analyzed this approximately continuous dependent variable using the linear regression method as well.

As evidenced by Heller's (2003) reanalysis of four NCES reports on college participation, the choice of statistical model can greatly impact the results of the analysis. Specifically, contradictory findings on the influences of family income and parental education level on college participation resulted when tabular versus multivariate analyses were used. In fact, results from the NCES multivariate analyses were in contrast to a large body of research in college access. Although the use of descriptive versus inferential statistics is not the specific issue at hand in the current study, Heller's (2003) findings provide an argument for the use of appropriate statistical techniques when results

[3] Tuition rates, rather than being a truly continuous measure, may instead be a non-continuous, limited outcome variable (Frone, 1997), but this distinction was not examined by Hearn (1988).

have policy implications. Only one study of influences on selectivity of college attended has used an ordered dependent variable paired with the appropriate ordered regression technique (Hilmer, 2001). Therefore, an analysis of possible differential results from the use of the classical linear regression model and regression techniques designed for categorical dependent variables is needed.

Chapter 3: Methodology

The current study examined the following research question: How does the use of alternative statistical approaches for analyzing a categorical dependent variable impact a model of the influences on selectivity of college attended? Based on previous research and data available from ACT®, Inc., variables were chosen as candidates for inclusion in a model of student college choice. The categorical dependent variable of selectivity of college was analyzed in SPSS version 19.0.0.1 with classical linear regression, multinomial logistic regression, and ordered logistic regression. The impact of adding predictor variables to the model was examined across regression method, and models were compared on the basis of overall significance, goodness-of-fit, predictive ability, behavior of predictors, and diagnostic analyses.

Data Sources

Data for this study come from ACT®, Inc. Three different datasets were merged to produce the final data source for analysis. Dataset 1 contains student-level information and was generated in part from a Student Profile Section students complete when they register to take the ACT® test. This 189-item questionnaire provides information about student background, high school experiences, and educational plans. Also included in Dataset 1 are ACT® test scores and information from high school transcripts such as grade point averages and courses taken.

Dataset 2 is an ACT®-generated questionnaire, the Institutional Data Questionnaire, that is sent to approximately 4,000 public and private colleges and universities. The 2005-2006 response rate was 69.1% (B. Reynolds, personal communication, Summer 2006). The responding institutions provide information on admissions requirements, student body composition, programs offered, and tuition costs. The third dataset contains student-level college enrollment information purchased from the National Student Clearinghouse ("the Clearinghouse") by ACT®, Inc. The Clearinghouse is a non-profit organization that keeps records on student enrollment for more than 3,300 postsecondary institutions enrolling about 93% of college students (National Student Clearinghouse, 2001).

Matriculation data from the National Student Clearinghouse were matched with the most recent Student Profile Section and ACT® test score data available, which is from the 2004-2005 academic year. Therefore, students in the current study filled out the questionnaire and took the ACT® in 2004-2005 and matriculated to college in the fall of 2005. Institutional-level data are from the 2005-2006 academic year, when these students would have been attending college.

The entire population of students taking the ACT® in the 2004-2005 academic year numbered 1,186,251. However, students electing to take the ACT® test may have self-selected on variables related both to the decision to take a college entrance exam and to college attendance. To circumvent this problem of endogeneity, I made use of state policies in place in Illinois and Colorado in 2004-2005 that required all high school students to take the ACT®. By limiting the population of ACT®-takers to students in Illinois and Colorado (N = 183,275), a more comprehensive picture of the student choice process is obtained by including all high school students, not just ones who presumably already had a disposition to attend college.

Furthermore, by studying all test takers rather than just those that matriculated to college, the problem of sample selection, or incidental truncation, is avoided. As previously noted, studying only the small sub-sample of all eligible college students who enroll limits the accuracy and usefulness of the findings for explaining the college choice process, as only part of the relationship between explanatory variables and college choice is studied. This restriction is primarily made due to the readily-available cross-section of post-secondary student enrollment numbers at any given point in time (Hossler & Stage, 1992; Toutkoushian, 2001). In fact, cross-sectional studies are especially prevalent in research on selectivity of college attended: Alexander et al. (1987), Davies and Guppy (1997), Hearn (1988; 1991), and McDonough and Antonio (1996) all limited their samples to include only college attenders. Although the current study is restricted to the final stage of the process, enrollment, and cannot capture the influences on the predisposition and search phases, by including college attenders and non-attenders, it is hoped that a more complete explanation of the final stage is obtained. This research is in line with Hilmer (2001) and Perna and Titus (2004; 2005) who included in their studies non-college attenders along with students who attended a two- or four-year institution.

Sample

The merged datasets resulted in 184,579 cases from Illinois and Colorado. However, this dataset contained 1,283 students with duplicate records, produced from the merging of Dataset 1 (the student-level test score and Student Profile Section questionnaire file) and the college attendance information from Dataset 3. Examination of the data showed that the majority of the duplicate files were identical except for the college enrollment variable (i.e., some students had two or more different colleges listed as having attended in 2005). Of the 1,283 students with duplicate records, 40 students had two of the exact same record; in this instance, one of the records was deleted. Of the remaining 1,243 students, 1,235 students had two colleges listed; 7 had three colleges listed; and one had four colleges listed.

Because of the nature of the National Student Clearinghouse dataset, the presence of multiple colleges for one student is not considered to be indicative of a transfer to a different institution. Rather, it is most likely that the students enrolled in more than one college during the fall of 2005. In fact, there were three commonly-occurring pairs of institutions, accounting for 28.6% of the non-identical duplicate records. In these instances, one institution was a two-year, “Open” (selectivity-level) institution and the other was a four-year, “Traditional” institution, both in the same state (one pair in Illinois, one pair in Colorado, and one pair in Wyoming). From this pattern, I determined that students enrolled in a four-year institution but also took courses at a nearby community college, a common practice of college students. Although it is not possible to know the reasons behind all duplicate records, especially those with three or four institutions listed, I reasoned that dual enrollment was a plausible explanation. Therefore, I deleted all but the record with the most selective institution attended for all duplicate cases. This produced a final merged dataset with $N = 183{,}275$ unique cases, equal to the number of ACT® test takers in Illinois and Colorado in 2004-2005.

After consideration of missing data, the dataset of $N = 183{,}275$ was reduced to $N = 91{,}632$. Details of this reduction are discussed in a later section entitled Missing Data. A sample size of 91,632 is still rather large; in fact, in this instance the dataset might be thought of as a population rather than a sample, since it includes all test takers (with complete data) in the two states. Large sample sizes are not only impractical due to computer processing requirements (Alexander et al., 1987; DesJardins et al., 1999), but they might result in erroneous interpretation of significance tests due to extremely large power (McDonough & Antonio, 1996). Therefore, data were reduced for my analysis.

As a guideline for data reduction, the primary studies of the enrollment stage used in my literature review on student college choice were consulted. Twelve out of the 15 studies examined reduced their sample sizes; the three that did not included one study of all freshman at eight selected institutions (Wolniak & Engberg, 2007) and two studies that utilized the population of students in prior longitudinal studies (Christensen et al., 1975; Mahoney et al., 2003). The remaining studies reduced their data for various reasons. Many chose a sample to fit their research interests, including, for example, only college attenders (Alexander et al., 1987; Davies & Guppy, 1997; Hearn, 1988, 1991), attenders at elite colleges (McDonough & Antonio, 1996), students of certain ethnicities (Kaufman & Gabler, 2004; Perna & Titus, 2005), or students who indicated intention to receive a bachelor's degree (Hilmer, 2001). Other studies attempted to match the population in an original dataset (Marsh, 1992; McDonough & Antonio, 1996), and half of the studies explicitly mentioned reducing the sample based on the presence of complete data (Alexander et al., 1987; Hearn, 1988, 1991; Hilmer, 2001; Perna & Titus, 2004, 2005). Two studies reduced their sample size but did not give a rationale for doing so (Fuller et al., 1982; Hearn, 1984).

The final sample sizes in these 15 studies ranges from 440 to 18,000 students. The mean is 7,104.3 students, but this is skewed upward by the presence of four studies with over 10,000 students. The median is 5,200 students. It is not common in educational research to find a sample size approaching 180,843 students[4]. Even the comparative

---

[4] One exception is DesJardins, et al. (1999), who also used the ACT® Student Profile Section. The researchers studied the college application rather than enrollment process, however, and reduced the ACT®

methods studies consulted for the present study, which come from the area of student retention research, have smaller sample sizes of 3,589 (Porter, 1999; all freshmen at one institution) and 947 (Dey & Astin, 1993; community college freshmen at institutions that returned a complete survey). Therefore, my data were also reduced to approximate more realistic research conditions in the relevant field.

The maximum likelihood (ML) estimation method used by two of the regressions in the present study requires larger sample sizes than estimation by ordinary least squares and relies on large-sample asymptotic normality (Garson, 2012a). Long (1997) suggests that at least 500 subjects is adequate. A minimum of 10 observations per explanatory variable is a commonly-recommend rule of thumb (Garson, 2012a; Long, 1997), as long as that number is at least 100 (Long, 1997). An alternate recommendation is a sample size of at least 30 times the number of parameters being estimated (Garson, 2012a), especially if the explanatory variables are highly collinear or there is little variation in the dependent variable. Using the median sample size ($N = 5{,}200$) from the 15 primary studies described above and the more stringent requirement of 30 cases per predictor variable would allow me to examine up to 173 predictor variables, which is more than enough to model the student college choice process. Therefore, I chose to reduce my complete Illinois and Colorado population of $N = 91{,}632$ by randomly selecting 5,200 cases (approximately 5.68%).

The random sample is comparable on relevant characteristics to the rest of the population of ACT® test takers in Illinois and Colorado with complete data (see Table 2).

---

data of over one million students to a sample of 110,491. The dependent variable of interest in DesJardins et al. (1999) was application at one particular institution, and the sample was composed only of ACT® test takers in adjacent states.

Table 2

Sample and Complete-Data Population Description

Mean Differences for Continuous Variables				
	Sample *n* = 5,200	Remainder of Complete-Data Population *n* = 86,432	*t*	*df*
ACT® Score	20.72 (5.05)[a]	20.78 (5.15)[a]	.900	5,868.95[b]
Income	53,373.08 (29,685.88)[a]	53,050.18 (29,684.47)[a]	-.762	91,630
Siblings	1.5 (1.36)[a]	1.52 (1.37)[a]	.674	91,630
GPA	2.95 (.690)[a]	2.96 (.703)[a]	.667	91,630
Total Extracurricular Participation	3.04 (2.29)[a]	3.05 (2.34)[a]	.471	91,630
Admissions Policy	1.3 (1.28)[a]	1.3 (1.29)[a]	-.297	91,630

Categorical Variable Percentages		
	Random Sample *n* = 5,200	Remainder of Complete-Data Population *n* = 86,432
White	3,481 (66.9%)	58,352 (67.5%)
Black	506 (9.7%)	8,394 (9.7%)
Hispanic	610 (11.7%)	10,139 (11.7%)
College Preparatory Track (took)	2,587 (49.8%)	43,004 (49.8%)

Advanced Courses (took)	2,939 (56.5%)	49,869 (57.7%)
Foreign Language (took 3+ years)	2,232 (42.9%)	37,460 (43.3%)
Attended College	3,296 (63.4%)	53,987 (62.5%)
Tested in Colorado	1,352 (26.0%)	22,352 (25.9%)
Tested in Illinois	3,848 (74.0%)	64,080 (74.1%)

[a] Standard deviations in parentheses. [b] Equal variances not assumed.

Variables

Potential variables for the model were defined based on previous student college choice research, specifically research on influences on selectivity level of college attended, and available data from ACT®. Following from the examples of the comparative methodological studies conducted by Dey and Astin (1993) and Dunn (1993), "The dependent variable is fit to a deliberately simple model, since our main goal is to investigate the comparative quality of data rather than to evaluate different models [of Army officer behavior]" (Dunn, 1993, p. 191). Similarly, Dey & Astin (1993) in their study from the field of student retention research state that "Since the purpose of this analysis is to explore the predictive capabilities of different statistical techniques, these variables are not designed to fit within any particular theoretical framework" (p. 575). Dey and Astin do note, though, that they chose variables shown by past research to predict retention. Therefore, although the variables considered for inclusion in my model have been shown to influence selectivity of college attended, mine is not intended to be a comprehensive explanation of the student college choice process. However, I did take into consideration the inclusion of variables that might be useful to students when attempting to increase their chances of attendance at selective institutions. The ACT® Student Profile Section questionnaire as well as the Institutional Data Questionnaire provide much more information than is used here that would be of interest to student college choice researchers in more fully describing the decision process.

This section describes my decision process for inclusion in the model of potential variables from the ACT® datasets. The analysis of missing data and assumptions of regression analysis are discussed in the next sections.

Admissions Policy

The dependent variable in this study comes from Dataset 2, the Institutional Data Questionnaire. Postsecondary institutions are asked to rate their admissions policy as either highly selective, selective, traditional, liberal, or open. ACT® suggests the following criteria for the categories, but the schools are not required to follow it: highly selective schools admit students from only the top 10% of their graduating class or above; selective schools admit from the top 25% of the graduating class of above; traditional schools admit from the top 50% and above; liberal schools admit from the top 75% of the class and above; and open schools accept everyone who applies.

For the purposes of this analysis, students who do not have college enrollment information in the merged dataset were coded with a “0,” for no college attended. College enrollment status comes from the National Student Clearinghouse (Dataset 3), and 39.8% of the students in the Illinois and Colorado population (N = 183,275) had missing enrollment data. It is not known whether these students did not enroll in college or are simply missing enrollment data. However, the enrollment rate of all 18- to 24-year-olds in degree granting postsecondary institutions in 2005 was 38.9% (Digest of Education Statistics, 2006), much lower than the enrollment rate in the present sample of 61%. Therefore, the assumption that the 39.8% of cases with missing enrollment data did not enroll in college and can be coded with a "0" for the dependent variable seems valid, as it under-represents the population percentage for that year (meaning that it likely does not include much missing enrollment data).

Students whose college code matched with an admissions policy of open were coded with a "1"; liberal admissions policies were coded "2"; traditional admissions policies were coded "3"; selective admissions policies were coded "4"; and highly selective admissions policies were coded "5"; thus creating a seemingly-ordered variable. The accuracy of the designation as an ordinal- or nominal-level variable will be tested during the main analyses (discussed below).

The percentage of students in the full dataset (N = 183,275) attending a college with a liberal admissions policy is strikingly small (.9%), indicating that "liberal" might not be a valid category. Therefore, to aid in analysis by increasing the chances of adequate cell counts in each level of the dependent variable, this category was collapsed into an adjacent category (Garson, 2011b). Furthermore, interpretation of results may be simplified slightly by modeling an outcome with five rather than six categories. In order to determine with which category to combine the liberal admissions policy responses, the percentages of the liberal institution classification and the two adjacent categories, open and traditional, that are two-year colleges were compared. The percentage of colleges classified as "liberal" that are also two-year institutions is 21.8%. The percentage of the lower-selectivity category, "open," that is two-year colleges is 96.1%, and the percentage of the category adjacent to “liberal” with higher selectivity, “traditional,” is 1.1%. Because 21.8% is closer to 1.1% (“traditional”) than it is to 96.1% (“open”), it was determined that liberal admissions policy institutions are more similar to traditional admissions policy institutions than to open institutions, and these two categories were merged. The new coding scheme is represented in Table 15.

Although different factors may influence the decision process for attending or not attending college than for attending various selectivity levels of college, limiting the sample to only college attenders may exclude a non-random portion of high school seniors. As previously noted, by including both college attenders and non-attenders, the problem of sample selection is avoided; furthermore, Hilmer (2001) and Perna and Titus (2004; 2005) included in their studies non-college attenders and college attenders alike, so there is a precedent for this method of defining the sample.

Student Background Characteristics

The most studied variables in student college choice research are ascriptive student background characteristics, specifically socioeconomic status and race. Sex was not shown to have any significant effects in all but one (Hearn, 1991) of the reviewed studies of influences on college attendance in general or on selectivity level of college attended; therefore, it was not included in my model.

SES. The majority of the student college choice studies reviewed defined socioeconomic status (SES) by including up to four separate measures in a model. Within the literature on selectivity of college attended, only Alexander et al. (1987), Davies and Guppy (1997), and Owings et al. (1998) defined SES as a composite measure. Hearn (1984; 1988; 1991), McDonough and Antonio (1996), McDonough et al. (1997) and

Hilmer (2001) included parent income, mother and father education level, and in some cases number of siblings as separate measures of SES.

One limitation of the data from ACT® is that no information on the commonly-studied variable of mother or father education level is included. Therefore, the two SES measures that are available, parents' income level and number of siblings, were included as two separate variables in the present model. Number of siblings is considered as a continuous[5] variable, following DesJardins et al. (1999) who coded the same ACT® Student Profile Section variable in this way, ranging from zero to nine in response to the question "How many brothers and sisters under 21 years of age do you have?" Additionally, Hearn (1988; 1991) treated his six-value (zero to five) sibling variable as continuous.

The ACT® Student Profile Section asks students to estimate the approximate total combined income of their parents before taxes in the last year. The students are given the option of choosing from 10 categories of income level ranges. This is an imperfect variable, to say the least. First of all, students' estimates of parents' income may or may not be accurate. Furthermore, the response categories have overlapping values: for example, category "1" is $18,000 to $24,000, and category "2" is $24,000 to $30,000 (the value of $24,000 is present in two categories). Additionally, the categories are not equal – the first category has a range of $18,000 (the description is "less than $18,000"), the next four categories have a range of $6,000, the next category has a range of $8,000, the next

[5] Indeed, number of siblings is a count measure, making it discrete and not continuous (Kirk, 1999; Lane, 2003). However, precedent has been set in student college choice research (see for example DesJardins et al., 1999; Hearn, 1988, 1991) for treating this explanatory variable as continuous rather than breaking it into dummy variables, and this is the procedure I chose to follow in the current study.

one has a range of $10,000, the next two have ranges of $20,000, and the final category is "greater than $100,000."

Although precise interpretation of unit changes in the explanatory variable is difficult due to varying interval sizes, conceptualizing the income variable as approximately continuous yields more information than breaking the original variable into numerous dichotomous variables as DesJardins et al. (1999) did. Furthermore, the income variable as it is measured by ACT® has numerous categories that represent a quantitative attribute, or a variable that is measured in terms of numbers (Lane, 2003) – annual income in dollars, in the present study. Dummy coding in CLRM is most appropriate for discrete variables which represent qualitatively different categories that do not fall along a quantitative continuum (Tabachnick & Fidell, 2001). For example, a significant, positive relationship between an approximately continuous income variable and the dependent variable (or the logit of the dependent variable) would indicate, roughly, that as income increases, selectivity of college increases. However, the finding of a significant, positive relationship between a dichotomous variable for "$24,000 to $30,000" could mean that students within that income range attend/are more likely to attend more selective colleges than students either above or below that range (depending on the reference category). More information about impact of income is gained from the former interpretation.

Additionally, there is precedent in the research for treating categorically-coded income variables as continuous (Hearn, 1988, 1991; Hilmer, 2001; McDonough & Antonio, 1996; and Perna & Titus, 2005), and it is common to treat variables as continuous when the number of categories is greater than seven (Tabachnick & Fidell, 2001), as in the present analysis. An alternate method, employed by Toutkoushian (2001) with his predictor variable of a student report of family income, is to use the midpoint value of each income category as a representative value. In such a manner, an approximately continuous manifest variable is created to represent the underlying quantitative construct of income.

Despite the flaws in the construction of the income variable, because there is no information in the ACT® data on mother and father education level, an SES measure shown to be significantly related to college attendance in many studies, exclusion of the income variable would leave only number of siblings (a lesser-studied variable) as the only measure of SES in the model. Income is included in the model as an approximately continuous explanatory variable with 10 values representing the midpoints of family income class as reported on the SPS questionnaire. The unit of measurement used in the regression analyses is $1,000.00 (rather than $1.00) in order to facilitate a more realistic interpretation of the results (a $1,000.00 increase in annual income is more likely than a $1.00 annual increase).

Race. Second in influence to, and far less significant than, the impact of SES on college attendance is race or ethnicity of the student. The most commonly-studied ethnicities in student college choice research are White, Black, and Hispanic. Although ACT® asks the students to categorize themselves as one of eight races/ethnicities in response to the question "Which phrase best describes your racial/ethnic background?," this variable was recoded to the four categories of White (reference category), Black, Hispanic (including the original categories of Mexican American/Chicano/Latino and

Puerto Rican/Cuban/other Hispanic), or "Other" (American Indiana/Alaskan Native, Asian American/Pacific Islander, other ethnicity, or multiracial).

Student Academic Characteristics

Behind ascriptive student characteristics, student academic characteristics are also commonly studied in student college choice research. Most often this is measured by composite SAT® or other standardized test score, high school performance (grade point average and/or class rank), and to a lesser extent course-taking patterns. Participation in extracurricular activities can be considered in this category, although fewer studies incorporate this variable. Wolniak and Engberg (2007), in their study of influences on attendance at eight particular private postsecondary institutions, used a composite measure of academic achievement in order to avoid multicollinearity, but all other studies of the student college choice process included separate measures in their models. In general, greater academic achievement is positively associated both with demand for postsecondary education and type of institution attended.

Standardized test score. For obvious reasons, composite ACT® rather than SAT® or any other standardized test score was used in the current study. This is an approximately continuous[6] measure with a possible range of 1 to 36 (ACT®, Inc., 2009). The composite score is the average of the four sub-tests in the areas of English, mathematics, reading, and science (ACT®, Inc., 2009).

High school performance. Class rank is not studied as much in the student college choice research as is high school grade point average (GPA). The ACT® Student Profile Section asks students to classify their class rank in one of four quartiles (top quarter,

[6] Although a truly continuous measure takes on an infinite range of real values (Kirk, 1999; Frone, 1997), Tabachnick and Fidell (2001) describe a continuous variable as one that takes on any value within the range of the scale; thus, ACT® score may be considered as an approximately continuous variable.

second quarter, etc.). Besides being an imprecise, coarsely-categorized, self-reported measure, this variable is also closely related to the dependent variable, admissions policy, by definition[7]. Therefore, class rank was not included in the present model.

In addition to the self-reports from the student questionnaire, ACT® also obtains data from student high school transcripts on grade point averages in six subject areas: English, mathematics, social studies, natural science, foreign language, and arts. These GPAs were "unweighted," meaning that honors or advanced placement courses were not awarded more points than traditional courses. The average GPA was calculated from these numbers, resulting in an approximately continuous[8] measure with a possible range of zero to four. However, it might be argued that this is an imperfect measure because grades can reflect differences in high schools (instruction as well as expectations and grading procedures) rather than true academic ability (Perna & Titus, 2005). Additionally, GPA is related to class rank (as class rank is determined by relative GPAs), which is the criterion by which the dependent variable is classified. However, GPA and class rank are not directly related: class rank is relative to the performance of other students in a high school. Therefore, and because some measure of high school performance is included in almost all models of student college choice, the GPA variable was included in the present model.

Course-taking patterns. The measurement of student academic characteristics with course-taking pattern is less well-defined in the literature; however, it is an important area of study in which students can potentially control their behaviors to

---

[7] Selectivity level is determined by institutions' policies for admitting students based on their class rank (i.e., highly selective institutions admit students from only the top 10% of their graduating class or above).
[8] Tabachnick and Fidell (2001) use grade point average as an example in their definition of a continuous variable as one that takes on any value within the range of the scale.

maximize their potential for attending selective colleges. Conceptualizations vary anywhere from the pursuit of a college preparatory track (Davies & Guppy, 1997; Hearn, 1988, 1991; Hilmer, 2001), enrollment in a particular subject such as foreign language (McDonough & Antonio, 1996; McDonough et al., 1997), number of years taken in certain subjects (DesJardins et al., 1999), or highest level of a subject taken (Perna & Titus, 2004; 2005).

The ACT® Student Profile Section asks students to describe the program of high school courses taken as either business/commercial,vocational-occupational, college preparatory, or other/general. The questionnaire also provides information on whether students took advanced placement, accelerated, or honors courses (included in the same question, not as separate measures) in five different subject areas. Four studies (Davies & Guppy, 1997; Hearn, 1998, 1991; Hilmer, 2001) found that following a college preparatory track was significantly related to selectivity of postsecondary institution attended; however, no studies reviewed for the current study used a variable indicating whether a student took advanced courses.

There are drawbacks to such course-taking pattern variables, however. Perna and Titus (2004) argue that participation in a college preparatory track is an unreliable measure of academic preparation. This could be due to the variability among high school courses offered and definitions of the different tracks. Given the coding of the program-of-studies ACT® questionnaire variable, this argument is plausible in the current study. When self-reporting their course program, students might have trouble categorizing their pattern, especially if they took courses in more than one category. Although the number of advanced courses taken is a less subjective measure (students can easily count up the number of courses designated as advanced or honors that they took), this variable does not get away from the critique that instruction and curriculum in honors courses varies across high schools.

Because of its significant influence in previous studies, the program of courses taken variable was chosen for inclusion in the present research. Although examination of the effects of each of the four tracks would be interesting, to keep the model simple and to mirror existing research, I recoded this question into a dichotomous variable representing whether the student took a college preparatory track or not. No studies have explored the effects of the other programs of high school courses (such as a business or vocational track), however, so these variables might be of interest for inclusion in future studies of influences of selectivity of college attended. Additionally, the examination of the novel advanced placement/ accelerated/ honors course variable could be of potential interest to student college choice research; therefore, a dichotomous variable representing whether a student took any advanced courses during high school was created and included as well (rather than five separate subject honors variables, in keeping with the goal of building a simple model).

The ACT® Student Profile Section questionnaire provides categorized information on number of years students studied certain subjects. For example, a response of “1” indicates a half-year of study in a subject, “2” indicates one year of study in a subject and so on through “8,” with “9” indicating that the student did not take any courses in the subject. ACT® also has information on courses taken from students’ high school transcripts. From this, the actual number of years a student studied English, mathematics, social studies, natural science, foreign language, and arts is provided.

Additionally, the specific combination of courses taken can be determined from the high school transcripts. ACT® has calculated four non-ordered, categorical variables which describe which particular English, math, social science, and natural science courses a student has taken. As an illustration, the social science course pattern variable has six categories: U.S. History, World History, American Government, and other history; Other combination of four or more years of social science; U.S. History, World History, and American Government (defined as the minimum core); Other combination of three or three and a half years of social science; Less than three years of social science; and No social science course information reported.

The information provided on number of years and specific courses taken in each subject could potentially inform student college choice research about influences of course-taking patterns on college attendance. However, such detail is beyond the scope of this study and would complicate the construction of a simple model for use with comparison of statistical methods; therefore, the number of possible variables should be reduced from the 10 (four of which having from six to nine response categories) that ACT® provides. The literature does not give consistent guidance for definition of this academic characteristic. McDonough and Antonio (1996) and McDonough et al. (1997) included one approximately continuous variable, years of foreign language. DesJardins et al. (1999), using the ACT® Student Profile Section, included five dichotomous variables – preparation (as defined by a minimum number of years taken) in all subject areas except for arts. Perna and Titus (2004; 2005) suggest that an appropriate measure of the quality of high school curriculum is the highest level of coursework taken and reason that the hierarchical sequence of math courses is the most well-defined and often-used in research. Therefore, their measure of academic characteristics of students is defined by four different math course-taking patterns.

Of these studies that included some measure of years or combination of courses taken, only McDonough and Antonio (1996), McDonough et al. (1997) and Perna and Titus (2004; 2005) were limited to the final stage of the student college choice process, enrollment. DesJardins et al. (1999), although finding significant impacts for college preparation in five academic subjects, conducted their study to examine the application process, specifically at one large, land-grant institution. McDonough and Antonio (1996) found that number of years of foreign language is significantly positively related to selectivity of college attended, within a subset of selective colleges used in their study. However, the methodology used in that study may not be appropriate: average freshman SAT® scores, converted to a five-level categorical variable, were used as the dependent variable in a linear regression analysis. McDonough et al. (1997), though, also found a significant impact of the dichotomous variable of three or more years of foreign language taken on attendance at a selective college in a binary logistic regression analysis.

Perna and Titus (2005), although not explicitly studying selectivity of college as the dependent variable, found that the math patterns of either Algebra I and Geometry, Algebra II, or another advanced math course as the highest level course taken were all significantly related to enrollment in either two or four year institutions (as opposed to non-enrollment in college). Similarly, Perna and Titus (2004) found significant effects on all five of their dependent variable categories (in-state public two- and four-year institutions, in-state private four-year universities, and any type of out-of-state institution) for the math course-taking patterns of Algebra I with Geometry, Algebra II, or any advanced math.

Based on these findings, a dichotomous variable for three or more years of foreign language taken[9] was included in the present model along with an unordered categorical variable representing math course-taking pattern as defined by ACT®: less than three years of math; any combination of three to three and a half years of math; any combination of four or more years of math; minimum core (Algebra I, Algebra II, and Geometry); the pattern of Algebra I, Algebra II, Geometry, and Trigonometry or another advanced math course; or the pattern of Algebra I, Algebra II, Geometry, Trigonometry, and Calculus or another advanced math course.

Extracurricular activities. The ACT® Student Profile Section questionnaire asks students whether they participated in 16 different extracurricular activities at any time during high school. There is some evidence to indicate that total participation in extracurricular activities is positively related to college attendance but that the effects of different types of activities vary (Hearn, 1988, 1991; Hilmer, 2001; Kaufman & Gabler, 2004; Mahoney et al., 2003; Marsh, 1992). Therefore, it would be informative to examine the effects of the activities included in the ACT® data on selectivity of institution attended.

In the portion of their study that included influences of extracurricular activities on matriculation at an elite college, Kaufman and Gabler (2004) found that, for White students only, participation in school government, yearbook/newspaper, and hobby clubs

[9] Self-reported foreign language data from the ACT® Student Profile Section rather than high school transcript data will be used to calculate this dummy variable due to a discrepancy in missing data percentages (only 28.5% of cases did not report years of foreign language taken, whereas 71.1% of the high school transcripts were missing this information).

had significant impacts – but the impacts depended on the gender of the student. Hearn (1991) found that participation in student government was significantly related to selectivity of college attended as measured by average SAT® score but that participation in drama, debate, journalism, or departmental clubs was not significantly related.

Alternately, Marsh (1992) did not find a significant impact of participation in student government, but his dependent variable was attendance at any type of college. Marsh (1992) did find significant influences of sports, music, hobby clubs, yearbook, honor society, and community service. DesJardins et al. (1999), using the ACT® Student Profile Section data but examining a different dependent variable than the present study, found significant effects on application to a particular university for participation in music, belief system clubs, and community service but not in varsity athletics.

These findings contain similarities and inconsistencies regarding some form of all of the 16 individual extracurricular variables in the ACT® data. Although these variables may yield useful information about specific extracurricular influences on the student college choice process, estimation of separate models for all specific extracurricular dichotomous variables on each of three statistical methods being compared in this study would be superfluous, and including all of them in the same model would result in a complicated model and possible multicollinearity. However, I do want to provide potentially useful information to students interested in attending selective institutions. A total participation variable, such as was found by Hilmer (2001) to be significantly negatively related to selectivity of college attended, would provide more information than a dichotomous variable – knowledge of the influence of number of activities is arguably more useful than knowledge of the influence of merely participating in any extracurricular activity – and was therefore opted for in the present model. This variable is a count measure, ranging from 0 to 16 extracurricular activities participated in during all four years of high school, but it was treated as approximately continuous rather than discrete, using the same rationale that was used for the sibling and income variables.

Once determination of the most appropriate statistical method for analyzing my categorical dependent variable is made, however, the influences of specific extracurricular activities can be examined in a future study. Results of such an analysis may further inform student college choice research by helping students to more specifically prepare during high school for attendance at selective institutions. Table 3 lists references of student college choice research that has examined the influence of predictor variables used in my study.

Table 3

References for Predictor Variables

Predictor Variable	Studies That Have Examined the Variable
Family Income (measured separately)	Christensen et al., 1975; Davies & Guppy, 1997; Hearn, 1984, 1988, 1991; DesJardins et al., 1999; Hilmer, 2001; Mahoney et al., 2003; McDonough & Antonio, 1996; McDonough et al., 1997; Teachman, 1987; Toutkoushian, 2001
Number of Siblings (measured separately)	Christensen et al., 1975; Davies & Guppy, 1997; DesJardins et al., 1999; Hearn, 1988, 1991; Mahoney et al., 2003; McDonough & Antonio, 1996; Teachman, 1987
Race	Akerhielm et al., 1988; Alexander et al., 1987; DesJardins et al., 1999; McDonough & Antonio, 1996; McDonough et al., 1997; Perna, 2000; Perna & Titus, 2004, 2005
ACT® Score	DesJardins et al., 1999
GPA	Christensen et al., 1975; Hand & Prather, 1987;

	Hearn, 1991; McDonough & Antonio, 1996; McDonough et al., 1997; Venti & Wise, 1982
College Preparatory Track	Davies & Guppy, 1997; Akerhielm et al, 1998; Hearn, 1988, 1991; Hilmer, 2001
Advanced Courses	N/A
Foreign Language	DesJardins et al., 1999; McDonough & Antonio, 1996; McDonough et al., 1997
Math Course-Taking Pattern	Perna & Titus, 2004, 2005
Extracurricular Participation	DesJardins et al., 1999; Hearn, 1988, 1991; Hilmer, 2001; Kaufman & Gabler, 2004; Mahoney et al., 2003; Marsh, 1992

Data Screening

Before running any analyses, the data must be examined for accuracy and missing values (Tabachnick & Fidell, 2001). An examination of the ranges of variables did not indicate any mis-entered values outside of the acceptable range. However, missing data are a problem. Table 4 shows the percentage of total cases (*N*=183,275) with missing values on each potential variable described above. Procedures employed to deal with this missing data are discussed below, followed by the results of screening for outliers in the variables and for normality in the continuous and approximately continuous variables.

Table 4

Missing Data Percentages

Variable	Percent Missing
Admissions Policy	3.2
Income	34.0
Siblings	25.7

Race	18.0
ACT® Score	0.0
GPA	21.3
College Preparatory Track	28.6
Advanced Courses	28.7
Foreign Language	28.5
Math Course-Taking Pattern	13.4
Total Extracurricular Participation	28.6

Note. *N* = 183,275.

Missing Data

The nonlinear model assumed by logistic regression requires a full set of data (Garson, 2012a). In fact, the default option for most statistical computer programs is to delete cases with missing values when running any regression analysis: for example, in logistic regression, SPSS provides only for LISTWISE deletion of cases with missing data, using the remaining full dataset to calculate logistic parameters (Garson, 2012a; Tabachnick & Fidell, 2001). Rather than let the computer program determine which cases to delete, the researcher should examine the dataset for any patterns in missing data and make the decision about how to proceed. By doing this at the start of the analysis, the researcher controls what happens to the data and can describe the properties of cases with missing data.

Data that are missing at random do not pose as much of a problem as non-randomly missing data, especially if 5% or fewer of the cases have randomly missing data points (Tabachnick & Fidell, 2001). Deleting cases with missing data may be a good option if they are few in number and seem to be a random subsample of the total sample;

however, one should not assume data are missing at random. The researcher should test for patterns in missing data, particularly on variables thought to be related to the variable with missing data (Tabachnick & Fidell, 2001). Non-randomly missing values can affect the generalizability of results if they are deleted. Even randomly missing data that are scattered through cases and variables can result in a substantial loss of subjects (Tabachnick & Fidell, 2001).

There are options for dealing with missing data. First of all, if missing values are concentrated in a certain variable or variables and those variables are not crucial to the model or are highly correlated with each other, the variables themselves can be excluded from analysis (Tabachnick & Fidell, 2001). If cases are missing data on different variables, deletion of cases is reasonable when there is a random pattern and if only a few cases are affected. However, if the pattern is non-random, methods that preserve all cases are preferred (Tabachnick & Fidell, 2001). One option for preserving cases is to estimate missing data on continuous variables through any of a variety of methods. Mean substitution should be avoided unless the proportion of missing values is extremely small. Regression methods may be implemented if the data contain good variables for predicting the missing values on a variable, but regression is less desired than principled methods such as expectation maximization (EM) (Peng, Harwell, Liou, & Ehman, 2006). EM methods are the preferred approach when the data are missing randomly; when there is a pattern to missing values, EM methods can be incorporated into multiple imputation for best results. When missing data are scattered over variables and there are no variables with a substantial number of missing values, a missing data correlation matrix can be used. Regardless of the method of imputation, repeating analyses with and without missing data is recommended, especially if the dataset is small and the proportion of missing values is high (Tabachnick & Fidell, 2001).

Most of the student college choice studies reviewed either did not mention how missing data were treated or stated simply that missing cases were deleted (or, alternately, that only cases with complete data on all variables were used; see Alexander et al., 1987; Hearn, 1988, 1991; Hilmer, 2001). Perna and Titus (2005), however, dealt with their non-randomly missing data through the method discussed by Cohen and Cohen in 1983 (cited in Perna & Titus, 2005 and Tabachnick & Fidell, 2001). Perna and Titus (2005) created a variable for tendency to have missing data, calculated as the number of explanatory variables on which data are missing for a particular case. Then, mean scores on continuous explanatory variables were imputed and the new missing data variable was used as another variable in the analysis. Peng and Naegle Nichols (2003) first ran their analysis with subjects who completed all questions on a survey. However, to check whether the missing data on the dependent variable were at random, the researchers imputed missing values on all variables through the EM method and analyzed the full dataset. Analyses with and without the missing data were compared for differences in order to test the validity of using only complete cases (in this case, results were similar except for smaller measures of association due to the lowered statistical power in the smaller sample resulting from listwise deletion; Peng et al., 2006; Peng & Naegle Nichols, 2003).

Percentages of missing data on my variables range from 3.2% to 34.0%. The proportion of cases with any missing values (in other words, without complete data) is 50% (much greater than the 5% recommended by Tabachnick & Fidell for safe deletion, if data are missing at random). Therefore, examination of the patterns of missing data is warranted in order to determine how to proceed (Tabachnick & Fidell, 2001). As recommended in Peng et al. (2006) and Tabachnick and Fidell (2001), *t*-tests were run on continuous or approximately continuous variables with the grouping variable as missing versus nonmissing data. Additionally, crosstabulations of categorical variables and variables with missing data were examined for differences in percentages among groups of the categorical variables (Tabachnick & Fidell, 2001). Missing data patterns for each data source are described below, concluding with a discussion of the construction of the final sample.

Admissions policy. Information on the dependent variable, admissions policy, comes from the Institutional Data Questionnaire (IDQ). Few institutions (110 out of 1,446, or 7.6%) are missing admissions policy data. Because the assumption that the dependent variable is continuous may not be accurate, imputation is not an appropriate option, whether the data are missing randomly or non-randomly. It is the institutions and not the students who provide information for this variable; therefore, examination of patterns of missing data would include institutional characteristics. The IDQ provides information on average ACT® score of entering freshmen. As noted, average college entrance exam scores of freshmen is often considered a proxy for selectivity (Davies & Guppy, 1997; Hearn, 1991; McDonough & Antonio, 1996).

The average ACT® score of entering freshmen for institutions with missing admissions policy data is 19.7 (*SD* = 2.84), and the average ACT® score for institutions providing admissions policy data is 20.4 (*SD* = 3.25). This difference is not statistically significant: $t(815) = -1.836$, $p = .067$. These results indicate that institutions not providing admissions policy information are likely similar in selectivity level to institutions that did provide this information. However, results may be affected by the differences in group size (*n* = 81 institutions with missing admissions policy data that had average freshman ACT® score data and *n* = 736 institutions with admissions policy data and average freshman ACT® score data). Although 7.6% of institutions are missing admissions policy data, the total number of cases (students) with missing admissions policy data is only 3.2%. Because fewer than 5% of student cases are missing admissions policy data (*n* = 5,937), deletion is not likely to significantly impact the analysis (Tabachnick & Fidell, 2001). Regression analyses will exclude these cases anyway, so deletion prior to analysis is recommended (resulting in a sample size of 177,338 students).

ACT® Student Profile Section. Besides the income variable, the variables with the highest percentage of missing data are from ACT® Student Profile Section (SPS) questions about high school experiences (track, advanced courses, years of foreign language, extracurricular participation). A variable representing tendency to have missing data on the relevant questions about high school experience (1 = missing all SPS high school experience data) was created in order to examine through *t*-tests group differences on other variables of interest between students who filled out any portion of this section of the SPS and the 26.7% (out of 177,338) who filled out none of the questions about program of high school courses followed, advanced courses and foreign language courses taken, and extracurricular activities participated in. Results of this missing data analysis on the reduced sample with no missing admissions policy data (*N* = 177,338) for continuous or approximately continuous variables are presented in Table 5. For the purpose of this analysis – to efficiently compare group differences – the ordered categorical variable of admissions policy was treated as continuous, although it has a limited range of values, and the use of parametric statistics such as calculation of means and *t*-tests is not advised by Stevens (1946). Meaningful interpretation of the means calculated on the admissions policy variable for use with the *t*-test is therefore not appropriate.

Table 5

Continuous Variable Means for Missing and Non-Missing SPS Data

	SPS High School Experience Questions			
	Missing All (n = 47,253)	Completed Any (n = 130,085)	t	df [a]
ACT® Score	19.48 (5.2)	20.46 (5.2)	35.09***	84,343.6[b]
GPA	2.89 (.731)	2.94 (.718)	9.44***	35,015.4[b]
Income	48,652.26 (29,850)	52,355 (29,853)	8.7***	117,003
Siblings	1.58 (1.42)	1.54 (1.39)	-2.45*	131,768
Admissions Policy	1.14 (1.25)	1.23 (1.29)	13.3***	85,822.9[b]

Note. Standard deviations in parentheses.
[a] Degrees of freedom vary due to missing values on the continuous variables. [b] Equal variances not assumed.
*p < .05, ***p < .001

Additionally, the relationship between missingness on the SPS questions and the categorical variables of race and a collapsed admissions policy variable (attended college or not) was examined through a crosstabulation. Results are presented in Table 6.

Table 6

Categorical Variable Percentages for Missing and Non-Missing SPS Data

Variable	SPS High School Experience Questions	
	Missing All (n = 47,253)	Completed Any (n = 130,085)
White	15,580 (33.0%)	81,864 (62.9%)
Black	4,382 (9.3%)	12,575 (9.7%)
Hispanic	3,852 (8.2%)	15,280 (11.7%)
Other	2,700 (5.7%)	13,785 (10.6%)
Missing Race Data	20,739 (43.9%)	6,581 (5.1%)
Attended College	26,739 (56.6%)	77,616 (59.7%)
Did Not Attend College	20,514 (43.4%)	52,469 (40.3%)

There does appear to be a nonrandom pattern for missing data on the SPS questions relating to high school experiences. Students who chose not to fill out this section have significantly lower ACT® scores and grade point averages, attend less selective postsecondary institutions, and come from families with lower annual incomes and more siblings than students who partially or fully completed this SPS section. Additionally, more White students than Black students filled out some or all of this section, but the portion of students who did not fill out any of the high school experience section also did not complete the race/ethnicity question at a much higher rate than students who did answer the high school experience questions. Therefore, due to the high amount of missing race data, conclusions about the racial make-up of SPS section non-completers cannot be made with confidence. The proportions of students attending college are similar for the two groups, though.

Although there seems to be a pattern for non-completion of the high school experience SPS questions, there is no agreed-upon recommended method for dealing with non-ignorable missing data (Peng et al., 2006). One recommendation is to eliminate these SPS variables (track, advanced courses, years of foreign language, extracurricular participation); however, this is only prudent when the variables are not critical to the analysis (Tabachnick & Fidell, 2001). In the present model, though, elimination of these variables would significantly reduce the explanatory power of the model for student college choice behavior. Additionally, it might be reasoned that the 26.7% of the students who did not fill out the high school experience questions of the SPS did not fill out any, or much of, the SPS questionnaire, including the questions about income, number of siblings, and race (since those variables also have high percentages of missing data). Dropping those variables would only leave ACT® score, GPA, and math course-taking pattern in the model. Such a reduced model would exclude variables found in existing research to be important for explaining student college choice, which could result in omitted variable bias.

Because the pattern of missingness is non-random, excluding these cases will affect the generalizability of results. Deleting students with missing data on the SPS questions may result in distorted sample values on ACT® score, GPA, income, number of siblings, racial make-up, and selectivity of college. It is important to note, however, that the *t*-tests are extremely sensitive due to the large sample size: significant results may have been found when the true difference is actually small. When the analyses are run, though, these cases will be excluded by the statistical computer package SPSS. Therefore, deleting the cases with missing high school experience SPS data is the best option, with the caveat that the resulting sample may not be representative of the total sample.

SES variables. Even after deleting cases that are missing admissions policy and all high school experience SPS data, the remaining dataset (N = 130,085) contains cases with missing income and sibling data. Patterns for missingness on income, the more personal of the SES questions (which students might be reluctant to answer), were explored. Students who did not fill out the income question of the ACT® Student Profile Section (14.0% of the remaining sample) have significantly lower ACT® scores and grade point averages, have more siblings, and attend institutions of a lower selectivity level than students who provided income information. Again, the *t*-tests might be overly sensitive to recognizing small differences due to the large, un-evenly split sample size of 130,085.

There does not appear to be a large difference between percentage of students providing and not providing income data for White and Hispanic students, but more Black students completed the income question than did not complete it. Students who did not provide income data also did not provide race data at a higher rate (8.6%) than students who did provide income data (4.5%). Overall, though, the percentages of students within racial groups providing and not providing income information is similar, especially in comparison to the racial differences seen between those completing and not completing the high school experience SPS questions.

Students who completed the income question attended college at a higher rate (60.6%) than students who did not provide income data (53.9%). Results of this missing data analysis for continuous, approximately continuous, and categorical variables are presented in Tables 7 and 8.

Table 7

Continuous Variable Means for Missing and Non-Missing Income Data

	Income Variable			
	Missing (n = 18,229)	Completed (n = 111,856)	t	df [a]
ACT® Score	20.12 (5.5)	20.52 (5.2)	9.121***	23,942.8[b]
GPA	2.85 (.788)	2.95 (.706)	14.739***	18,787.6[b]
Siblings	1.57 (1.41)	1.53 (1.39)	-2.935**	126,117
Admissions Policy	1.15 (1.31)	1.25 (1.28)	9.630***	24,208.1[b]

Note. Standard deviations in parentheses.
[a] Degrees of freedom vary due to missing values on the continuous variables. [b] Equal variances not assumed.
**p* < .05, ***p* < .01

Table 8

Categorical Variable Percentages for Missing and Non-Missing Income Data

Variable	Missing (n = 18,229)	Completed (n = 111,856)
White	11,430 (62.7%)	70,434 (63.0%)

Black	1,449 (7.9%)	11,126 (9.9%)
Hispanic	1,988 (10.9%)	13,292 (11.9%)
Other	1,793 (9.8%)	11,992 (10.7%)
Missing Race Data	1,569 (8.6%)	5,012 (4.5%)
Attended College	9,825 (53.9%)	67,791 (60.6%)
Did Not Attend College	8,404 (46.1%)	44,065 (39.4%)

Based on the results of the *t*-tests, the group of students not providing income data appear to be of lower academic ability (as measured by ACT® score and GPA) and attend institutions of lower selectivity level than the students who did provide income information. Income values could be imputed; however, the recommended method – EM – is only valid for data missing at random (Tabachnick & Fidell, 2001). Tabachnick and Fidell (2001) only recommend mean substitution if the percent of cases is "very small" (p. 66). Although the word "very" is not defined, SPSS Missing Values Analysis examines variables with greater than 5% of cases missing (Tabachnick & Fidell, 2001), so 5% might be a reasonable guideline. The income variable has 14.0% of its data missing, so mean substitution is likely not a viable option. Alternately, dropping a variable is a method of dealing with missing data if the percentages are large and the variables are not critical to the analysis. Although 14.0% is not very small, it is also not large. Furthermore, income is a critical variable, so dropping it is not an option. Therefore, the cases with missing SES (income and sibling) values were deleted.

Race. After deleting cases with missing SES values, 4.4% of the remaining sample of 111,362 did not have race data. Similar analyses as above were run to determine the pattern of missingness. Results are show in Tables 9 and 10.

Table 9

Continuous Variable Means for Missing and Non-Missing Race Data

	Race Variable			
	Missing (n = 4,940)	Completed (n = 106,422)	t	df^{a}
ACT® Score	20.81 (5.8)	20.51 (5.2)	3.527***	5,305.6[b]
GPA	2.93 (.753)	2.95 (.704)	-1.995*	4,587.7[b]
Income	52,152.83 (30,735.24)	52,393.81 (29,806.52)	-.539	5,379.2[b]
Siblings	1.57 (1.47)	1.53 (1.39)	1.939	5,351.9[b]
Admissions Policy	1.23 (1.34)	1.25 (1.28)	-.837	5,360.9[b]

Note. Standard deviations in parentheses.
[a] Degrees of freedom vary due to missing values on the continuous variables. [b] Equal variances not assumed.
*p < .05, ***p < .001

Table 10

Categorical Variable Percentages for Missing and Non-Missing Race Data

Variable	Missing (n = 4,940)	Completed (n = 106,422)
Attended College	2,812 (56.9%)	64,721 (60.8%)
Did Not Attend College	2,128 (43.1%)	41,701 (39.2%)

Students who did not complete the ACT® Student Profile Section question on race had statistically significantly higher ACT® test scores than students who did provide race data, but the grade point averages of the question non-completers were significantly lower than those of question completers. No other significant differences for continuous variables were found between students who provided race data and those who did not. The mixed results regarding academic ability (as measured by ACT® score and GPA) between the two groups are echoed by the similar college attendance rates of those who complete the race question and those who did not (60.8% and 56.9%, respectively). Because of the diminished importance, as compared to SES, of race or ethnicity as a student ascriptive characteristic for predicting college attendance, in addition to the fact that only 4.4% of the sample has missing race data, these cases were deleted.

High school transcripts. In addition to the Institutional Data Questionnaire and ACT® Student Profile Section, the source of missing data could also be high school transcripts. Information for the high school grade point average (GPA) and math course-taking pattern variables comes from this source. Deleting cases with missing admissions policy and all ACT® Student Profile Section data resulted in a sample of 106,422, of which 10.2 and 3.4 percent, respectively, have missing data on the two high school transcript variables of GPA and math course-taking pattern. In order to determine any patterns for missing transcript information, *t*-tests and percentages were examined for a combined group of cases with any of or none of the two transcript variables.

The high school transcript data do not appear to be missing at random. Students with no high school transcript data have significantly lower ACT® scores, attend institutions of lower selectivity levels, and come from lower income households with more siblings than do students with high school transcript data. Additionally, more White students have present transcript data and more Black students have missing transcript data, and students with transcript data attend college at a much higher rate than those without transcript data (62.3% and 48.2%, respectively). Results are presented in Tables 11 and 12.

As noted in the above section regarding the income variable, the recommended EM method for imputation is only valid for data missing at random (Tabachnick & Fidell, 2001), which does not apply to the GPA variable, and mean substitution is either not recommended at all (Peng et al., 2006) or is only recommended if the percent of cases is "very small" (Tabachnick & Fidell, 2001, p. 66). Using the 5% guideline, the high school transcript variables exceed this value with 10.3% of their data missing, so mean substitution (on the continuous GPA variable) is likely not a viable option. Alternately, dropping a variable is a method of dealing with missing data if the percentages are large and the variables are not critical to the analysis. Again, the percentage (10.3%) is not very small, but it is also not large, and at least the variable of GPA is critical to the analysis. Therefore, the cases with missing high school transcript data were deleted, even though this results in a sample that likely has distorted ACT® score, SES, and admissions policy means.

Table 11

Continuous Variable Means for Missing and Non-Missing High School Transcript Data

	High School Transcript Variables			
	Missing (n = 11,004)	Present (n = 95,418)	t	df[a]
ACT® Score	18.61 (4.9)	20.73 (5.2)	-42.591***	13,924.9[b]

Income	47,577.79 (30,284.6)	52,949.21 (29,700.9)	-17.927***	106,420
Siblings	1.63 (1.47)	1.52 (1.37)	7.471***	13,300.5[b]
Admissions Policy	.89 (1.13)	1.29 (1.29)	-34.942***	14,521.5[b]

Note. Standard deviations in parentheses.
[a] Degrees of freedom vary due to missing values on the continuous variables. [b] Equal variances not assumed.
***$p < .001$

Table 12

Categorical Variable Percentages for Missing and Non-Missing High School Transcript Data

Variable	Missing (n = 11,004)	Present (n = 95,418)
White	6,105 (55.5%)	64,109 (67.2%)
Black	1,665 (15.1%)	9,370 (9.8%)
Hispanic	1,915 (17.4%)	11,316 (11.9%)
Other	1,319 (12.0%)	10,623 (11.1%)
Attended College	5304 (48.2%)	59,417 (62.3%)
Did Not Attend College	5,700 (51.8%)	36,001 (37.7%)

Final sample. After methodologically comparing missing data on variables in my model and deciding how to deal with these data, the resulting sample size was 95,418.

My initial missing data check of the sample with no missing admissions policy data (*N* = 177,338) resulted in the deletion of the 47,253 cases that did not fill out all of the SPS high school experience questions. The cases were grouped this way in order to examine differences between students who chose not to answer any of the questions and those who responded to at least some of the questions. However, due to the fact that SPSS will not analyze cases with any missing data (through LISTWISE deletion in regression analyses), the rest of the cases with any missing SPS high school question values (the track, advanced courses, foreign language, and extracurricular variables) were also deleted. This resulted in a final sample size of 91,632 complete cases.

Although deletion of such a large number of cases (50% of my initial sample) affects the generalizability of results due to non-random patterns of missingness, this method is in line with much of the existing student college research: most of the studies reviewed that specifically mentioned treatment of missing data deleted all cases with any missing values, without any discussion of randomness or non-randomness to the pattern of missingness (Alexander et al., 1987; Hearn, 1988, 1991; Hilmer, 2001). In order to facilitate comparisons between my sample and samples in other studies, a description of the cases with complete data versus those that were excluded from analysis (had any missing data on variables in the model) is presented in Table 13.

As noted, a random sample of 5,200 cases was selected for the final analysis, to more closely approximate research in the student college choice field.

Table 13

Description of Cases with Complete Data on All Variables and of Cases Excluded from Analysis

Mean Differences for Continuous Variables				
	Complete Data (*n* = 91,632)	Missing Any Values[a] (*n* = 91,643)	*t*	*df*[b]
ACT® Score	20.78 (5.15)[c]	19.78 (5.28)[c]	-41.084***	183,163.99[d]
Income	53,068.5 (29,684.48)[c]	50,683.12 (30,424.75)[c]	-11.753***	48,454.48[d]
Siblings	1.52 (1.37)[c]	1.57 (1.42)[c]	7.084***	85,660.46[d]
GPA	2.96 (.702)[c]	2.90 (.741)[c]	-14.291***	104,983.58[d]
Total Extracurricular Participation	3.05 (2.33)[c]	2.90 (2.43)[c]	-10.319***	71,360.37[d]
Admissions Policy	1.30 (1.289)[c]	1.11 (1.257)[c]	-30.536***	177,026.18[d]

Categorical Variable Percentages		
	Complete Data (*n* = 91,632)	Missing Any Values[a] (*n* = 91,643)
White	61,833 (67.5%)	39,458 (43.1%)
Black	8,900 (9.7%)	8,569 (9.4%)
Hispanic	10,749 (11.7%)	8,821 (9.6%)
Other	10,150 (11.1%)	6,825 (7.5%)

Missing Race Data	-	27,970 (30.5%)
College Preparatory Track (took)	45,591 (49.8%)	17,614 (19.2%)
Missing Track Data	-	52,445 (57.2%)
Advanced Courses (took)	52,808 (57.6%)	20,653 (22.5%)
Missing Advanced Course Data	-	52,617 (57.4%)
Foreign Language (took 3+ years)	39,692 (43.3%)	16,381 (17.9%)
Missing Foreign Language Data	-	52,195 (57.0%)
Attended College	57,283 (62.5%)	53,009 (57.8%)
Missing Admissions Policy Data	-	5,937 (6.5%)
Tested in Colorado	23,704 (25.9%)	23,701 (25.9%)
Tested in Illinois	67,928 (74.1%)	67,942 (74.1%)

[a] These cases were excluded from the final sample. [b] Degrees of freedom vary due to missing values on the continuous variables. [c] Standard deviations in parentheses. [d] Equal variances not assumed.
***$p < .001$

Outliers in the Variables

An outlier, or a case with an extreme value on a variable or a combination of variables, can distort the results of an analysis. In addition to leading to Type I and Type

II errors, due to the disproportionate impact on the regression coefficient, outliers affect the generalizability of results (Tabachnick & Fidell, 2001).

Tabachnick and Fidell (2001) recommend labeling as univariate outliers cases with very large continuous variable standardized scores that are disconnected from the other standardized scores. For large samples, standardized scores beyond ±3.29 indicate possible outliers (Tabachnick & Fidell, 2001). Standardized scores on the sibling variable extend to 5.52, but visual examination of the data, histograms, and boxplots indicated that these cases were not disconnected from the rest of the sample. Similarly, total extracurricular participation standardized scores ranged to 5.66 but were not disconnected from the rest of the data. The lowest standardized average GPA scores were -3.84 but were also connected to the sample.

Because scores in the smaller category of an unevenly split dichotomous variable are more influential than those in the larger category, causing the same effect as an outlier, dichotomous variables with greater than 90-10 splits between categories should be deleted (Tabachnick & Fidell, 2001). Bar graphs and frequencies of the dichotomous variables College Preparatory Track, Advanced Courses, and Foreign Language revealed no splits greater than 14%; therefore, these variables were retained for analysis.

In order to detect multivariate outliers, the Mahalanobis distances can be examined for each case using the χ^2 distribution. Additionally, leverage is a statistical measure related to Mahalanobis distance that can identify multivariate outliers, although it is not based on a χ^2 distribution. When case number is regressed on the 10 predictor variables and one dependent variable in my data, as recommended by Tabachnick and Fidell (2001), the critical Mahalanobis distance value ($df = 11$, $p < .001$) is 31.264, and the critical value for leverage (calculated from the critical χ^2 value for Mahalanobis distance; α = .001) is .0062. Twenty-eight cases (.54%) are above the critical Mahalanobis value, and 25 cases (.48%) are above the critical value for leverage when CLRM is used.

Because the cases with extreme standardized scores are connected to the rest of the cases, I determined that they are not univariate outliers and rather products of skewed distributions. Additionally, the results of the multivariate outlier analysis are not convincing enough to label any cases as multivariate outliers. Accuracy of data entry has been confirmed; therefore, all cases are determined to be legitimate parts of the sample.

Normality of Variables

Although not required for analysis, normality in the predictors may enhance power and result in more stable solutions for logistic and for classical linear regression (Garson, 2012a, 2012b; Tabachnick & Fidell, 2001). Additionally, interpretation of linear relationships is enhanced when predictor variables have the same underlying distribution as the dependent variable. Therefore, multivariate normality is desirable in CLRM (Garson, 2012b), although OLS estimation is relatively robust to minor deviations from normality (while ML estimation is even more robust against moderate deviations in normality; Garson, 2011b).

The dichotomous predictor variables College Preparatory Track, Advanced Courses, and Foreign Language have a bimodal distribution, and the categorical predictor variables Race and Math Course-Taking Pattern have a multimodal distribution. The continuous (and approximately continuous) predictor variables and the ordered dependent variable were examined for normality. Tabachnick and Fidell (2001) recommend assessing normality of variables by graphical methods, rather than formal inference tests of skewness and kurtosis, when the sample is large. Even minor deviations from normality may cause rejection of the null hypothesis of a normal distribution in large samples (Garson, 2011b; Tabachnick & Fidell, 2001).

Bar graphs of the income and sibling SES variables showed a slightly concave distribution for income (highest frequencies in the tails) and a positively skewed, leptokurtic distribution for number of siblings. Although significance tests were not run, the skewness (.257) and kurtosis (-1.21) values for the income variable are not far from zero, the value for a normal distribution. The income variable is slightly platykurtic, as evidenced by the bar graph shape of a near-plateau and the negative kurtosis value. Interestingly, the highest frequency is for the annual income category of $100,000 or greater. The sibling variable, however, is positively skewed (1.56) and appears leptokurtic (3.71). The value of "1" sibling has the highest frequency.

The composite ACT® score variable appears to be approximately normally distributed, as indicated from histograms and skewness (.326) and kurtosis (-.524) values. The average GPA variable is slightly negatively skewed (-.495), but kurtosis value (-.366) and a visual examination of the histogram indicate an approximately normal shape. Of the predictor variables graphed with a histogram, total extracurricular participation varies the most from a normal distribution with a skewness value of .913 (positively skewed) and a kurtosis value of 1.18 (leptokurtic).

The dependent variable, selectivity of college attended, is moderately positively skewed (.629) and slightly platykurtic (-.722), although the bar graph revealed a mesokurtic shape. The value of "0" (did not attend college) has the highest frequency.

Analysis

Assumptions

The classical linear regression model (CLRM) makes assumptions about the way data are generated, and when all of these assumptions hold, OLS is the best linear unbiased estimator (Kennedy, 1998; Kreiberg, n.d.; Long, 1997). Logistic regression has few restrictions, allowing for a variety and complexity of datasets that can be analyzed with this method (Tabachnick & Fidell, 2001). Assumptions regarding the distributions of predictors (normality, homogeneity of variance, normally distributed errors) are not required. The dependent variable does not need to be linearly related to the explanatory variables, but it must be discrete (Garson, 2012a; Tabachnick & Fidell, 2001). Although many of the restrictive assumptions of ordinary least squares regression do not apply in logistic regression, other practical issues and limitations must be considered (Garson, 2012a; Tabachnick & Fidell, 2001). Table 14 highlights the formal assumptions of the classical linear, the multinomial logistic, and the ordered logistic regression models, and the following sections discuss these considerations and assumptions. Results of evaluation of the formal assumptions are given in Chapter 4.

Table 14

Assumptions Required for Three Regression Models

	CLRM	Multinomial Logistic	Ordered Logistic
Zero mean error $E(\varepsilon_i) = 0$	X		
Non-stochasticity $Cov(\varepsilon_i, x_i) = 0$	X		
Normality of errors $\varepsilon_i \sim N(0, \sigma^2)$	X		

Homoskedasticity $Var(\varepsilon_i) = \sigma^2 < \infty$	X		
Non-autocorrelation $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$	X	X	X
Linearity in the DV	X		
Linearity in the logit of the DV		X	X
Absence of multicollinearity	X	X	X
Large samples		X	X
Adequate expected cell frequencies		X	X
Parallel regression			X

Proper sampling and specification of the model. Many of the assumptions of all types of regression can be met by properly specifying the model and selecting the sample (Garson, 2012a; Peng, Lee, & Ingersoll, 2002; Tabachnick & Fidell, 2001). The goal of regression is to properly predict an outcome using the most parsimonious model ("Logistic Regression," 2002; R. Toutkoushian, personal communication, Fall 2005; Tabachnick & Fidell, 2001). The inclusion or deletion of variables in a model may cause changes in magnitude and/or direction of parameters. If a relevant variable is excluded, non-stochasticity may result, and the variance it shares with an included variable may be wrongly attributed to that included variable, thus inflating the error term. However, variance may be wrongly attributed to included irrelevant variables if they share a common variance with included relevant variables. High correlations of irrelevant variables with other explanatory variables lead to high standard errors for these explanatory variables (Garson, 2012a). Ideally, all included explanatory variables will be strongly correlated with the dependent variable but not with other explanatory variables in order to avoid problems of endogeneity and multicollinearity (Tabachnick & Fidell, 2001).

There are many alternative student choice models from which to choose. I attempted to build an appropriate model based on the existing student college choice research. As noted, I am not trying to guide my research with a particular theory, such as human or cultural capital or economic theory. In addition to the recommendations of constructing a model based on previous research or a theoretical framework, it is recommended to include too many variables than to risk excluding an important variable, which might result in the omitted variable bias (R. Toutkoushian, personal communication, Fall 2005). Unfortunately, ACT® does not provide information on parental education level, resulting in the exclusion of this important variable from my model.

Another consideration is the selection of the sample. Restricting the range in the sampling of cases such that the dependent variable is only observed if certain conditions are met may lead to deflated correlations, or incidental truncation, as only a truncated portion of the relationship between the predictors and the dependent variable is observed (R. Toutkoushian, personal communication, Fall 2005; Tabachnick & Fidell, 2001). By including both college attenders and non-attenders, my sample largely avoids the sample selection problem because a non-random portion of high school seniors (those who did not attend college) is not excluded.

The CLRM assumption of non-stochasticity, or independence of a predictor and the error term (Kreiberg, n.d.), may also be violated if an unobserved variable is responsible for selection into the sample and is also related to the dependent. Random sampling of the entire population of interest is desired for this assumption to be met. When the sample is random, observations are independent from each other, and not only are the assumptions of non-stochasticity and non-autocorrelation met, but the assumption of a binomial distribution for logistic regression is robust as well (Peng, Lee, & Ingersoll, 2002).

Had my data been made up of only students who chose to take the ACT®, an unobserved variable may have caused students to self-select into this population, and this variable might likely be related to the dependent variable of college attendance. However, by making use of state policies in place in Illinois and Colorado in 2004-2005 that required all high school students to take the ACT®, the problem of endogeneity is avoided. The primary issue with selectivity bias that may lead to endogeneity in the current study is the presence of a large amount of missing data. Half of the population of Illinois and Colorado test takers was eliminated from the final sample due to missing data, and it is unknown whether an unobserved variable influences both the tendency to have missing data and the dependent variable of selectivity of college attended.

Distribution of errors. One assumption of the classical linear regression model (CLRM) necessary for the desirable properties of the ordinary least squares (OLS) estimator to hold is that the population errors have a mean of zero (Kreiberg, n.d.). Additionally, there must be no relationship between the population error and each of the predictor variables (Garson, 2012b; Kreiberg, n.d.). Together, these assumptions refer to the concept of exogeneity. If regressors are correlated with the error term, they are referred to as endogenous, a condition which leads to biased coefficients estimated through OLS. This assumption cannot be tested statistically, as the population regression line is not known for sample data. In order for a predictor variable and the population error to be independent of each other (exogenous), variables not included in the equation should not affect the outcome variable and should not be correlated with included variables. Additionally, the outcome variable should not be an influence on one or more explanatory variables (Garson, 2012b; R. Toutkoushian, personal communication, Fall 2005).

The distribution of the error term in CLRM is assumed to be normal, with a mean of zero and a variance of σ^2 for all values. In other words, the residuals of prediction should be normally distributed around each predicted score (Tabachnick & Fidell, 2001). This normality of errors assumption is particularly important in small samples, but non-normality of errors – often caused by outliers (Kreiberg, n.d.) – is essentially inconsequential in large samples (Garson, 2012b; Kreiberg, n.d.). The normality assumption can be tested in various manners, including examination of residuals plots, histograms of standardized residuals, normal probability plots (a P-P Plot), and skewness and kurtosis values (Garson, 2011b; Tabachnick & Fidell, 2001).

Homoskedasticity, or the assumption made by the CLRM that the error term has the same variance, σ^2, in each observation, is more often violated in cross-sectional rather than time-series data (Kreiberg, n.d.). Heteroskedastic data could indicate an interaction between a measured predictor variable and one not included in the model or the presence of skewed predictor variables and will result in biased standard errors (Garson, 2012b;

Kreiberg, n.d.). This assumption can be checked by viewing scatterplots of various combinations of standardized and unstandardized observed and predicted values.

Typically a problem of time-series data, non-autocorrelation refers to the concept that error terms should not be correlated, also referred to as independence of observations (Garson, 2012b, 2011b; Kreiberg, n.d.). This assumption that error terms are independent, or that each response comes from a different, unrelated case, applies to both classical linear and logistic regression. Essentially, this requirement of independent sampling prevents the use of matched samples or repeated measures designs, such as time-series data. If subjects provide multiple observations at different time points, usual regression procedures are inappropriate because of correlated errors, but special adaptations are available in statistical packages to handle non-independent data (Garson, 2012a; Tabachnick & Fidell, 2001).

Linearity. First and foremost, the CLRM is a linear procedure, meaning that the dependent variable is a linear function of a specific set of explanatory variables and an error term (Garson, 2012b; Greene, 2000; Kennedy, 1998; Kirk, 1999; Long, 1997). When the assumption of linearity is violated, R^2 will underestimate the variance explained, and the coefficients will underestimate the importance of the variables. Substantial violation of the linearity assumption, as in the case of a dichotomous dependent variable, renders the results virtually unusable. For this reason, the dependent variable in CLRM is assumed to be continuous, interval-level data (Garson, 2012b).

Although use of a dichotomous outcome variable with CLRM is not recommended, researchers often use ordinal-level data to approximately meet the assumption of linearity, as minor violations will not substantially affect the interpretation of the results (Garson, 2012b). This is based primarily on a conceptualization of the observed discrete dependent variable as a manifestation of a latent continuous variable (Harwell & Gatti, 2001; Kennedy, 1998; McKelvey & Zavoina, 1975; Winship & Mare, 1984). The appropriateness of assuming a linear relationship between the predictors and an ordered categorical dependent variable can be informally inspected through scatter plots of each predictor variable with greater than two levels against the outcome variable and through a plot of the residuals against the predicted values (Garson, 2011b). Additionally, if the standard deviation of the residuals exceeds the standard deviation of the dependent variable, this is often considered an indicator of nonlinearity in CLRM (Garson, 2011b).

Logistic regression assumes a linear relationship between continuous predictors and the logit transformation of the dependent variable (Tabachnick & Fidell, 2001). When this assumption is violated, logistic regression will underestimate the degree of relationship of the predictors to the dependent and will lack power (Garson, 2012a). The Box-Tidwell approach is one of the simplest methods for testing this assumption (Garson, 2012a; Tabachnick & Fidell, 2001). In this approach, the crossproducts (interactions) of each continuous explanatory variable and its natural logarithm are calculated as new variables and then added to the logistic regression model. If an interaction term is significant, then the assumption is violated and there is nonlinearity in the logit (Garson, 2012a; Tabachnick & Fidell, 2001).

Absence of multicollinearity. All varieties of multiple regression are sensitive to extremely high correlations (r =.70 or greater) among predictor variables (Tabachnick & Fidell, 2001). Multicollinearity does not change the estimates of the coefficients, but it does affect their reliability (Garson, 2012a). Standard errors of the logit coefficients become inflated when there is a high multicollinearity among the explanatory variables (Garson, 2012a, 2012b), resulting in small *t*-ratios. In such a case, the effects of the predictors cannot be separated and are unreliable.

There is no universally-accepted test for multicollinearity, although tolerance values and other collinearity diagnostic statistics such as variance inflation factors can be examined in CLRM. Additionally, there are indicators for which to watch. If two similar variables have widely different impacts on the dependent variable, multicollinearity might be an issue (R. Toutkoushian, personal communication, Fall 2005). Additionally, a large R^2 value and significant F tests of the classical linear regression model and exceedingly high standard errors for parameter estimates but few significant *t*-tests of coefficients flag possible multicollinearity (Garson, 2012a, 2011b; Tabachnick & Fidell, 2001). If multicollinearity is suspected, the source can be determined through various means for discrete and continuous predictors (Garson, 2012a; Tabachnick & Fidell, 2001), and one of the correlated variables can be dropped to both test and correct for the problem (R. Toutkoushian, personal communication, Fall 2005).

Large samples. Estimation through maximum likelihood requires a large sample size as it relies on large-sample asymptotic normality (Garson, 2012a). Extremely large parameter estimates and standard errors may be produced or failure of convergence may result when the ratio of cases to variables is too small (too many cells with no cases). Additionally, perfect separation of outcome groups by a discrete predictor renders a ML solution impossible and is caused by either too small a sample or too many variables relative to few cases in one outcome. Indication of these problems is evident by extremely high parameter estimates and standard errors or failure to converge and is therefore determined after the analysis is run (Garson, 2012a; Tabachnick & Fidell, 2001).

Adequate expected cell frequencies. Goodness-of-fit tests that compare full or partial models to a hypothetical, perfect model assume adequate expected cell frequencies since they are based on differences between observed and expected frequencies (Tabachnick & Fidell, 2001). Such a goodness-of-fit analysis may have little power if the expected frequencies are too small. In order to test for this, expected cell frequencies for all pairs of discrete variables including the outcome variables should be evaluated. It is recommended that researchers run crosstabulations to assure that the requirement of no more than 20% of the cells have a count less than five and that all expected frequencies are greater than one (Garson, 2012a; Tabachnick & Fidell, 2001).

Parallel regression. In the ordered logistic regression model used in this study, the odds ratio, or slope (β), of a predictor is assumed to be the same across all response categories (Liu, 2007; Long, 1997; Peng & Naegle Nichols, 2003). This is the assumption of proportional odds, or parallel regression (Long, 1997). The proportional odds model has invariance to the choice of response category (Agresti, 1996), so there is a common effect of any explanatory variable regardless of response value. Proportionality is restrictive because equal log-odds for each explanatory variable are assumed across all cumulative logits, but in many research situations this is a reasonable assumption (O'Connell et al., 2008) which can be tested for (Long, 1997; O'Connell et al., 2008). A score test, or Lagrange multiplier test, evaluates how the log likelihood would change if the constraint that all coefficients are equal across all regressions was removed (Long, 1997; O'Connell et al., 2008; Peng & Naegle Nichols, 2003). If the test result is significant, the assumption has been violated, and a multinomial logit model can be used (Garson, 2011a; Long, 1997).

Regressions

Three models for each of the following regression types were estimated using SPSS version 19.0.0.1: 1) classical linear regression with OLS estimation; 2) multinomial logistic regression with ML estimation; and 3) ordered logistic regression with ML estimation. Roughly simulating the methodology of Hearn (1991), a model including only student background characteristics (income, siblings, and race) was run for each of the regression methods. A second model which added the approximately continuous student academic characteristics of ACT® score and GPA was also run for each regression type. With the exception of the unordered categorical race variable, this second model consisted of only approximately continuous variables. This most closely fits the type of model typically employed in classical linear regression in which the predictors are assumed to be continuous, interval variables (although it is common to use ordinal data; Garson, 2012b). The third, full model estimated for each regression included all 10 explanatory variables. Definitions of the variables included in the models are given in Table 15.

Regression equations for all three statistical models are presented below. Overall model significance tests (analysis of variance *F*-tests and likelihood ratio χ^2 tests) and the Pearson χ^2 and the deviance-based inferential tests of goodness-of-fit were performed for each appropriate equation. In addition to parameter estimates, effect sizes (R^2), measures of association (R_L^2, Somers' D_{yx}), and classification tables were obtained and are reported in the next chapter.

Table 15

Definitions of Variables

Variable Name	Definition
Dependent Variable	
Admissions Policy (Y)	0 = Did not attend any postsecondary institution 1 = Attended an Open institution 2 = Attended a Liberal or Traditional institution 3 = Attended a Selective institution 4 = Attended a Highly Selective institution
Student Background Characteristics	
Income (X_1)	An approximately continuous variable representing the midpoints of the annual family income categories. This is measured in units of $1,000 for the regression analyses. $9,000 = Less than $18,000 $21,000 = About $18,000 to $24,000 $27,000 = About $24,000 to $30,000 $33,000 = About $30,000 to $36,000 $39,000 = About $36,000 to $42,000 $44,000 = About $42,000 to $50,000 $55,000 = About $50,000 to $60,000 $70,000 = About $60,000 to $80,000 $90,000 = About $80,000 to $100,000 $100,000 = More than $100,000
Siblings (X_2)	The number of siblings under 21 years of age. An approximately continuous (count) measure ranging from 0-9
Race/Ethnicity (X_3)	An unordered categorical variable 1 = Other (American Indiana/Alaskan Native, Asian American/Pacific Islander, other ethnicity, or multiracial) 2 = African American/Black (non-Hispanic) 3 = Mexican American/Chicano/Latino/Puerto Rican/Cuban or other Hispanic 4 = Reference category: Caucasian American/White (non-Hispanic).
Student Academic Characteristics	
ACT® Score (X_4)	A continuous measure ranging from 1-36
GPA (X_5)	A continuous measure ranging from 0.0-4.0 The average of a student's grades in six subject

	areas.
College Preparatory Track (X_6)	A dichotomous variable 1 = student described his/her program of courses as college preparatory
Advanced Courses (X_7)	A dichotomous variable 1 = student indicated that he/she took any advanced placement (AP), accelerated, or honors courses during high school
Foreign Language (X_8)	A dichotomous variable 1 = student took 3 or more years of any combination of foreign language
Math Course-Taking Pattern (X_9)	An unordered categorical variable 1 = any combination of three to three and a half years of math 2 = any combination of four or more years of math 3 = Algebra I, Algebra II, and Geometry (minimum core) 4 = Algebra I, Algebra II, Geometry, and Trigonometry or another advanced math course 5 = Algebra I, Algebra II, Geometry, Trigonometry, and Calculus or another advanced math course 6 = Reference category: less than three years of math
Total Extracurricular Participation (X_{10})	An approximately continuous (count) variable ranging from 0-16. The number of extracurricular activities a student participated in during high school.

Classical linear regression. This model is represented by the following equations and is estimated with OLS. As required by SPSS, all variables were entered as continuous covariates; the categorical Race and Math Course-Taking Pattern variables were converted into dummy variables and the greatest category (4 = Caucasian American/White and 6 = less than three years of math) omitted.

Equation C1

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{3.1} X_{3.1} + \beta_{3.2} X_{3.2} + \beta_{3.3} X_{3.3} + \varepsilon$$

Equation C2

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{3.1} X_{3.1} + \beta_{3.2} X_{3.2} + \beta_{3.3} X_{3.3} + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

Equation C3

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{3.1} X_{3.1} + \beta_{3.2} X_{3.2} + \beta_{3.3} X_{3.3} + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_{9.1} X_{9.1} + \beta_{9.2} X_{9.2} + \beta_{9.3} X_{9.3} + \beta_{9.4} X_{9.4} + \beta_{9.5} X_{9.5} + \beta_{10} X_{10} + \varepsilon$$

Multinomial logistic regression. The equation for the multinomial logistic regression as an odds model is described as the odds of outcome *m* versus outcome *n*, given the explanatory variables. In order to identify the model, the effect of the predictor variables (coefficients) on one of the outcome categories, chosen arbitrarily, is constrained to equal zero. In SPSS, the default is to constrain the effects of the highest category (the one with the largest number; "highly selective" in the present study). Therefore, the resulting coefficients are the effects on the odds of being in the given category of the dependent variable, where the odds represent the probability of being in that category versus the probability of being in the reference category (Garson, 2012a). In the present study, with the reference group, highly selective, being coded "4," the equations are:

Equation M1.

$$\text{logit}\left[\pi\left(Y = m \mid x_1, x_2, \ldots x_p\right)\right] = \ln\left(\frac{\pi\left(Y = m \mid x_1, x_2, \ldots x_p\right)}{\pi\left(Y = 4 \mid x_1, x_2, \ldots x_p\right)}\right) = \alpha_m + \left(-\beta_m X_1 - \beta_m X_2 - \beta_m X_3\right) + \varepsilon$$

Equation M2.

$$\text{logit}\left[\pi\left(Y = m \mid x_1, x_2, \ldots x_p\right)\right] = \ln\left(\frac{\pi\left(Y = m \mid x_1, x_2, \ldots x_p\right)}{\pi\left(Y = 4 \mid x_1, x_2, \ldots x_p\right)}\right) = \alpha_m + \left(-\beta_m X_1 - \beta_m X_2 - \beta_m X_3 - \beta_m X_4 - \beta_m X_5\right) + \varepsilon$$

Equation M3.

$$\text{logit}\left[\pi\left(Y = m \mid x_1, x_2, \ldots x_p\right)\right] = \ln\left(\frac{\pi\left(Y = m \mid x_1, x_2, \ldots x_p\right)}{\pi\left(Y = 4 \mid x_1, x_2, \ldots x_p\right)}\right) = \alpha_m + \left(-\beta_m X_1 - \beta_m X_2 - \beta_m X_3 \ldots - \beta_m X_{10}\right) + \varepsilon$$

Because of the constraint SPSS imposes, the intercept (α_4) and coefficients (βs) in the equation for $Y = 4$ are zero, which results in the equation for the response category of interest (m) being estimated with only the coefficients for m (for $m = 0$ to 3). The multinomial logistic regression constitutes one complex multinomial logistic model in which *J-1* pairs of categories, four in the present study, are estimated simultaneously (Agresti, 1996; Long, 1997).

Ordered logistic regression. Although the choice between the use of a logit or probit curve is largely one of convenience and convention, the estimate of the parameter is not comparable between the models because they are on different scales (Agresti, 1996; Kennedy, 1998; Long, 1997; Maddala, 1983). Only one study within the student college choice research literature conceptualized the dependent variable as ordered (Hilmer, 2001); therefore, there is not a lot of guidance based on previous research. I used ordered logistic regression in order to keep the parameter estimates on the same scale as the multinomial logistic regression also used in my study. Additionally, the logit model offers relatively simple interpretation of logit coefficients as odds ratios (Agresti, 1996; Long, 1997). Furthermore, Liu (2007) offers guidance for ordinal regression in SPSS and uses a logit link; using the same logit link in my study will facilitate interpretation of this guidance.

In SPSS, the ordered logistic regression model assumes a latent variable and is expressed as follows (Liu, 2007):

Equation O1.

$$\text{logit}\left[\pi\left(Y \le j \mid x_1, x_2, \ldots x_p\right)\right] = \ln\left(\frac{\pi\left(Y \le j \mid x_1, x_2, \ldots x_p\right)}{\pi\left(Y > j \mid x_1, x_2, \ldots x_p\right)}\right) = \alpha_j + \left(-\beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3\right) + \varepsilon$$

Equation O2.

$$\text{logit}\left[\pi\left(Y \le j \mid x_1, x_2, \ldots x_p\right)\right] = \ln\left(\frac{\pi\left(Y \le j \mid x_1, x_2, \ldots x_p\right)}{\pi\left(Y > j \mid x_1, x_2, \ldots x_p\right)}\right) = \alpha_j + \left(-\beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 - \beta_4 X_4 - \beta_5 X_5\right) + \varepsilon$$

Equation O3.

$$\text{logit}\left[\pi\left(Y \le j \mid x_1, x_2, \ldots x_p\right)\right] = \ln\left(\frac{\pi\left(Y \le j \mid x_1, x_2, \ldots x_p\right)}{\pi\left(Y > j \mid x_1, x_2, \ldots x_p\right)}\right) = \alpha_j + \left(-\beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 \ldots - \beta_{10} X_{10}\right) + \varepsilon$$

where

α_j (j = 0 to 3) are the thresholds, or *Y* intercepts

$\beta_1, \beta_2, \ldots \beta_p$ are the logit coefficients/odds ratios

Because the odds ratio, or slope (β), of any predictor is assumed to be the same across all response categories, this is a proportional odds model (Liu, 2007; Long, 1997; Peng & Naegle Nichols, 2003). The proportional odds model works with the logit, or natural log, of the odds and estimates the log odds of being at or below a given category (Liu, 2007; Peng & Naegle Nichols, 2003). This model predicts four cumulative logits in the present study, representing the J-1 response categories (Liu, 2007; Long, 1997; Peng & Naegle Nichols, 2003).

Diagnostic Analyses

Outliers in both classical linear and logistic regression can affect results significantly (Garson, 2012a). Cases that are poorly predicted by the solution – in logistic regression, for example, a case in one category of the outcome may have a high probability for being in another category – are considered outliers. If there are many outliers, the model has poor fit (Tabachnick & Fidell, 2001). The identification and removal of outliers occurs after the analysis is run. If the CLRM model is poor-fitting, the researcher should examine standardized residuals to find outlying cases (Garson, 2012b; Tabachnick & Fidell, 2001). Outliers should be considered for removal or separate modeling if the standardized residuals are greater than ±3.3 (this value indicates an outlier at the α =.001 level; Garson, 2012b). Standardized Pearson residuals in binary logistic regression can be calculated using leverage statistics (Peng & So, 2002), but SPSS does not provide these measures for multinomial and ordered logistic regression. Therefore, the difference between observed and predicted probabilities was calculated and examined for large values that are separated from the rest of the values.

Several diagnostic statistics exist on which decisions about potential outliers and a model's poor fit can be made. However, these are best computed by covariate patterns (Peng & So, 2002), and SPSS calculates most diagnostic statistics by individual observations. Nevertheless, an examination of the available diagnostic statistics is warranted. In addition to residuals, SPSS provides various measures of distance and influence for the CLRM. The only diagnostic statistic available in SPSS for multinomial and ordered logistic regression is the Pearson residual, which is a component of the Pearson χ^2 goodness-of-fit statistic. Fortunately, SPSS does report this value by covariate pattern, but the more informative change in Pearson χ^2 diagnostic statistic is not reported (Peng & So, 2002). In the absence of this preferred diagnostic statistic, the Pearson residuals were examined for large values which indicate a poorly explained covariate pattern. An analogous measure in CLRM is Cook's distance, *D*, an influence statistic which measures the change to the solution experienced by deleting a given observation.

A plot of Cook's *D* against observations was examined in order to identify any cases with unusually high influence.

Chapter 4: Results

In order to determine the impact of alternate statistical approaches for analyzing the categorical dependent variable of selectivity of postsecondary institution attended, models of student college choice were estimated in SPSS version 19.0.0.1 with classical linear regression, multinomial logistic regression, and ordered logistic regression. Results of the examination of the assumptions of the models are presented first, followed by the results of each regression. Within each regression method, results of three models are presented: a model including only student ascriptive characteristics; a model with the student ascriptive characteristics and the academic characteristics of ACT® score and grade point average; and a full model with all 10 ascriptive and academic explanatory variables. The chapter concludes with a comparison across regression method of the effects of adding variables to the model, resulting in the selection of the most appropriate specification of the influences on selectivity of postsecondary institution attended.

Assumptions

The classical linear, the multinomial logistic, and the ordered logistic regression models all make assumptions about the data generating process, although more restrictive assumptions apply to the CLRM than to the logistic models. Before regression analysis was run in the current study, the assumptions of the models were examined. Results are discussed in the following sections.

Distribution of Errors

Zero mean error. The ordinary least squares (OLS) estimator used in classical linear regression will always minimize the sum of squared residuals, by definition, but in certain estimating situations it might not possess other properties researchers deem important such as unbiasedness or efficiency. One assumption of the classical linear regression model (CLRM) necessary for the desirable properties of the OLS estimator to hold is that the population errors have a mean of zero (Kreiberg, n.d.). The population regression line is not known for sample data; therefore, assessment of this assumption is done by determining whether selectivity bias is an issue. Selectivity bias can cause certain populations to be over- or under-represented in the sample (Garson, 2011b).

Initially, the data for the present study largely avoided the selectivity bias issue as the data were composed of the entire population of high school students in two states. Selectivity bias may result when a subpopulation of students who choose to take a college entrance exam because of a predisposition to attend college is studied, but Illinois and Colorado's requirement that all high school students take the ACT® test creates an unbiased dataset. However, half of the students were eliminated from the initial dataset due to the presence of missing data. The remaining cases used in this analysis differ from the whole population of Illinois and Colorado test takers, but the differences are known and have been examined.

Non-stochasticity. In order for the desired properties of the OLS estimator to hold, there must be no relationship between the population residuals and each of the predictor variables (Garson, 2012b; Kreiberg, n.d.). This assumption cannot be tested statistically, as the population regression line is not known for sample data. In order for a predictor variable and the population error to be independent of each other, variables not included in the equation should not be causes of the outcome variable and should not be correlated with included variables. Additionally, the outcome variable should not be an influence on one or more explanatory variable (Garson, 2012b).

In the present data, the outcome variable of selectivity of college attended could not be considered a cause of any of the predictor variables. Ascriptive student characteristics and all academic characteristics included in the full CLR model which uses OLS estimation (Equation C3) are determined and measured prior to the outcome variable. In regards to omitted variables, this relates to the assumption of proper specification of the model (relevant to all regression methods) discussed in Chapter 3. Based on the literature review, all available predictor variables thought to be influences on the outcome variable of selectivity of college attended are included in the full model, Equation C3. However, as noted, ACT®, Inc. does not provide information on parent education level, which has been shown to be related to college attendance.

Together with the assumption of zero mean error, the assumption of non-stochasticity relates to the concept of exogeneity. As such, the selectivity bias discussion in the above section is relevant here as well. My data have addressed the concern that an omitted variable might be responsible for selection into the sample of ACT® test takers and related to the dependent variable of college selectivity by sampling from states in which the ACT® test is required. However, it is not known whether an unobserved variable may influence both selection into my sample on the basis of complete data and the dependent variable of selectivity of postsecondary institution attended. This unobserved variable problem is relevant for both classical linear and logistic regressions.

Normality of errors. Errors, or residuals, should be normally distributed in the CLRM. Although the CLRM does not require measurement of the dependent variable at a particular level, the normality of errors assumption required for the desired properties of the OLS estimator to hold is difficult to meet with ordinal data due to its discrete nature.

At extreme values of the explanatory variables, errors will likely depart from a normal distribution (McKelvey & Zavoina, 1975). Although my dependent variable, admissions policy, is discrete and ordinal-level, the assumption is worth examining in order to discern the extent to which it is violated.

The assumption of normality of errors is particularly important in small samples, but non-normality of errors – often caused by outliers (Kreiberg, n.d.) – is essentially inconsequential in large samples (Garson, 2012b; Kreiberg, n.d.). Although my data sample is large and no substantial univariate or multivariate outliers were detected during data screening, I tested this assumption necessary for the properties of the OLS estimator to hold anyway by examining residuals plots (Tabachnick & Fidell, 2001), normal probability plots (a P-P Plot), histograms of standardized residuals, and by skewness and kurtosis values of the standardized residuals (Garson, 2011b) for all equations. When sample size is large, Tabachnick and Fidell (2001) recommend assessing normality of variables by graphical methods rather than formal inference tests of skewness and kurtosis. Even minor deviations from normality may cause rejection of the null hypothesis of a normal distribution in large samples (Garson, 2011b; Tabachnick & Fidell, 2001).

Examination of the unstandardized residuals plot showed slight positive skew for Equation C1, the model with only student ascriptive characteristic predictors. The residuals plots for Equations C2 and C3 are very similar to each other and show an approximately normal distribution. Histograms of standardized residuals supported these observations: Equation C1 is moderately positively skewed whereas Equations C2 and C3 look approximately normal. The distributions of errors for Equations C2 and C3 are normal by visual examination of the P-P Plots, which resulted in 45-degree lines, indicating a good match between observed and expected cumulative probabilities of occurrence of the standardized residuals. However, the line in the P-P Plot of Equation C1 veers from a 45-degree angle. Finally, standardized residuals skewness and kurtosis values decrease as variables are added to the model. In Equation C1, the skewness value is .513 and the kurtosis value is -.504, but the values for Equations C2 and C3 (skewness = -.114 and -0.174; kurtosis = -.162 and -.040, respectively) are close to zero, the value for a normal distribution. Therefore, it appears that errors in Equations C2 and C3 are normally distributed but that there may be a violation of this assumption for Equation C1.

Homoskedasticity. As heteroskedasticity is typically associated with cross-sectional rather than time-series data (Kreiberg, n.d.), this is an important assumption of the CLRM that needs to be tested in the current study. Residuals should be dispersed randomly across all values of the estimated dependent variable (Garson, 2012b; Tabachnick & Fidell, 2001). Although homoskedasticity is another assumption about the distribution of the error term that is difficult to meet with a discrete dependent variable (McKelvey & Zavoina, 1975), I examined the properties of the residuals for informative purposes. Heteroskedastic data could indicate an interaction between a measured explanatory variable and one not included in the model or the presence of skewed explanatory variables and will result in biased standard errors (Garson, 2012b; Kreiberg, n.d.).

Although regression is relatively robust to modest violations of homoskedasticity (Garson, 2012b), this assumption was checked with each of the CLR models (Equations C1-C3) by first examining residuals scatterplots of the unstandardized predicted dependent variable by the unstandardized residuals (Garson, 2011b, 2012b; Tabachnick & Fidell, 2001). Under homoskedasticity, the width of the residuals band is approximately equal for all values of the predicted dependent variable. Statistical tests of homoskedasticity (e.g. White's test, Breusch-Pagan-Godfrey test) are not readily available in SPSS (Garson, 2012b).

The residuals plot should appear largely as a "cloud" around the center in order to indicate homoskedasticity. Equation C1 appears to be homoscedastic across the predicted dependent variable range (.157 to 2.07; actual observed dependent variable range is 0 to 4). However, the residuals plot for Equation C2 shows a greater range of error (unstandardized residuals) for the higher values of the predicted dependent variable, and this discrepancy is even more noticeable in the residuals plot for Equation C3. Examination of scatterplots of various combinations of standardized and unstandardized observed and predicted values, residuals, and explanatory variables confirmed this assessment.

To roughly assess the assumptions of normality and homoskedasticity [$\varepsilon_i \sim N(0, \sigma^2)$; $Var(\varepsilon_i) = \sigma^2 < \infty$] for the predicted dependent variable, I rounded the continuous predicted values to the nearest whole value and devised categories based on the observed, discrete admissions policy variable. The means and the skewness and kurtosis values of the residuals should be close to zero for each category ("0" through "4") for each model (Equations C1, C2, and C3) and standard deviations should be very similar for each dependent variable category within model.

As expected (McKelvey & Zavoina, 1975), the residuals departed from a normal distribution at the extreme values of the predicted dependent variable, although this departure is more marked in Equations C2 and C3. Equation C1 only predicts categories "0," "1," and "2," and the unstandardized residuals are distributed approximately normally with a mean close to zero for the dependent variable categories of "1" and "2." Equations C2 and C3 only predicted one case for dependent variable category "4," so there is no error distribution to assess. However, the extreme categories of "0" and "3" (the former more so than the latter) departed from normal in both models.

Within each model, standard deviations of the unstandardized residuals increased as predicted value of the dependent variable increased, although to a lesser extent in Equation C1 (the range of standard deviations is the same for Equations C2 and C3), confirming the observations from the residuals plots that the assumption of homoskedasticity is not met in Equations C2 and C3 and casting doubt on the validity of this assumption for Equation C1.

Non-autocorrelation. Typically a problem of time-series data, non-autocorrelation refers to the concept that error terms should not be correlated (either conceptually or linearly), also referred to as independence of observations (Garson, 2011b, 2012b; Kreiberg, n.d.). This assumption is required of all three regression methods in the current study and can be verified conceptually as well as empirically. Although it is unlikely that this assumption is violated due to the cross-sectional nature of my data, the Durbin-Watson coefficient, d, was examined in the CLR full model, Equation C3. A d value between 1.5 and 2.5 is generally thought to indicate the absence of autocorrelation (Garson, 2012b); the value in the current study ($d = 1.92$) falls within this range and suggests, as expected, independence of observations. Although the Durbin-Watson statistic is not available for multinomial or ordered logistic regression, the fact that it was within the acceptable range in CLRM, along with the fact that the data are not time-series or otherwise paired, indicates that this assumption is not violated for the logistic models.

Linearity

The CLRM assumes a linear relationship between the dependent variable and a specific set of explanatory variables and an error term (Garson, 2012b; Greene, 2000; Kennedy, 1998; Kirk, 1999; Long, 1997). This assumption is best met with at least interval-level data on the dependent variable, but researchers often conceptualize an observed discrete dependent variable as a manifestation of a latent continuous variable and use ordinal-level data to approximately meet the assumption of linearity. The validity of this practice with the ordinal-level dependent variable in the current study was examined.

An indication of linearity is a plot of the residuals (standardized or unstandardized) against the standardized or unstandardized predicted values (Garson, 2011b; Tabachnick & Fidell, 2001). If linearity is present, the scatterplot should be random. Examination of the plot for my data showed a random pattern ($R^2 = 0$) for all three of the equations, suggesting linear relationships between various combinations of the explanatory variables and the ordered categorical dependent variable. Additionally, the standard deviation of the residuals does not exceed the standard deviation of the dependent variable ($SD_Y = 1.28$) in any of the models, although Equation C1 comes close ($SD_{res} = 1.2$). Granted, this standard deviation comparison "test" is just a rule of thumb (Garson, 2011b), but combined with the results of the scatterplot of standardized residuals and standardized estimates, it may indicate that the assumption of linearity is not violated, at least not enough to make the results of CLRM uninterpretable.

As noted, researchers often use ordinal-level data to approximately meet the assumption of linearity, as minor violations will not substantially affect the interpretation of the results (Garson, 2012b). There is evidence from the examination of the data that perhaps the ordered categorical dependent variable of Admissions Policy does not severely violate the assumption of linearity even though it is measured on an ordinal instead of interval level.

Logistic regression assumes a linear relationship between continuous predictors and the logit transformation of the dependent variable, and the Box-Tidwell approach is one of the simplest methods for testing this assumption (Garson, 2012a; Tabachnick & Fidell, 2001). Significant interaction terms for the crossproducts of variables and their natural logarithms indicate a violation of this assumption and the presence of nonlinearity in the logit (Garson, 2012a; Tabachnick & Fidell, 2001).

The natural logarithms of the five continuous explanatory variables income, siblings, ACT® score, GPA, total extracurricular participation were calculated to create new variables. The crossproducts of these variables with the original variables were then calculated to construct interaction variables. In the multinomial logistic regression for the full model, Equation M3, the interaction variables for income, ACT® score, GPA, and total extracurricular participation were significant (only the interaction of siblings with its natural log was not significant). Tabachnick and Fidell (2001) recommend remedying the nonlinearity through a transformation, such as the natural logarithm, of the predictors with significant interactions and replacement of the original variables with the transformed variables in another Box-Tidwell test. This approach was taken, and three of the five interactions were significant (only the interaction of total extracurricular participation with its natural log became non-significant after transformation). However, due to undefined values for the natural log of zero (students who indicated participating in zero extracurricular activities are excluded from analyses of the full model), the sample size was decreased from 5,200 to 4,588 when the log of total extracurricular participation was used. Additionally, Tabachnick and Fidell (2001) recommend comparing results from the regression analyses with and without the transformed variable in order to determine its necessity. Results regarding overall model evaluation, goodness-of-fit, and parameter estimates were not substantively different for Equation M3 with and without the extracurricular participation variable transformation. Because of the non-random reduction in sample size caused by using the natural log transformation, along with the fact that the results do not drastically differ when the transformed variable is used, I decided to continue with the original extracurricular variable for the final analysis. Therefore, the problem of non-linear relationships between income, ACT® score, GPA, and total extracurricular participation and the logit of the dependent variable still remains.

When the assumption of linearity in the logit of the dependent variable was tested with the Box-Tidwell transformation approach in the ordered logistic regression for the full model, Equation O3, the interaction variables for income, ACT® score, and total extracurricular participation were significant (the interactions of siblings and GPA with their natural logs were not significant). Including the natural logarithms of income, ACT® score, and total extracurricular participation in place of the original variables in Equation O3 resulted in the interactions of income and ACT® score still being significant. However, like in the multinomial logistic regression analysis, replacement of the total extracurricular participation variable with its natural logarithm did not alter the results substantially. Therefore, the untransformed extracurricular participation variable was used in the final ordered logistic regression analysis in order to maintain consistent sample sizes. The assumption of linearity in the logit is violated for income, ACT® score, and total extracurricular participation in the ordered regression full model.

Absence of Multicollinearity

High multicollinearity among the explanatory variables inflates standard errors in all three regression methods (Garson, 2012b; Tabachnick & Fidell, 2001). In such a case, the effects of the predictors cannot be separated and are unreliable. In CLRM, a sign of high multicollinearity is a large R^2 value and significant *F* tests of the model with extremely high standard errors and few significant *t*-tests of coefficients (Garson, 2011b). This is not the case for any of my models, Equations C1-C3. The R^2 values are moderate, ranging from .117 to .391, and the *F* tests of the models are significant. However, rather than a few significant *t*-tests of coefficients, 20 of the 28 (71%) coefficient tests across all three models are significant, and all standard errors are less than .075.

Additionally, a tolerance value of less than .20 for any predictor variable is generally accepted to indicate that the violating variable be dropped from the analysis (Garson, 2011b). None of the tolerance values for any of the predictor variables in my three CLRM equations is less than .20, indicating an absence of any high multicollinearity. However, a value of the collinearity diagnostic condition index of greater than 15 indicates possible multicollinearity problems (Garson, 2012b), and one of my variables in Equation C2 (GPA, 15.47) and two variables in Equation C3 (Math Course-Taking Pattern 5, 20.15; and Total Extracurricular Participation, 22.15) have condition indices over 15. These condition indices do not approach 30, though, which is the value generally viewed as signaling serious collinearity problems (Garson, 2012b),

If two or more variables have a variance proportion of greater than .50 on a factor with a high condition index, multicollinearity is generally thought to be a problem (Garson, 2012b). Examination of the variance proportion values for GPA, Math Course-Taking Pattern 5, and Total Extracurricular Participation, available from the collinearity diagnostics, showed that one variable has a variance proportion of greater than .50 with GPA (Math Course-Taking Pattern 5 in Equation C3) and that Math Course-Taking Pattern 2 has a variance proportion of greater than .50 with Total Extracurricular Participation. No variables have variance proportions greater than .50 in the Math Course-Taking Pattern 5 factor. Since the factors with high condition indices do not have more than one variable with sizable variance proportions, coupled with the fact that the condition indices were not much over 15, multicollinearity was not determined to be a problem in my data when assessed with the CLRM.

Using the presence of large standard errors as an indicator of multicollinearity, results from the multinomial and ordered logistic regressions were consistent with the findings of the CLRM. In the multinomial equations (Equations M1-M3), all the coefficient standard errors are less than 1.0, and in the ordered logistic regression equations, all standard errors are less than 0.23. Additionally, if two similar variables have widely different effects, this could signal the presence of high multicollinearity (R. Toutkoushian, personal communication, Fall 2005). The most similar of my variables, ACT® score and GPA, were significant in every equation in which they appeared in both types of regression, further indicating that the assumption of the absence of multicollinearity is met in my data.

Large Samples

Estimation through maximum likelihood, used in the multinomial and ordered logistic regressions in the present study, requires a large sample size (Garson, 2012a). Long (1997) suggests that at least 500 subjects is adequate. A minimum of 10 observations per explanatory variable is a commonly-recommend rule of thumb (Garson, 2012a; Long, 1997), as long as that number is at least 100 (Long, 1997). An alternate recommendation is a sample size of at least 30 times the number of parameters being estimated (Garson, 2012a), especially if the explanatory variables are highly collinear or there is little variation in the dependent variable. There are 10 explanatory variables in my study; 30 times this number is 300. As previously discussed in Chapter 3, the final sample size of 5,200 should be large enough to accommodate ML estimation.

To further verify that this assumption of large sample size has been met, parameter estimates were examined after the analyses were run. Perfect separation of outcome groups by a discrete predictor renders a ML solution impossible and can be caused by too small a sample. Indication of these problems is evident by extremely high parameter estimates and standard errors or failure to converge (Garson, 2012a; Tabachnick & Fidell, 2001). All equations for the multinomial and ordered regressions converged, and as discussed in the section on multicollinearity above, no extremely high standard errors were found in any of my regressions. Furthermore, parameter estimates are not extremely high: all explanatory variable and threshold estimates are less than 10, and the intercepts in multinomial logistic regression are less than 20. Evidence indicates that the large sample size assumption has been met in the current study (*N* = 5,200).

Adequate Expected Cell Frequencies

Another assumption required by both the multinomial and ordered logistic regression models is that expected cell frequencies are adequate to support the goodness-of-fit tests that compare full or partial models to a hypothetical, perfect model (Tabachnick & Fidell, 2001). Crosstabulations were run for all pairs of discrete variables in order to assure that the requirement of no more than 20% of the cells have a count less than five and that all expected frequencies are greater than one (Garson, 2012a; Tabachnick & Fidell, 2001).

Cell counts and expected cell counts were examined for all pairs of discrete variables (Race, College Preparatory Track, Advanced Courses, Foreign Language, Math Course-Taking Pattern, and Admissions Policy). No expected frequencies less than or equal to one were found, and only one cell (0.57% of the 176 cells examined) had a count of less than five [Math Course-Taking Pattern 3 (minimum core) by Admissions Policy value “4,” or “highly selective”]. Based on this analysis, the dataset meets the adequacy of expected cell frequencies assumption.

Parallel Regression

The ordered logistic regression model used in this study makes the assumption of proportional odds, or parallel regression, meaning there is a common effect of any explanatory variable regardless of response value (Agresti, 1996; Garson, 2011a; Liu, 2007; Long, 1997; Peng & Naegle Nichols, 2003). Before interpreting the results of Equations O1, O2, and O3, this assumption of proportional odds was examined for each equation through the χ^2 likelihood ratio test of parallel lines in SPSS version 19.0.0.1. Results for all three equations were significant at the $\alpha = .001$ level, indicating that the assumption is violated. However, large samples may cause the test of parallel lines to be significant even if differences are trivial and potentially ignorable for substantive interpretation (Garson, 2011a). Following a recommendation of Garson (2011a), a smaller sample size for which the parallel lines tests discriminates ($n = 200$) was taken and the test run again for all three equations.

For the likelihood ratio test of the model constrained to have equal slopes for the predictor variables of income, number of siblings, and race/ethnicity (Equation O1) against an unconstrained model, $\chi^2(15) = 12.71$, $p = .625$, indicating that the parallel regression assumption was upheld in this smaller sample size. The result of the test for parallel lines in Equation O2 was very close to being non-significant at the $\alpha = .05$ level [$\chi^2(21) = 32.94$, $p = .047$], and the test for Equation O3 (which includes all 10 explanatory variables) was non-significant [$\chi^2(48) = 14.88$, $p = 1.0$]. Therefore, it was concluded that violations of the parallel regression assumption observed in the full dataset ($N = 5{,}200$) are likely due to the over-sensitivity of the parallel lines test in large samples. The slopes are assumed to be sufficiently parallel for interpreting the other results of the ordered logistic regressions.

Classical Linear Regression

Equation C1

The classical linear regression model in which admissions policy is regressed on the student ascriptive characteristics of income, siblings, and race/ethnicity is statistically significant [$F(5, 5194) = 137.27$; $p < .001$]. The R^2 value is .117. Results of this analysis indicate that, as a model, income, number of siblings, and race/ethnicity predict selectivity of postsecondary institution attended. The model accounts for almost 12% of the variance in admissions policy. Results of Equation C1 are presented in Table 16.

Income, number of siblings, and being of Hispanic ethnicity make statistically significant unique contributions to the model. A $1,000 increase in annual family income is associated with a small increase in selectivity of college attended (a change of .014 in the dependent variable value for every unit increase in annual income), but as the number of siblings increases, selectivity of college decreases. Additionally, Hispanic students systematically attend postsecondary institutions of lower selectivity than do White students.

Table 16

Classical Linear Regression Results: Equation C1

Variable	*B*	SE *B*	β (*standardized beta*)	*t*	Sig. (*p*)
Income[a]	.014	.001	.315	22.67	<.0001
Siblings	-.045	.012	-.048	-3.63	<.001
Black	-.065	.060	-.015	-1.10	.272
Hispanic	-.239	.054	-.060	-4.40	<.0001
Constant	.679	.046		14.85	<.0001

Note. $N = 5{,}200$. $R^2 = .117$. [a] Measured in units of $1,000.

Equation C2

The classical linear regression model in which admissions policy is regressed on student ascriptive characteristics as well as the student academic characteristics of ACT® score and grade point average (GPA) is also statistically significant [$F(7, 5192) = 428.67$;

$p < .001$]. The R^2 value is .366. Results of this analysis indicate that, as a model, student ascriptive characteristics, ACT® score, and GPA predict selectivity of postsecondary institution attended. The model accounts for almost 37% of the variance in admissions policy. Results of Equation C2 are presented in Table 17.

Income and number of siblings still make statistically significant unique contributions to the model once ACT® score and GPA are added. However, the nature of the relationship between being of Hispanic ethnicity and admissions policy changes once the academic characteristics of ACT® score and GPA are added, and indicating a race/ethnicity of "Black" is now also statistically significantly related to the dependent variable. Black or Hispanic students systematically attend postsecondary institutions of higher selectivity than White students once ACT® score and GPA are taken into account. Finally, ACT® score and GPA are statistically significant predictors of selectivity of college attended. Students with higher ACT® score or GPA systematically attend more selective postsecondary institutions.

Table 17

Classical Linear Regression Results: Equation C2

Variable	*B*	SE *B*	*β* (*standardized beta*)	*t*	Sig. (*p*)
Income[a]	.006	.001	.134	10.70	<.0001
Siblings	-.030	.011	-.031	-2.81	.005
Black	.420	.052	.097	8.13	<.0001
Hispanic	.136	.047	.034	2.90	.004
ACT® Score	.097	.004	.385	25.31	<.0001
GPA	.442	.026	.238	16.90	<.0001
Constant	-2.35	.078		-30.27	<.0001

Note. N = 5,200. R^2 = .366. [a] Measured in units of $1,000.

Equation C3

The classical linear regression model in which admissions policy is regressed on all student ascriptive and student academic characteristics is statistically significant [$F(16, 5183) = 208.41$; $p < .001$]. The R^2 value is .391. Results of this analysis indicate that the full model predicts selectivity of postsecondary institution attended, accounting for 39.1% of the variance in admissions policy. Results of Equation C3 are presented in Table 18.

When all predictor variables are included, the socioeconomic ascriptive characteristics of income and siblings still make statistically significant unique contributions to the model. Being of Hispanic race/ethnicity is no longer significantly related to admissions policy, but being of Black race/ethnicity still significantly indicates attendance at institutions of higher selectivity level relative to White students. The student academic characteristics of ACT® score, GPA, college preparatory track (enrolled), foreign language (three or more years), math course-taking pattern “2” (four or more years), math course-taking pattern “5” (Algebra I, Algebra II, Geometry, Trigonometry, and Calculus or another advanced math course) and total extracurricular participation are all statistically significantly positively associated with an increase in selectivity of college attended.

Table 18

Classical Linear Regression Results: Equation C3

Variable	B	SE B	β[a]	t	Sig. (p)
Income[b]	.005	.001	.110	8.83	<.0001

Siblings	-.033	.010	-.035	-3.24	.001
Black	.294	.052	.068	5.67	<.0001
Hispanic	.068	.047	.017	1.46	.145
ACT® Score	.077	.004	.302	18.46	<.0001
GPA	.330	.027	.178	12.09	<.0001
College Preparatory Track	.191	.031	.075	6.06	<.0001
Advanced Courses	.048	.034	.019	1.42	.155
Foreign Language	.208	.032	.080	6.54	<.0001
Math Course-Taking Pattern 1	-.002	.074	.000	-.028	.997
Math Course-Taking Pattern 2	.157	.053	.050	2.95	.003
Math Course-Taking Pattern 3	-.069	.047	-.021	-1.46	.144
Math Course-Taking Pattern 4	.021	.048	.007	.442	.658
Math Course-Taking Pattern 5	.180	.054	.053	3.32	.001
Total Extracurricular Participation	.038	.007	.068	5.77	<.0001
Constant	-1.89	.086		-21.81	<.0001

Note. $N = 5{,}200$. $R^2 = .391$. [a] β is the standardized beta coefficient. [b] Measured in units of $1,000.

Diagnostic Analyses

Among the three CLRM equations, Equation C1 explains the least amount of variance in the dependent variable ($R^2 = .117$). Examination of the standardized residuals, however, indicates that outliers are likely not the cause of this poor fit. The range of standardized values is -1.72 to 2.97, and graphical measures (histogram and boxplot) indicate that all values are closely connected. Additionally, the plot of Cook's distance, *D*, reveals only one observation that is separated from the group in Equation C1.

Equations C2 and C3, which explain 36.6% and 39.1% of the variance respectively, also have no standardized residual values greater than ±3.3. Although the range of Equation C3's standardized residual values is the greatest (-3.2 to 3.2), a histogram and a boxplot of the residuals show that the points are closely connected. Plots of Cook's distance, *D*, reveal one potential outlier in Equation C2 (this is a different observation than the potential outlier in Equation C1). Equation C3's scatterplot of observation number against Cook's *D* shows no obvious outliers which exert undue influence on the regression.

Outliers do not appear to be a problem in the CLRM equations in the present study. All standardized residuals are within the acceptable range at the α =.001 level. Although the plot of Cook's *D* indicates that one observation has a larger influence on the solution in Equation C2 than do other observations, this equation explains the data moderately well (R^2 = .366); therefore, I would not consider removing this observation. Equation C1 is less well-fit to the data (R^2 = .117), and one potential Cook's *D* outlier was seen in the graphical measure. There are 15 observations with higher standardized residuals than the case with the extreme Cook's *D* value, though, so I determined that this case did not need to be deleted either.

Summary

All three models estimated with classical linear regression were found to be statistically significant, although they varied on amount of variance explained (R^2). Equation C1 includes only student background characteristics and explains 12% of the variance in selectivity of postsecondary institution attended. Adding the academic characteristics of ACT® score and GPA in Equation C2 resulted in the greatest increase in

R^2 from .117 to .366 (a significant R^2 change of .249, or almost 25% of variance explained). Equation C3 includes all explanatory variables but only represents a 2.5% increase in variance explained.

Among the student ascriptive characteristics, the income and sibling variables remain statistically significant across all three models, whereas the “Black” variable becomes significant only once the academic characteristics are included. Being of Hispanic race/ethnicity decreases in its influence on selectivity of institution attended, going from a significant negative influence in Equation C1 to a significant positive influence of lower magnitude (as measured by standardized beta coefficient, or the effect of a one standard deviation change in a predictor on the standard deviation change in the outcome variable) in Equation C2 to a non-significant positive influence in C3. ACT® score and GPA are significant in both equations in which they are included. Of the remaining academic characteristics included only in Equation C3, being enrolled in a college preparatory track; taking three or more years of foreign language; taking four or more years of math; taking the math course pattern of Algebra I, Algebra II, Geometry, Trigonometry, and Calculus or another advanced math course; and participating in an increasing number of extracurricular activities are all significantly positively related to selectivity level. Tables 19 and 20 present comparative results of Equations C1, C2, and C3.

Table 19

Summary of Classical Linear Regression Models

Criterion	Equation C1	Equation C2	Equation C3
Assumptions Violated[a]	Non-stochasticity Normality of errors	Homoskedasticity	Homoskedasticity

R^2	.117	.366	.391
F for change in R^2	137.27*** (df = 5, 5194)	1022.24*** (df = 2, 5192)	23.88*** (df = 9, 5183)
% Correctly Predicted	25.9	36.6	38.6
Predicted Dependent Variable Range[b]	1.91 (.157 to 2.07)	4.26 (-.760 to 3.50)	4.42 (-.615 to 3.80)
Diagnostic Analyses	No outliers	No outliers	No outliers

[a] The assumption of linearity between the predictors and the observed, discrete dependent variable is theoretically violated for all models.
[b] Observed dependent variable range is 4, from the values of 0 to 4.
***p < .001

Table 20

Behavior of Predictors across Classical Linear Regression Models

Variable	Equation C1 β[a]	Equation C2 β	Equation C3 β
Income[b]	.315***	.134***	.110***
Siblings	-.048***	-.031**	-.035**
Black	-.015	.097***	.068***
Hispanic	-.060***	.034**	.017
ACT® Score		.385***	.302***
GPA		.238***	.178***
College Preparatory Track			.075***
Advanced Courses			.019
Foreign Language			.080***
Math Course-Taking Pattern 1			.000
Math Course-Taking Pattern 2			.050**
Math Course-Taking Pattern 3			-.021

Math Course-Taking Pattern 4	.007
Math Course-Taking Pattern 5	.053**
Total Extracurricular Participation	.068***

Note. N = 5,200. The test of the hypotheses of beta = 0 are based on t ratios. [a] β is the standardized beta coefficient. [b] Measured in units of $1,000.

Multinomial Logistic Regression

Equation M1

A multinomial logistic regression model was fitted to examine the influence of the student ascriptive characteristics of income, siblings, and race/ethnicity on the likelihood of one of five categories of outcome: not attending college or attending a postsecondary institution with either an open, liberal/traditional, selective, or highly selective admissions policy. Results of the Likelihood Ratio test indicate that the multinomial logistic model with three predictors is more effective than an intercept-only, or null, model [$\chi^2(20) = 692.52$; $p < .001$].

The model is fit to the data well when a deviance criterion is used [$\chi^2(1068) =$ 1,118.53; $p = .138$]. However, the inferential Pearson goodness-of-fit test was significant [$\chi^2(1068) = 1,180.5$; $p < .01$], presenting contrasting evidence regarding the fit of the multinomial logistic regression model against actual outcomes. The Pearson and deviance statistics are based on different χ^2 distributions, but for large samples the results usually do not differ substantively (Garson, 2012a). Furthermore, Peng & So (2002) caution that because SPSS calculates these goodness-of-fit statistics from raw data rather than covariate patterns, the statistics should not be assessed against a χ^2 distribution at all.

Therefore, descriptive measures of association or effect size are particularly helpful in the case of Equation M1 where the results of the inferential tests are ambiguous.

The coefficient of determination, or R^2, is the preferred measure of overall effect size in the classical linear regression model because of its clear interpretation in terms of explained variation (Allen & Le, 2008; Menard, 2000). However, for models with categorical dependent variables, there is not a consensus on how to calculate corresponding measures of strength of association between a dependent variable and a set of predictors (Long, 1997; Menard, 2000; Peng & Naegle Nichols, 2003). The likelihood ratio index, often denoted R_L^2, was suggested by McFadden (as cited in Allen & Le, 2008; Long, 1997; and Menard, 2000) and can be applied to any model estimated with maximum likelihood (Long, 1997). It is generally accepted as the standard pseudo- R^2 measure for categorical dependent variables, particularly because it is invariant to base rate, naturally varies between zero and one, and is easy to interpret (Allen & Le, 2008; Menard, 2000; Peng & So, 2002). As an analogy to the commonly-used calculation of R^2 as the percentage of explained variation, R_L^2 can be interpreted as the proportional reduction in error when going from an intercept only model to a model with regressors included (Menard, 2000). McFadden's R_L^2 for Equation M1 is .045.

The five-way classification table reveals that the prediction for students who did not attend college was more accurate than for any other outcome category (87.1% specificity). The overall correct prediction was low at 37.6%. There appears to be over-prediction in the categories of "No College" (77.4% of all students) and "Liberal/Traditional" institution (21.0%) relative to the categories of "Selective" and "Highly Selective" institutions, and none of the 5,200 cases was predicted to attend a postsecondary institution with an “Open” admissions policy. In addition to the classification table, measures of association indicate the degree to which predicted probabilities correspond to actual outcomes. For the present model, Somers’ D_{yx} is .349, indicating a 34.9% reduction in errors when predicting outcome (admissions policy) category using estimated probabilities over chance alone.

Statistical tests of individual predictors using the Wald χ^2 statistic show that income has a statistically significant ($p < .001$) negative effect on the probability of either not attending college, attending an “Open” institution, a “Liberal/Traditional” institution, or a “Selective” institution relative to the base category of “Highly Selective” institution. Being Hispanic, as opposed to being White, increases the odds of not attending a postsecondary institution compared to attending a highly selective institution by a factor of 2.2. In other words, the odds of a student of Hispanic ethnicity not attending college are 2.2 (= $e^{0.7818}$) times greater than the odds for a White student, holding family income and number of siblings constant. There were no other statistically significant effects of race on admissions policy. Detailed results of Equation M1 are presented in Table 21.

Table 21

Multinomial Logistic Regression Results: Equation M1

Predictor Variable	Dependent Variable Category[a] (Admissions Policy)							
	No College		Open		Liberal		Selective	
	B	e^B (*odds ratio*)	*B*	e^B (*odds ratio*)	*B*	e^B (*odds ratio*)	*B*	e^B (*odds ratio*)
Income[b]	-.035*** (.002)	.97	-.031*** (.002)	.97	-.019*** (.002)	.98	-.008*** (.002)	.99
Siblings	.072 (.045)	1.1	-.019 (.047)	.98	-.018 (.048)	.98	-.083 (.054)	.92

Black	.114 (.236)	1.1	-.293 (.246)	.75	-.117 (.248)	.89	-.183 (.279)	.83
Hispanic	.781** (.247)	2.2	.187 (.257)	1.2	.304 (.259)	1.4	.202 (.285)	1.2
Constant	3.47*** (.187)	NA	3.15*** (.192)	NA	2.36*** (.195)	NA	1.15*** (.214)	NA

	χ^2	*df*	*p*
Likelihood Ratio Test	692.52	20	<.0001
Goodness-of-Fit Statistics			
Pearson	1,180.50	1,068	.009
Deviance	1,118.53	1,068	.138

Note. N = 5,200. Standard errors in parentheses. The test of the hypotheses of $B = 0$ are based on Wald's X^2, $df = 1$. McFadden's $R_L^2 = .045$. Somers' $D_{yx} = .349$.
[a] Highly selective admissions policy is the reference category. [b] Measured in units of $1,000.
**$p < .01$, ***$p < .001$

Equation M2

A second multinomial logistic regression model was fitted to examine the influence of the student ascriptive characteristics as well as the student academic characteristics of ACT® score and grade point average (GPA) on the likelihood of one of five categories of outcome: not attending college or attending a postsecondary institution with either an open, liberal/traditional, selective, or highly selective admissions policy. Results of the Likelihood Ratio test indicate that the multinomial logistic model with five predictors is more effective than an intercept-only, or null, model [$\chi^2(28) = 2{,}597.85$; $p < .001$]. According to both the deviance criterion [$\chi^2(18080) = 11{,}476.96$; $p = 1.0$] and the Pearson χ^2 test [$\chi^2(18080) = 16{,}834.38$; $p = 1.0$], the model is fit to the data well. The McFadden's R_L^2 descriptive measure of strength of association for Equation M2 is .169.

The classification table reveals that the prediction for students who did not attend college was more accurate than for any other outcome category (80.1% specificity). The next highest specificity is 48.0% correct predictions for students who attended highly selective institutions. The overall correct prediction was low at 44.0%, although this is an improvement over the multinomial model with only ascriptive predictors. As with Equation M1, there appears to be over-prediction in the categories of "No College" and "Liberal/Traditional" institution (59.1% and 23.9% of all students, respectively) relative to the categories of "Selective" and "Highly Selective" institutions, although to a lesser degree. In contrast to Equation M1 in which none of the 5,200 cases was predicted to attend a postsecondary institution with an "Open" admissions policy, Equation M2 predicted 3.2% of cases to fall in this category. In addition to the classification table, measures of association indicate the degree to which predicted probabilities correspond to actual outcomes. For the present model, Somers' D_{yx} is .513, indicating a 51.3% reduction in errors when predicting outcome (admissions policy) category using estimated probabilities over chance alone.

Statistical tests of individual predictors using the Wald χ^2 statistic show that income has a statistically significant ($p < .001$) negative effect on the probability of either not attending college, attending an "Open" institution, or attending a "Liberal/Traditional" institution – but not a "Selective" institution – relative to the base category of "Highly Selective" institution.

Being Black or Hispanic, as opposed to being White, statistically significantly decreases the odds of not attending a postsecondary institution, attending an "Open" institution, or attending a "Liberal/Traditional" institution compared to attending a highly selective institution. The odds of attending a "Selective" institution are decreased for Black (but not Hispanic) students by a factor of 0.243. In other words, the odds of a student of African American ethnicity attending a postsecondary institution with a selective admissions policy, as opposed to a highly selective institution, are 0.243 (=$e^{-1.41}$) times less than the odds for a White student.

ACT® score and grade point average (GPA) are statistically significantly ($p <$ .001) negatively related to each category of outcome relative to the reference category of "Highly Selective" postsecondary institution. As a student's ACT® score or GPA increases (GPA more so than ACT® score), he or she is less likely to either not attend college or to attend an "Open," "Liberal/Traditional," or "Selective" institution than he or she is to attend a "Highly Selective" postsecondary institution. Detailed results of Equation M2 are presented in Table 22.

Table 22

Multinomial Logistic Regression Results: Equation M2

Predictor Variable	Dependent Variable Category[a] (Admissions Policy)							
	No College		Open		Liberal		Selective	
	B	e^B (*odds ratio*)	*B*	e^B (*odds ratio*)	*B*	e^B (*odds ratio*)	*B*	e^B (*odds ratio*)
Income[b]	-.022*** (.003)	.98	-.019*** (.003)	.98	-.012*** (.003)	.99	-.004 (.003)	1.0
Siblings	.036 (.056)	1.0	-.052 (.057)	.95	-.043 (.055)	.96	-.100 (.058)	.91
Black	-2.62*** (.302)	.07	-2.95*** (.310)	.05	-2.09*** (.301)	.12	-1.41*** (.313)	.24
Hispanic	-.816** (.294)	.44	-1.38*** (.301)	.25	-.839** (.293)	.43	-.501 (.304)	.61

ACT® Score	-.410*** (.021)	.66	-.422*** (.022)	.66	-.321*** (.021)	.73	-.188*** (.021)	.83
GPA	-2.33*** (.189)	.10	-1.95*** (.190)	.14	-1.16*** (.188)	.31	-.904*** (.194)	.41
Constant	20.43*** (.783)	NA	19.34*** (.786)	NA	14.22*** (.769)	NA	9.18*** (.771)	NA

	χ^2	*df*	*p*
Likelihood Ratio Test	2,597.85	28	<.0001
Goodness-of-Fit Statistics			
Pearson	16,834.38	18,080	1.0
Deviance	11,476.96	18,080	1.0

Note. N = 5,200. Standard errors in parentheses. The test of the hypotheses of $B = 0$ are based on Wald's X^2, $df = 1$. McFadden's $R_L^2 = .169$. Somers' $D_{yx} = .513$.
[a] Highly selective admissions policy is the reference category. [b] Measured in units of $1,000.
**$p < .01$, ***$p < .001$

Equation M3

The multinomial logistic regression model in which admissions policy is regressed on all student ascriptive and student academic characteristics is more effective than an intercept-only, or null, model [$\chi^2(64) = 2{,}938.50$; $p < .001$]. According to both the deviance criterion [$\chi^2(20688) = 12{,}446.76$; $p = 1.0$] and the Pearson χ^2 test [$\chi^2(20688) = 19{,}877.91$; $p = 1.0$], the model is fit to the data well. McFadden's R_L^2 for Equation M3 is .191.

The classification table reveals that the prediction for students who did not attend college was more accurate than for any other outcome category (77.0% specificity). As in Equation M2, the next highest specificity is 52.5% correct predictions for students who attended highly selective institutions. The overall correct prediction was low at 46.3%, representing a 2.3% increase over the multinomial model with ascriptive predictors and the academic predictors of ACT® score and grade point average. As with Equations M1 and M2, there appears to be over-prediction in the categories of "No College" and "Liberal/Traditional" institution (54.3% and 23.5% of all students, respectively) relative to the categories of "Selective" and "Highly Selective" institutions, although to a lesser degree. For the model M3, Somers' D_{yx} is .503, indicating a 50.3% reduction in errors when predicting outcome (admissions policy) category using estimated probabilities over chance alone. This is a decrease in error reduction from Equation M2.

Statistical tests of individual predictors using the Wald χ^2 statistic show that income has a statistically significant ($p < .001$) negative effect on the probability of either not attending college, attending an "Open" institution, or attending a "Liberal/Traditional" institution – but not a "Selective" institution – relative to the base category of "Highly Selective" institution.

When all explanatory variables are in the regression equation, being Black as opposed to being White statistically significantly decreases the odds of not attending a postsecondary institution, attending an "Open" institution, attending a "Liberal/Traditional" institution, or attending a "Selective" institution compared to attending a highly selective institution. With all variables in the equation, being of Hispanic ethnicity is significantly negatively related to non-attendance ($p = .041$), attendance at an "Open" institution ($p < .001$), and attendance at a "Liberal/Traditional" institution ($p = .016$). The inclusion of the additional student academic characteristics from Equation M2 to Equation M3 does not affect the impact of ACT® score and grade point average (GPA): both are still statistically significantly ($p < .001$) negatively related

(GPA more so than ACT® score) to each category of outcome relative to the reference category of "Highly Selective" postsecondary institution.

Of the additional student academic characteristics present in Equation M3, College Preparatory Track, Advanced Courses, Foreign Language, Math Course-Taking Pattern 1 (any combination of three to three and a half years of math), 3 [Algebra I, Algebra II, and Geometry (minimum core)], 4 (Algebra I, Algebra II, Geometry, and Trigonometry or another advanced math course), and Total Extracurricular Participation had statistically significant effects on one or more category of outcome. Enrolling in a college preparatory track decreases the odds of not attending college ($p < .01$), as opposed to attending an institution with a highly selective admissions policy, by a factor of 0.574, but this does not have a significant effect on attendance at any other level of selectivity. Having taken one or more advanced placement, accelerated, or honors course decreases the odds of attending an "Open" institution ($p = .049$) by a factor of 0.60. Taking three or more years of a foreign language is negatively related both to non-attendance and to attendance at an "Open," "Liberal/Traditional" ($p < .001$), or "Selective" ($p < .01$) institution.

Students who took the minimum core of math courses are more likely to either not attend college or to attend open, liberal/traditional, or selective institutions than to attend highly selective institutions, by factors ranging from 6.5 ("No College") to 12.2 ("Liberal/Traditional"). Taking the combination of Algebra I, Algebra II, Geometry, and Trigonometry (or another advanced math course) or simply taking any combination of three to three and a half years of math is also positively related to attendance at selective or liberal/traditional institutions. As the number of extracurricular activities increases, the odds of not attending a postsecondary institution, attending an "Open" institution, attending a "Liberal/Traditional" institution, or attending a "Selective" institution rather than a "Highly Selective" institution decrease. Detailed results of Equation M3 are presented in Table 23.

Table 23

Multinomial Logistic Regression Results: Equation M3

Predictor Variable	Dependent Variable Category[a] (Admissions Policy)							
	No College		Open		Liberal		Selective	
	B	e^B (*odds ratio*)	*B*	e^B (*odds ratio*)	*B*	e^B (*odds ratio*)	*B*	e^B (*odds ratio*)
Income[b]	-.018*** (.003)	.98	-.016*** (.003)	.99	-.010*** (.003)	.99	-.003 (.003)	1.0
Siblings	.051 (.058)	1.1	-.035 (.059)	.97	-.036 (.057)	.97	-.094 (.060)	.91
Black	-2.27*** (.313)	.10	-2.62*** (.320)	.07	-1.98*** (.311)	.14	-1.38*** (.322)	.25
Hispanic	-.612* (.300)	.54	-1.15*** (.308)	.32	-.725* (.299)	.49	-.445 (.308)	.64
ACT® Score	-.332*** (.023)	.72	-.345*** (.023)	.71	-.284*** (.022)	.75	-.167*** (.022)	.85
GPA	-1.91*** (.198)	.15	-1.51*** (.200)	.22	-.928*** (.197)	.40	-.797*** (.202)	.45
Col. Prep. Track	-.556*** (.173)	.57	-.185 (.176)	.83	.066 (.172)	1.1	-.034 (.175)	.97
Adv. Courses	-.315 (.255)	.73	-.504* (.256)	.60	-.195 (.254)	.82	-.135 (.264)	.87
For. Lang.	-.942*** (.175)	.39	-1.05*** (.178)	.35	-.806*** (.172)	.45	-.512** (.176)	.60
Mat. Crse.	.560	1.8	.372	1.5	1.24*	3.5	1.57*	4.8

Pat. 1	(.583)		(.591)		(.588)		(.618)	
Mat. Crse. Pat. 2	-.150 (.409)	.86	-.438 (.414)	.65	.253 (.417)	1.3	.732 (.457)	2.1
Mat. Crse. Pat. 3	1.87** (.704)	6.5	2.05** (.705)	7.8	2.50*** (.710)	12.2	2.34** (.731)	10.4
Mat. Crse. Pat. 4	.085 (.424)	1.1	.350 (.426)	1.4	1.06* (.431)	2.9	1.09* (.472)	3.0
Mat. Crse. Pat. 5	-.461 (.412)	.63	-.333 (.415)	.72	.572 (.418)	1.8	.730 (.458)	2.1
Tot. EC Partic.	-.179*** (.0330	.84	-.161*** (.033)	.85	-.122*** (.032)	.89	-.088** (.032)	.92
Constant	18.75*** (.860)	NA	17.52*** (.864)	NA	12.85*** (.850)	NA	8.20*** (.867)	NA

	χ^2	*df*	*p*
Likelihood Ratio Test	2,938.50	64	<.0001
Goodness-of-Fit Statistics			
Pearson	19,877.91	20,688	1.0
Deviance	12,446.76	20,688	1.0

Note. N = 5,200. Standard errors in parentheses. The test of the hypotheses of $B = 0$ are based on Wald's X^2, $df = 1$. McFadden's R_L^2 = .191. Somers' D_{yx} = .503.
[a] Highly selective admissions policy is the reference category. [b] Measured in units of $1,000.
*p < .05, **p < .01, ***p < .001

Diagnostic Analyses

Although the deviance goodness-of-fit test was non-significant for Equation M1, the Pearson χ^2 test indicated that the regression with only student ascriptive characteristics as explanatory variables may not be well-fit to the data. A scatterplot and boxplot of the residuals (calculated by subtracting the predicted category probability from the actual category probability) indicate two possible outliers, but their residual values are only separated from the previous value by .03.

Examination of the Pearson residuals for large values identifies, albeit subjectively, three covariate patterns (representing six cases) that are poorly explained by Equation M1. The covariate patterns of the two potential outliers, identified by the residuals criteria, do not match the patterns of the three most extreme Pearson residual values; however, the potential residuals outliers' covariate patterns do fall in the top 15 largest Pearson residuals. Due to the inconsistent findings in the analysis of residuals calculated from actual and predicted probabilities and Pearson residuals, I did not conclude that outliers were the cause of the poor fit of the model as indicated by the Pearson χ^2 test.

Both the deviance and Pearson criteria show good fit of the observed to expected frequencies for Equations M2 and M3; therefore, outlier analysis is not necessary (Tabachnick & Fidell, 2001).

Summary

All three models estimated with multinomial logistic regression represented improvement over an intercept-only model as evidenced by significant Likelihood Ratio χ^2 tests. Results of the inferential χ^2 goodness-of-fit tests largely indicate that the models are fit to the data well, the only exception being the significant ($p < .01$) Pearson χ^2 test in Equation M1. There is a large increase in values for the descriptive measure of strength of association, McFadden's R_L^2, from Equation M1 to Equation M2 when ACT® score and GPA are added, but the increase is much less from Equation M2 to M3 with the addition of other academic variables.

In terms of predicted probabilities, classification tables show an increase in percent correctly predicted from 37.6% in Equation M1 to 44.0% in Equation M2. The Somers' D_{yx} measure of association also increased, from $D_{yx} = .349$ in Equation M1 to $D_{yx} = .513$ in Equation M2. When additional academic characteristics were included in the model (Equation M3), the correct prediction rate increased to 46.3%, but Somers' D_{yx} actually decreased to .503.

The negative influence of the student ascriptive characteristic of income remains statistically significant across all three models for the response categories of "0" (No College), "1" (Open), and "2" (Liberal/Traditional). Income is significantly negatively related to attendance at a selective institution (Admissions Policy = 3) versus a highly selective institution (reference category) in Equation M1, but once academic characteristics are included (Equations M2 and M3), the effect becomes non-significant.

The only significant effect of race/ethnicity in Equation M1 is a negative relationship between "Hispanic" and "No College" (relative to the reference category of "Highly Selective"). Once the academic characteristics of ACT® score and grade point average (GPA) are included in Equation M2, being Black or Hispanic is negatively related to non-college attendance, attendance at an open institution, or attendance at a liberal/traditional institution. Only being Black is related (negatively) to attendance at a selective institution relative to a highly selective institution with ACT® score and GPA in the model.

ACT® score and GPA are significantly negatively related to attendance at all categories of institution relative to highly selective institutions in both equations in which they are included (Equations M2 and M3). Of the remaining academic characteristics included only in Equation M3, College Preparatory Track, Advanced Courses, Foreign Language, Math Course-Taking Pattern 1 (any combination of three to three and a half years of math), 3 [Algebra I, Algebra II, and Geometry (minimum core)], 4 (Algebra I, Algebra II, Geometry, and Trigonometry or another advanced math course), and Total Extracurricular Participation had statistically significant effects on one or more category of outcome. Only Math Course-Taking Patterns 2 (any combination of four or more years of math) and 5 (Algebra I, Algebra II, Geometry, Trigonometry, and Calculus or another advanced math course) were not significantly related to any level of outcome relative to attendance at a highly selective institution. Tables 24 and 25 present comparative results of Equations M1, M2, and M3. In order to simplify the presentation of the predictor variable comparisons across models in Table 25, one level of dependent variable was chosen. The reference category is “Highly Selective” admissions policy; I chose the adjacent category of “Selective” admissions policy on which to illustrate the behavior of the variables.

Table 24

Summary of Multinomial Logistic Regression Models

Criterion	Equation M1	Equation M2	Equation M3
Assumptions Violated	None	None	Linearity in the logit
Likelihood Ratio Test	p <.0001	p <.0001	p <.0001
Pearson χ^2 Test	p <.01	n.s.	n.s.
Deviance χ^2 Test	n.s.	n.s.	n.s.
McFadden’s R_L^2	.045	.169	.191
% Correctly Predicted	37.6	44.0	46.3

Predicted Dependent Variable Values[a]	0, 2, 3, 4	0, 1, 2, 3, 4	0, 1, 2, 3, 4
Somers' D_{yx}	.349	.513	.503
Diagnostic Analyses	No outliers	Not performed	Not performed

[a] Observed dependent variable values are 0, 1, 2, 3, and 4.

Table 25

Behavior of Predictors across Multinomial Logistic Regression Models for Selective Admissions Policy

Variable	Equation M1 e^B (*odds ratio*)	Equation M2 e^B (*odds ratio*)	Equation M3 e^B (*odds ratio*)
Income[a]	0.99***	1.00	1.00
Siblings	0.92	0.91	0.91
Black	0.83	0.24***	0.25***
Hispanic	1.22	0.61	0.64
ACT® Score		0.83***	0.85***
GPA		0.41***	0.45***
College Preparatory Track			0.97
Advanced Courses			0.87
Foreign Language			0.60**
Math Course-Taking Pattern 1			4.82*
Math Course-Taking Pattern 2			2.08
Math Course-Taking Pattern 3			10.37**
Math Course-Taking Pattern 4			2.97
Math Course-Taking Pattern 5			2.08

Total Extracurricular Participation	0.92**

Note. N = 5,200. Highly selective admissions policy is the dependent variable reference category. The test of the hypotheses of beta = 0 are based on Wald's X^2, df = 1.
[a] Measured in units of $1,000.
*p < .05, **p < .01, ***p < .001

Ordered Logistic Regression

Equation O1

An ordered logistic regression model in which admissions policy is regressed on the student ascriptive characteristics of income, siblings, and race/ethnicity was fitted to the data. Estimated thresholds between the adjacent categories of responses (Admissions Policy = 0 to 3) are significant at the α = .001 level, suggesting that the implied ordering of the dependent variable, selectivity level, is reasonable. The Likelihood Ratio χ^2 test is significant [$\chi^2(5) = 626.76$, $p < .001$], indicating that the ordered logistic model with three predictors is more effective than an intercept-only, or null, model at predicting cumulative probabilities for each level of the dependent variable.

The deviance criterion [$\chi^2(1083) = 1{,}184.29$; $p < .05$] and Pearson [$\chi^2(1083) = 1{,}210.41$; $p < .01$] goodness-of-fit tests were both significant, suggesting that the ordered logistic regression model is ill fit against actual outcomes. However, the χ^2 goodness-of-fit tests will find even small differences significant in large samples; therefore, following the recommendation of Garson (2011a), a random sample of 200 was taken and goodness-of-fit tests run again. With the reduced sample, the χ^2 values were non-significant (p = .825; p = 1.0), indicating that perhaps the model fits the actual outcomes better than originally indicated, at least by the inferential goodness-of-fit tests criterion. A descriptive measure of effect size, McFadden's R_L^2, is .041 for the full sample (N = 5,200) in Equation O1, which is close to the smallest possible value of R_L^2 (0.0).

The classification table reveals that the ordered regression model with student ascriptive characteristics as explanatory variables correctly predicted students not attending college with 84.0% accuracy. The only other category predicted was "Liberal/Traditional" institutions (Admissions Policy = 2); no students were predicted to attend open, selective, or highly selective institutions. The overall correct prediction was low at 37.3%. In addition to the classification table, measures of association indicate the degree to which predicted probabilities correspond to actual outcomes. For the present model, Somers' D_{yx} is .340, indicating a 34.0% reduction in errors when predicting outcome (admissions policy) category using estimated probabilities over chance alone.

Statistical tests of individual predictors using the Wald χ^2 statistic show that income, number of siblings, and being of Hispanic ethnicity all have statistically significant ($p < .001$) effects on the log of the odds of selectivity of postsecondary institution attended. Income and number of siblings are significantly positively related to selectivity of institution attended, but being Hispanic, as opposed to being White, has a significant negative effect in the model. The odds of a Hispanic student attending an institution of higher rather than lower selectivity are 0.639 (= $e^{-0.449}$) times lower than for a White student. Detailed results of Equation O1 are presented in Table 26.

Table 26

Ordered Logistic Regression Results: Equation O1

Variable	*B*	SE *B*	e^B (*odds ratio*)	Wald's χ^2 (*df*=1)	Sig. (*p*)
Income[a]	.020	.001	1.02	447.70	<.0001

Siblings	-.076	.019	.93	15.84	<.0001
Black	-.153	.092	.86	2.76	.097
Hispanic	-.449	.085	.64	27.61	<.0001
Threshold Values					
Admissions Policy = 0	.249	.070	NA	12.63	<.001
Admissions Policy = 1	1.29	.072	NA	318.08	<.0001
Admissions Policy = 2	2.40	.078	NA	945.53	<.0001
Admissions Policy = 3	3.54	.089	NA	1,584.71	<.0001

	χ^2	*df*	*p*
Likelihood Ratio Test	626.76	5	<.0001
Goodness-of-Fit Statistics			
Pearson	1,210.41	1,083	.004
Deviance	1,184.29	1,083	.017

Note. N = 5,200. McFadden's R_L^2 = .041. Somers' D_{yx} = .340. [a] Measured in units of $1,000.

Equation O2

A second ordered logistic regression model was fitted to examine the influence of the student ascriptive characteristics as well as the student academic characteristics of ACT® score and grade point average (GPA) on the log of the odds of selectivity of postsecondary institution attended. Estimated thresholds between the adjacent categories of responses (Admissions Policy = 0 to 3) are significant at the α = .001 level, suggesting that the implied ordering of the dependent variable, selectivity level, is reasonable. The Likelihood Ratio χ^2 test is significant [$\chi^2(7)$ = 2,254.67, p < .001], indicating that the ordered logistic model with five predictors is more effective than an intercept-only, or null, model at predicting cumulative probabilities for each level of the dependent variable.

The deviance criterion [$\chi^2(18101) = 11{,}820.14$; $p = 1.0$] and Pearson [$\chi^2(18101) = 16{,}081.42$; $p = 1.0$] goodness-of-fit tests were both non-significant, suggesting that the ordered logistic regression model is fit to the data well. Additionally, McFadden's R_L^2 is .146.

The classification table reveals that the prediction for students who did not attend college was more accurate than for any other outcome category (71.5% specificity). The next highest specificity is 47.2% for correctly predicting attendance of those students who attended "Liberal/Traditional" institutions. The overall correct prediction was low at 43.0%, although this is an improvement over the ordered model with only ascriptive predictors. In contrast to Equation O1 in which only two of the admissions policy response categories were predicted, all five levels of selectivity were predicted in Equation O2. There does appear to be over-prediction in the two categories that were predicted in Equation O1, "No College" (49.0% of all students) and "Liberal/Traditional" institution (31.1% of all students), though. In addition to the classification table, measures of association indicate the degree to which predicted probabilities correspond to actual outcomes. For the present model, Somers' D_{yx} is .504, indicating a 50.4% reduction in errors when predicting outcome (admissions policy) category using estimated probabilities over chance alone.

Income and number of siblings are statistically significantly related to selectivity of institution attended in Equation O2, as they are in Equation O1. Once the academic characteristics of ACT® score and GPA are added in Equation O2, being of Hispanic ethnicity is no longer statistically significantly related to the log odds of the outcome. However, Black students are more likely ($e^{0.769} = 2.16$) than White students to attend institutions of a certain selectivity (or beyond) once ACT® scores and GPA are taken into consideration.

The academic characteristics of ACT® score and GPA are both significantly positively related to the log odds of selectivity of postsecondary institution attended. With a one-point increase in ACT® score, the odds of attending a college in a particular category of admissions policy or above increase by a factor of 1.18. The effect for a one-point increase in GPA was even greater: students with higher GPAs are 2.4 times more likely than their lower-GPA counterparts to attend institutions of increasing selectivity. Detailed results of Equation O1 are presented in Table 27.

Table 27

Ordered Logistic Regression Results: Equation O2

Variable	*B*	SE *B*	e^B (*odds ratio*)	Wald's χ^2 (df=1)	Sig. (*p*)
Income[a]	.011	.001	1.01	123.18	<.0001
Siblings	-.072	.020	.93	13.16	<.001
Black	.769	.098	2.16	61.92	<.0001
Hispanic	.155	.091	1.17	2.92	.088
ACT® Score	.169	.007	1.18	519.56	<.0001
GPA	.873	.051	2.40	289.53	<.0001
Threshold Values					
Admissions Policy = 0	5.81	.168	NA	1200.76	<.0001
Admissions Policy = 1	7.09	.177	NA	1610.21	<.0001

Admissions Policy = 2	8.51	.189	NA	2040.22	<.0001
Admissions Policy = 3	9.90	.200	NA	2445.69	<.0001

	χ^2	*df*	*p*
Likelihood Ratio Test	2,247.61	7	<.0001
Goodness-of-Fit Statistics			
Pearson	16,081.42	18,101	1.0
Deviance	11,820.14	18,101	1.0

Note. N = 5,200. McFadden's R_L^2 = .146. Somers' D_{yx} = .504. [a] Measured in units of $1,000.

Equation O3

The ordered logistic regression model in which admissions policy is regressed on all student ascriptive and student academic characteristics is more effective than an intercept-only, or null, model [$\chi^2(16) = 2,435.38$; $p < .001$]. According to both the deviance criterion [$\chi^2(20732) = 12,951.88$; $p = 1.0$] and the Pearson χ^2 test [$\chi^2(20732) = 18,715.22$; $p = 1.0$], the model is fit to the data well. McFadden's R_L^2 descriptive measure of effect size is .158.

The classification table reveals that the prediction for students who did not attend college was more accurate than for any other outcome category (71.2% specificity). The next highest specificity is 48.6% for correctly predicting attendance of those students who attended "Liberal/Traditional" institutions. The overall correct prediction rate (44.0%) represents a one-percent increase from Equation O2. As with Equation O2, there appears to be over-prediction in the categories of "No College" and "Liberal/Traditional" institution (47.8% and 30.7% of all students, respectively) relative to the other three categories. For the model O3, Somers' D_{yx} is .524, indicating a 52.4% reduction in errors when predicting outcome (admissions policy) category using estimated probabilities over chance alone. This is an increase in error reduction from Equation O2.

A one-unit change in the predictor variables of income or number of siblings has the same significant positive effect on the outcome in Equation O3 as it does in Equations O1 and O2. When all explanatory variables are in the regression equation, being Black as opposed to being White statistically significantly increases the odds of attending an institution of a given selectivity level or above by a factor of 1.78. The inclusion of the additional student academic characteristics from Equation O2 to Equation O3 does not affect the impact of ACT® score and grade point average (GPA): both are still statistically significantly ($p < .001$) positively related (GPA more so than ACT score; odds ratios of 1.99 and 1.15 respectively) to selectivity of postsecondary institution attended.

Of the additional student academic characteristics present in Equation O3, College Preparatory Track, Foreign Language, Math Course-Taking Pattern 2 (any combination of four or more years of math), 5 (Algebra I, Algebra II, Geometry, Trigonometry, and Calculus or another advanced math course), and Total Extracurricular Participation have statistically significant positive effects on selectivity of postsecondary institution attended. Detailed results of Equation O3 are presented in Table 28.

Table 28

Ordered Logistic Regression Results: Equation O3

Variable	B	SE B	e^B (*odds ratio*)	Wald's χ^2 (df=1)	Sig. (p)
Income[a]	.009	.001	1.01	87.05	<.0001
Siblings	-.082	.020	.92	16.61	<.0001
Black	.575	.100	1.78	33.13	<.0001

Hispanic	.049	.092	1.05	.281	.596
ACT® Score	.138	.008	1.15	298.81	<.0001
GPA	.686	.054	1.99	160.74	<.0001
College Preparatory Track	.372	.059	1.45	40.18	<.0001
Advanced Courses	.039	.064	1.04	.364	.546
Foreign Language	.340	.059	1.41	33.02	<.0001
Math Course-Taking Pattern 1	-.002	.145	1.0	.000	.989
Math Course-Taking Pattern 2	.296	.103	1.34	8.24	.004
Math Course-Taking Pattern 3	-.013	.095	.99	.018	.894
Math Course-Taking Pattern 4	.118	.095	1.13	1.53	.217
Math Course-Taking Pattern 5	.334	.104	1.40	10.32	.001
Total Extracurricular Participation	.072	.012	1.08	34.20	<.001
Threshold Values					
Admissions Policy = 0	5.16	.184	NA	788.20	<.001
Admissions Policy = 1	6.48	.191	NA	1,144.55	<.001
Admissions Policy = 2	7.95	.202	NA	1551.71	<.001
Admissions Policy = 3	9.36	.212	NA	1942.97	<.001

	χ^2	*df*	*p*
Likelihood Ratio Test	2,435.38	16	<.0001
Goodness-of-Fit Statistics			
Pearson	18,715.22	20,732	1.0
Deviance	12,951.88	20,732	1.0

Note. N = 5,200. McFadden's R_L^2 = .158. Somers' D_{yx} = .523. [a] Measured in units of $1,000.

Diagnostic Analyses

Results of the deviance and Pearson χ^2 goodness-of-fit tests indicated that Equation O1 does not fit the data well. Although the inferential tests were re-run on a smaller sample and found to be non-significant, residuals were calculated in the complete dataset by subtracting the predicted category probability from the actual category probability and examined to identify potential outliers. A scatterplot and boxplot of theresiduals indicate two or three possible outliers, but these are separated from each other and the rest of the data by merely .03.

Two Pearson residuals (representing seven cases) were noticeably larger than the rest and may indicate covariate patterns that are not explained well by Equation O1. As in the multinomial logistic regression, the covariate patterns of the observation with the two largest residual values do not match the patterns of the two most extreme Pearson residual values; however, the potential residuals outliers' covariate patterns do fall in the top 15 largest Pearson residuals. No covariate patterns or individual observations consistently presented as outliers, so I concluded that outliers were not the cause of the poor fit of the model found in the Pearson χ^2 and deviance test results in the complete dataset (N = 5,200).

Both the deviance and Pearson criteria indicate that Equations O2 and O3 are fit to the data well; therefore, outlier analysis is not necessary (Tabachnick & Fidell, 2001).

Summary

All three models estimated with ordered logistic regression represented improvement over an intercept-only model as evidenced by significant Likelihood Ratio χ^2 tests. Results of the inferential χ^2 goodness-of-fit tests largely indicate that the models are fit to the data well. The deviance criterion and Pearson χ^2 goodness-of-fit tests were significant in Equation O1, but once a random sample of 200 was drawn to correct for the sensitivity of the χ^2 test in large samples, these inferential goodness-of-fit tests were non-significant. There is a large increase in values for the descriptive measure of strength of association, McFadden's R_L^2, from Equation O1 to Equation O2 when ACT® score and GPA are added, but the increase is much less from Equation O2 to O3 with the addition of other academic variables.

In terms of predicted probabilities, classification tables show an increase in percent correctly predicted from 37.3% in Equation O1 to 43.0% in Equation O2. The Somers' D_{yx} measure of association also increased, from D_{yx} = .34 in Equation O1 to D_{yx} = .504 in Equation O2. When additional academic characteristics were included in the model (Equation O3), the correct prediction rate increased, but only by a small factor (43.0% to 44.0%). However, Somers' D_{yx} increased to .523 in Equation O3 when all 10 explanatory variables are included.

Among the student ascriptive characteristics, the income and sibling variables remain statistically significant across all three models. The "Black" variable becomes significantly positively related to the log odds of selectivity of postsecondary institution attended only once the academic characteristics are included in the model. Being of Hispanic race/ethnicity decreases in its influence on selectivity of institution attended, going from a significant negative influence in Equation O1 to a non-significant positive influences of decreasing magnitude in Equations O2 and O3.

ACT® score and GPA have significant positive effects in both equations in which they are included. Of the remaining academic characteristics included only in Equation

O3, being enrolled in a college preparatory track; taking three or more years of foreign language; taking four or more years of math (Math Course-Taking Pattern 2); taking the math course pattern of Algebra I, Algebra II, Geometry, Trigonometry, and Calculus or another advanced math course (Math Course-Taking Pattern 5); and participating in an increasing number of extracurricular activities are all significantly positively related to selectivity level. Tables 29 and 30 present comparative results of Equations O1, O2, and O3.

Table 29

Summary of Ordered Logistic Regression Models

Criterion	Equation O1	Equation O2	Equation O3
Assumptions Violated	Parallel regression	Parallel regression	Parallel regression Linearity in the logit
Likelihood Ratio Test	p <.0001	p <.0001	p <.0001
Pearson χ^2 Test	p <.01	n.s.	n.s.
Deviance χ^2 Test	p <.05	n.s.	n.s.
McFadden's R_L^2	.041	.146	.158
% Correctly Predicted	37.3	43.0	44.0
Predicted Dependent Variable Values[a]	0, 2	0, 1, 2, 3, 4	0, 1, 2, 3, 4
Somers' D_{yx}	.340	.504	.524
Diagnostic Analyses	No outliers	Not performed	Not performed

[a] Observed dependent variable values are 0, 1, 2, 3, and 4.

Table 30

Behavior of Predictors across Ordered Logistic Regression Models

Variable	Equation O1 e^B (*odds ratio*)	Equation O2 e^B (*odds ratio*)	Equation O3 e^B (*odds ratio*)
Income[a]	1.02***	1.01***	1.01***
Siblings	0.93***	0.93***	0.92***
Black	0.86	2.16***	1.78***
Hispanic	0.64***	1.17***	1.05
ACT® Score		1.18***	1.15***
GPA		2.40***	1.99***
College Preparatory Track			1.45***
Advanced Courses			1.04
Foreign Language			1.41***
Math Course-Taking Pattern 1			1.00
Math Course-Taking Pattern 2			1.34**
Math Course-Taking Pattern 3			0.99
Math Course-Taking Pattern 4			1.13
Math Course-Taking Pattern 5			1.40**
Total Extracurricular Participation			1.08**

Note. N = 5,200. The test of the hypotheses of beta = 0 are based on Wald's X^2, df = 1.
[a] Measured in units of $1,000.
**p < .01, ***p < .001

Effects of Model Specification

In an attempt to correctly model the student college choice decision process, three regression models were run for each statistical method (classical linear, multinomial logistic, and ordered logistic regression). Roughly simulating the methodology of Hearn (1991), the first model included only student background characteristics, and the second model added the approximately continuous student academic characteristics of ACT® score and GPA. The third, full model included all 10 student ascriptive and academic explanatory variables.

When choosing a model, one strategy is to compare competing models on such characteristics as overall model significance, goodness-of-fit, predictive power and accuracy, statistical tests of individual predictors, and results of diagnostic analyses (Peng & So, 2002). Additionally, assumptions of the regression methods may be met or violated based on the inclusion or exclusion of predictor variables. The effects of model specification (adding variables to the model) on evaluation of assumptions, significance of the model, inferential tests of goodness-of-fit, strength of association, predictive ability, behavior of predictors, and diagnostic analyses are discussed below, with emphasis given to whether these effects differ across regression method. Table 31 summarizes the criterion for selecting an appropriate model for describing influences on the student college choice process of selectivity of college attended.

Assumptions

The classical linear regression model (CLRM) requires many assumptions to be met in order for the desired properties of the ordinary least squares estimator to hold. In the present study, many of these assumptions are met, but Equation C1 violates more of them than do Equations C2 and C3 (see Table 19). Although non-stochasticity cannot be tested statistically, for a predictor variable and the population error to be independent of each other, variables not included in the equation should not be causes of the outcome variable. Equation C1 only includes student background characteristics, and it is reasonable (and proved through results of the main analysis in Equations C2 and C3) that omitted student academic characteristics relate to the outcome variable, admissions policy. Equation C1 also violated the normality of errors assumption, as evidenced by plots and descriptive statistics. It would stand to reason that the residuals were not normally distributed around each predicted score since Equation C1 only predicted about half the actual range of the dependent variable (see Table 19).

Table 31

Model Selection for Describing Influences on Selectivity of College Attended

Criterion	Superior Model		
	CLRM	Multinomial Logistic	Ordered Logistic
Assumptions	C3	—	—
Significance of Model	—	—	—
Effect Size (R^2 or R_L^2)	C2/C3	M2/M3	O2/O3
Predictive Ability	C2/C3	M2/M3	O2/O3
Diagnostic Analyses	—	—	—
Utility of Model	C3	M3	O3

Note. "—" indicates that no model is clearly superior on the given criterion.

Equations C2 and C3 theoretically meet the assumption of non-stochasticity and empirically meet the normality of errors assumption, but they both appeared to be heteroskedastic in residuals scatterplots. Heteroskedastic data could indicate an interaction between a measured predictor variable and one not included in the model or the presence of skewed predictor variables. Equation C3 includes all available variables and is the best attempt out of all three models at meeting homoskedasticity. Data screening procedures identified several possible skewed continuous (or approximately continuous) predictor variables, but this would impact all three models (including Equation C1). Additionally, the assumption of a linear relationship between the predictors and the dependent variable is an issue in all three classical linear regression models and is therefore not a discernable criterion on which to judge the superiority of any CLR model. Based on the evaluation of assumptions, the full model (Equation C3) appears to have a slight advantage over the model with continuous academic predictors (Equation C2) and a great advantage over the model with only student ascriptive characteristics (Equation C1).

Logistic regression methods make fewer assumptions than the CLRM, and all of the multinomial and ordered logistic models in the present study meet the required assumptions of non-autocorrelation, absence of multicollinearity, large sample size, and adequate expected cell frequencies. However, the assumption of a linear relationship between the predictor variables and the logit of the dependent variable is violated in Equations M3 and O3. Analysis with a transformed predictor variable did not produce substantively different results from analysis with the untransformed variable; therefore, this violation can be considered non-important for determining the superiority of competing logistic models. The additional assumption of parallel regression in the ordered logistic method is violated by all three models (Equations O1 through O3) and is therefore not a determining factor in model specification.

Significance of Model

All nine models in the present study represented improvement over intercept-only models, as indicated by significant analysis of variance *F*-tests in the CLRM and significant likelihood ratio χ^2 tests in the logistic models. Therefore, no model emerged as superior based on this criterion.

Goodness-of-Fit

Effect size, or strength of association, is measured by the coefficient of determination (R^2) in the CLRM. Equation C1 has the smallest R^2 by far, explaining 25% less of the variance in the dependent variable than the equation with the two continuous academic characteristics added (see Table 19). Equations C2 and C3 are only separated by a .025 change in R^2, although this change is statistically significant (perhaps due to large sample size). Although Equation C3 has the largest R^2 value (.391), the addition of the five remaining academic variables does not appear to contribute meaningfully to the model.

The Pearson χ^2 and the deviance-based inferential tests measure the fit of logistic models against the data. Of the six inferential goodness-of-fit tests performed in the multinomial logistic regression models (two for each of three models, Equations M1 through M3), only one result indicated poor fit of the model to the data: the Pearson χ^2 in Equation M1. Similar to the results in CLRM, the descriptive measure of effect size (McFadden's R_L^2) increases dramatically (by .124) from Equation M1 to M2 when ACT® score and GPA are added but only increases slightly (by .022) when the remaining five variables are included (see Table 24).

Both inferential tests of goodness-of-fit were significant in the ordered logistic regression model with only student ascriptive characteristics (Equation O1), indicating that this model may not be fit to the data well (Pearson and deviance χ^2 tests on a smaller random sample were not significant). However, these tests were non-significant in Equations O2 and O3. Additionally, McFadden's R_L^2 increases by .105 from Equation O1 to O2 but only increases by .012 to Equation O3. Although these are smaller increases than in the multinomial models, the pattern of increase is the same.

Overall, it appears that ACT® score and GPA contribute greatly to model fit and that the remaining five academic variables contribute little. This effect is consistent across regression method and eliminates Equations C1, M1, and O1 as viable models. The differences between the models with only the two continuous or with all seven academic variables are small and provide inconclusive evidence on a superior model based on this criterion alone.

Predictive Ability

Predictive ability is the accuracy of the model for predicting outcomes – the proportion of outcomes correctly predicted (Porter, 1999). Equation C1 has a truncated range of predicted dependent variable values (see Table 19), indicating an obvious problem in accuracy. Equations C2 and C3 predict outside the observed values of "0" to "4" (Equation C3 more so than Equation C2), but rounding brings the predicted range down to four, as it is in the actual data. In order to validate the predicted values with the actual outcomes, the continuous dependent variable values predicted with CLRM were rounded to the nearest whole number, representing discrete admissions policy categories.

Comparisons of predicted and actual dependent variable categories reveal that Equation C1 has an accuracy rate of 25.9%, with the most correct predictions occurring for the category of “Open” admissions policy. The categories of “Selective” and “Highly Selective” were not predicted at all. Equation C2 makes an improvement in prediction at an overall correct rate of 36.6% with the most correct also occurring in the “Open” category. Only one observation was predicted for the “Highly Selective” category (this happened to be a correct prediction). The results for Equation C3 were quite similar: 38.6% correct prediction, with the “Open” admissions policy being the most accurately predicted and only one observation predicted for “Highly Selective.” The Somers’ D_{yx} measure of association improves greatly from .336 in Equation C1 to .505 in Equation C2 and less so from Equation C2 to Equation C3 (.522).

When comparing predicted and observed discrete values of the dependent variable in the multinomial logistic regression models, Equation M3 predicts the most-accurately (46.3% correct). This is a slight improvement over Equation M2 (44.0%) but a larger improvement over Equation M1 (37.6%). Additionally, Equation M1 fails to predict any observations in the “Open” dependent variable category. Although this was the category with the most predictive accuracy in Equation C1, the two equations are similar in that Equation C1 predicts no cases in the categories of “Selective” and “Highly Selective,” and Equation M1 only predicts 1.6% of all cases to fall in these two categories. Equations M2 and M3 predict for all five levels of the dependent variable like Equations C2 and C3 do, but the most accurate category in the multinomial regressions is “No College” (as opposed to “Open” in CLRM). The Somers’ D_{yx} measure of association improves greatly from .349 in Equation M1 to .513 in Equation M2 and actually decreases to .503 in Equation M3.

The correct classification rate in the ordered logistic regression models increases by 5.7% from 37.3% in Equation O1 to 43.0% in Equation O2. The increase is much less (1.0%) when the five remaining academic characteristic variables are added in Equation O3. As in the multinomial logistic regressions, the category of “No College” is the most correctly-predicted across all three models. Equation O1, like Equation M1, fails to predict any observations in the “Open” category, and like Equation C1, no observations are predicted in the categories of “Selective” and “Highly Selective” admissions policy either. The Somers’ D_{yx} measure of association improves greatly from .340 in Equation O1 to .504 in Equation O2 and less so from Equation O2 to Equation O3 (.524).

The effects of adding predictor variables to the model are the same across regression method, with the exception of the effect on Somers’ D_{yx} from Equation M2 to M3. When adding the continuous academic predictors ACT® score and GPA to the reduced, student ascriptive characteristics-only model, the effects are substantial in all three regression methods. There is an increase in correct prediction rate of 10.7% in CLRM, 6.4% in multinomial logistic regression, and 5.7% in ordered logistic regression. The increase in the Somers’ D_{yx} measure of association is virtually the same in all three methods (.164 increase in the logistic regressions and .169 increase in the CLRM).

The reduced models omitted anywhere from one to three categories when predicting dependent variable values, but when ACT® score and GPA were added, all categories are predicted in each regression method. There was no change in most accurately-predicted category when the variables were added. Based on measures of predictive ability alone, the multinomial logistic regression Equation M1 has a slight advantage over Equation O1 (and both are superior to Equation C1), but results of analyses across methods make it clear that the equation with only student ascriptive predictor variables is inferior to the equations with more variables in terms of predicting outcomes accurately.

Effects of adding the remaining five academic predictor variables are mostly consistent across method: in general, they have little effect. They increase the total correct prediction rate by about the same percentage in all three methods (2.0% in CLRM, 2.3% in multinomial logistic, and 1.0% in ordered logistic). Somers' D_{yx} increases by .02 from Equation C2 to C3 and from Equation O2 to O3. However, the measure of association actually decreases by .01 in the multinomial logistic regression from Equation M2 to M3. All categories are predicted in the models with either the two continuous or all seven academic variables for each regression method, and the most-accurately predicted category remains consistent for all models and methods as well.

Because the changes in predictive ability when adding the five remaining predictor variables are small, choosing between the second and third models is difficult. Equations C3 and O3 have the slight advantage over Equations C2 and O2, but results about Equations M2 and M3 are inconclusive due to the differential impact of adding variables on classification rate and Somers' D_{yx}. On the criterion of predictive ability at least, the choice of model (between the one with only ACT® score and GPA and the one with all academic predictor variables) does have some impact on the decision about appropriate regression method. Equation M2 appears superior to Equations C2 and O2, and Equation M3 predicts outcome category more accurately than the others. However, the measure of association is stronger in both Equations C3 and O3 than in Equation M3.

Behavior of Predictors

The ascriptive student characteristics of income, number of siblings, and race/ethnicity are present in all three models. As academic characteristics are added to the model, the impact of these ascriptive characteristics has the potential to change. Comparisons of the effects of adding variables to the model are most easily made between the classical linear regression and ordered logistic models because a latent, ordered dependent variable is assumed. The multinomial logistic model estimates different coefficients for all but one of the five categories of the admissions policy dependent variable. The category of "Selective" admissions policy is adjacent to the reference category and is chosen for comparison (see Table 25), but results can be expected to differ slightly from those of the CLR and ordered logistic models.

Income is significantly positively related to admissions policy in all three CLR models. Its importance, measured by effects on standard deviation change, decreases by more from Equation C1 to C2 (from $\beta = .315$ to $\beta = .134$) when the continuous academic variables are introduced than when the other five academic variables are added in Equation C3 ($\beta = .110$). The behavior of the income variable is similar across the ordered logistic regressions: it is significantly related to admissions policy in all three models. Income is significantly related to attending a selective versus highly selective postsecondary institution in Equation M1, but once academic characteristics are added to the model, the relationship becomes non-significant.

Number of siblings has a significant, but small (β = -.048), relationship with the dependent variable in Equation C1. This impact remains significant but decreases when ACT® score and GPA are added to the model in Equation C2 and then remains essentially the same when the remaining variables are added in Equation C3. Number of siblings is also significantly negatively related to admissions policy in all three ordered logistic regression models (Equations O1, O2, and O3), and the small impact on the odds of attendance is about the same in each model (see Table 30). Number of siblings is not significantly related to attendance at a selective versus highly selective institution in Equation M1, M2, or M3.

Being Black, as opposed to being White, is not significantly related to admissions policy in Equation C1, but once the academic characteristics of ACT® score and GPA are taken into account, the effect of being Black is significantly positively related to the dependent variable. The same is true for Equation C3, although to a lesser extent (a change in β of -.029). The effects on the race/ethnicity variable category "Black" of adding variables to the model in the ordered logistic regressions mirror the effects in the CLRM. Similar behavior of this variable is observed in the multinomial logistic models: being Black is not significantly related to attendance at a selective versus highly selective institution in Equation M1, but once ACT® score and GPA are taken into account, Black students are significantly less likely than White students to attend selective rather than highly selective institutions. The impact is about the same in Equations M2 and M3 (e^B = 0.24 and 0.25 respectively).

Hispanic students are predicted to attend postsecondary institutions of lower selectivity levels than are White students when only ascriptive characteristics are in the

CLR model (Equation C1). Once ACT® score and GPA are added as predictor variables, this effect remains significant but changes direction: Hispanic students systematically attend institutions of higher selectivity than do White students when these academic characteristics are taken into account, although the impact is not as great (β = -.060 in Equation C1 and β = .034 in Equation C2). The effect of being Hispanic on admissions policy is non-significant in Equation C3. The effects of adding academic variables to the ordered logistic models on the "Hispanic" race/ethnicity category are almost identical to the effects in the CLRM. There is no significant effect of being Hispanic versus being White in the multinomial logistic models for the "Selective" dependent variable category, but the coefficients do change direction in the same pattern that they do in the CLRM and ordered logistic models.

ACT® score and GPA are the only other variables that are included in more than one model. They are significantly associated at the α = .001 level with higher values of the dependent variable in all six models in which they occur. The impact of ACT® score – measured in terms of standardized beta coefficients – diminishes somewhat in the CLRM when the remaining five academic variables are included (from Equation C2 to C3). When the impact of ACT® score is assessed in terms of odds ratios (effect of a one-unit change in the predictor variable on the odds of the dependent variable), however, it remains consistent when variables are added to create the full multinomial and ordered logistic models. Although the relationship between GPA and admissions policy remains significant and positive from Equations C2 to C3 and from Equations O2 to O3, its impact (assessed by either standardized beta coefficients or odds ratios) is greatly reduced in both instances when other academic variables are introduced. The impact of GPA actually increases slightly (from e^B = .041 to e^B = .045) from Equation M2 to M3 but is virtually the same.

Although not useful on its own for choosing among competing models within a regression method, examination of the predictor variables for differential patterns of behavior yields information that is helpful in determining whether choice of statistical method impacts the effects of adding variables to a model. In the present study, the effects of adding academic variables on the student ascriptive characteristics of income, number of siblings, and race are essentially the same for the classical linear, multinomial logistic, and ordered logistic regression methods. Changes in the weight of a variable as others are added may vary slightly across method, but results of the significance tests of predictors (including the ACT® score and GPA variables) are identical for the CLRM and the ordered logistic regression models. The structure of the multinomial logistic method makes it such that direct comparison is not possible with the CLRM and ordered logistic models. In the few instances where predictor significance results differ between the multinomial and other two methods, the direction of the relationship is the same.

Diagnostic Analyses

In the logistic models with poor fit on which diagnostic analyses were performed (Equations M1 and O1 specifically), no outliers were found. Residuals analysis for all three CLRM equations did not indicate the presence of outliers either. Therefore, diagnostic analyses are not an informative criterion on which to base specification of model in any of the regression methods.

Final Model

The effects of adding variables to the model are largely invariant across regression method. The model with only student ascriptive characteristics (Equations C1, M1, and O1) is inferior on every criterion in the CLR, multinomial logistic, and ordered logistic regression methods. The two models with student academic characteristics perform similarly in all three regression methods. The full model with all ascriptive and academic variables included (Equations C3, M3, and O3) surpasses the reduced model in which the categorical academic variables are excluded (Equations C2, M2, and O2) on goodness-of-fit and predictive ability, but only by a small margin, making the differences mostly trivial.

If the only goal in the current study was to choose the most parsimonious model, Equations C2, M2, and O2 would have the advantage over the full models. However, in the interest of informing the student choice literature, Equations C3, M3, and O3 are the most useful for making recommendations to students. In addition to reporting the influences of ACT® score and GPA on selectivity of postsecondary institution attended, the full model also provides information on how other student-controlled variables can impact post-secondary plans. Therefore, regardless of regression method employed, the full model (Equations C3, M3, and O3) surpasses the others on the majority of the criteria for model selection (Table 31) and will be used as the example when making comparisons among the CLR, multinomial logistic, and ordered logistic regression methods.

Chapter 5: Discussion

The current study used a model of student college choice with a categorical dependent variable representing five ordered levels of selectivity of postsecondary institution to examine the effects of using classical linear regression versus parametric regression techniques specifically designed and recommended for categorical dependent variables. Specifically, within-method comparisons were made and it was determined that the impacts of adding variables to a reduced model did not differ based on regression method. Therefore, only one model – the full model with all 10 ascriptive and academic explanatory variables – was chosen for comparisons across three statistical methods: classical linear regression with ordinary least squares estimation and multinomial and ordered logistic regression with maximum likelihood estimation.

Although the full models – Equations C3, M3, and O3 – are comparable on the basis that they used the same sample of students, the use of different estimation techniques prevents direct comparisons on all of the recommended criterion for selecting a superior model. Nevertheless, model performance is analyzed on the basis of assumptions, overall model significance, inferential tests of goodness-of-fit, strength of association, predictive ability, behavior of predictors, and usefulness in practical applications. Implications for student college choice research are discussed, followed by an assessment of the limitations of the current study.

Model Comparisons

As discussed previously, many of the assumptions of the classical linear regression model (CLRM) may be violated for ordinal outcomes. For example, the relationship between explanatory variables and an ordinal dependent variable is not truly linear because the observed ordered variable is discrete and bounded. This violation of the linearity assumption is associated with the violation of the assumptions of normally distributed errors and homoskedasticity. However, the practice of treating ordered categorical dependent variables as if they were measured on an interval scale and using the CLRM is common in the social sciences. The present study compared results of the CLRM with an ordinal-level dependent variable with recommended parametric regression procedures especially designed to support categorical dependent variables in order to determine the degree to which the theoretical violations in the CLRM impacted inferences that can be made about the relationship between student characteristics and selectivity of postsecondary institution attended.

On the basis of assumptions, the CLRM Equation C3 is at a disadvantage theoretically, although empirical evidence (P-P plot; histogram of standardized residuals; skewness and kurtosis values) indicates that the errors are normally distributed. Furthermore, even if the errors vary from normal in the extreme values of the predicted dependent variable, the CLRM is robust to non-normality of errors in large samples (Kreiberg, n.d.). Equation C3 appears to violate the assumption of heteroskedasticity (error variances are greater for the larger values of the predicted dependent variable), but again, the CLRM is relatively robust to modest violations of homoskedasticity (Garson, 2012b). Theoretically, the relationship between the predictors and the discrete admissions policy dependent variable cannot be linear because of the variable's non-continuous nature, but the residuals plot did not indicate a substantial departure. Additionally, linear regression may be appropriate for use with observed discrete variables when information about the interval sizes of the latent continuous variable is known. Threshold values obtained in ordered logistic regression provide the information for determining whether the distances between observed values are equal. The distances between the threshold values in the full model, Equation O3, are very similar (1.32, 1.47, and 1.41), indicating that the underlying construct of admissions policy (selectivity of institution) may be interval-level.

The logistic regression models do not make as many assumptions about data generation as does the CLRM, but Equations M3 and O3 do not escape scrutiny on this criterion. Proper sampling and specification of the model are important for all regressions in order to prevent problems of selectivity bias, endogeneity, multicollinearity, and autocorrelation (Garson, 2012a; Peng, Lee, & Ingersoll, 2002; R. Toutkoushian, personal communication, Fall 2005; Tabachnick & Fidell, 2001). Specific to logistic regression, there was evidence of violation of the assumption of a linear relationship between explanatory variables and the logit of the dependent variable in both the multinomial and the ordered logistic models, and when the complete sample was used (N = 5,200), Equation O3 violated the very important parallel regression assumption. However, the parallel lines test is sensitive in large samples, and a random sample of 200 observations yielded a non-significant result.

The CLRM, multinomial logistic, and ordered logistic models all represented improvement over intercept-only models, as indicated by a significant analysis of variance F-test of Equation C3 and significant likelihood ratio χ^2 tests of Equations M3 and O3.

Direct comparisons of effect size are not possible across models as the OLS coefficient of determination, R^2, does not have an equivalent for logistic regressions.

Equation C3 explains 39.1% of the variance in the dependent variable by the 10 explanatory variables. The Pearson χ^2 and the deviance-based inferential tests, which measure the fit of logistic models against the data, were non-significant in Equations M3 and O3, indicating a good fit to the data. McFadden's R_L^2, which represents the strength of association or effect size, is greater for Equation M3 (R_L^2 = .191) than for Equation O3 (R_L^2= .158).

Table 32

Model Significance and Effect Size for Classical Linear, Multinomial Logistic, and Ordered Logistic Regression Models

	Equation C3	Equation M3	Equation O3
Overall Model Evaluation	*F*-test: Sig.	LR χ^2 test: Sig.	LR χ^2 test: Sig.
Effect Size	R^2 = .391	R_L^2 = .191	R_L^2 = .158

Note. N = 5,200. R^2 and R_L^2 are not directly comparable.

Predictive Ability

In addition to overall model evaluation and goodness-of-fit information, models can be compared based on their predictive ability, or the accuracy of the model for predicting outcomes. For example, upon finding no statistical differences in fit or coefficient behavior among the logit, probit and linear regression models, Dey and Astin (1993) assessed the models on their ability to predict college student retention. The researchers found that all three techniques had similar levels of success at predicting retention and therefore concluded that there was little practical difference among the techniques (Dey & Astin, 1993).

Porter (1999) also studied effects on college student retention and compared a logistic regression model and a multinomial logit model. Although the likelihood ratio index pseudo-R^2 (R_L^2) resulted in a larger value for the multinomial logit model, analyses of predictive ability caused Porter (1999) to conclude that the multinomial logit model is not useful for actually predicting outcomes. Thus, examination of classification tables and measures of association provides valuable information for assessing and comparing the predictive ability of models (Dey & Astin, 1993; Hilmer, 2001; Menard, 2000; Peng & Naegle Nichols, 2003; Peng & So, 2002; Porter, 1999).

Whereas discussion of the statistical significance of predictors is only valid when there is a random sample, which is not the case in the present data, analysis of predictive ability is relevant to research inferences in any sample (Garson, 2010). The classical linear regression model achieves less predictive accuracy (38.6%) than do the logistic models, which is not surprising considering that the CLRM estimates continuous values outside of the actual discrete values of 0 to 4. The multinomial logistic regression model predicted with 46.3% accuracy, and the ordered logistic model performed slightly worse, achieving 44.0% accuracy. Evidence from the Somers' D_{yx} measure of association presents a different picture of accuracy: the CLRM and the ordered logistic models have nearly identical D_{yx} values, whereas the multinomial logistic model has the lowest D_{yx} value. See Table 33 for more details regarding the classifications of the three models.

Table 33

Classification Tables for Classical Linear, Multinomial Logistic, and Ordered Logistic Regression Models

Actual Admissions Policy	Predicted Admissions Policy	Predictive Efficiency

Equation C3[a]	No College	Open	Liberal/ Trad.	Selective	Highly Selective	
No College	**652**	900	289	63	0	% correct = 38.6 D_{yx} = .522
Open	242	**711**	256	6	0	
Liberal/ Traditional	28	455	**530**	57	0	
Selective	5	105	393	**112**	0	
Highly Selective	0	5	184	206	**1**	
Equation M3	No College	Open	Liberal/ Trad.	Selective	Highly Selective	
No College	**1467**	102	229	46	60	% correct = 46.3 D_{yx} = .503
Open	804	**163**	215	21	12	
Liberal/ Traditional	403	95	**448**	77	47	
Selective	124	24	241	**122**	104	
Highly Selective	25	1	91	71	**208**	
Equation O3	No College	Open	Liberal/ Trad.	Selective	Highly Selective	
No College	**1356**	164	303	50	31	% correct = 44.0 D_{yx} = .523
Open	764	**159**	280	11	1	
Liberal/ Traditional	313	138	**520**	84	14	
Selective	47	44	348	**132**	44	
Highly Selective	3	2	143	129	**120**	

Note. N = 5,200. Values in bold represent correct predictions.

[a] The continuous predicted dependent variable values were rounded to the nearest whole number to represent the discrete admissions policy categories.

In addition to overall predictive accuracy of the models, accuracy within the categories of the dependent variables varies by model. Table 34 depicts the predictive accuracy of the CLRM, the multinomial logistic, and the ordered logistic models with respect to each category of the dependent variable, admissions policy. For example, only one student is predicted to attend a highly selective postsecondary institution (Admissions Policy rounds to “4”) in the CLRM although 397 students in the dataset actually attended a highly selective institution (0.25% accuracy). The CLRM is best at predicting for the lower three categories (45.2% combined accuracy) than for the categories of selective and highly selective admissions policy (11.2% combined accuracy).

The multinomial and ordered logistic models, although representing improvements in total predictive accuracy over the CLRM, differ in their abilities to correctly predict attendance among the levels of the dependent variable as well. The multinomial model predicts non-college attendance (Admissions Policy = “0”) with the greatest accuracy (77.0%), followed by prediction of attendance at highly selective postsecondary institutions (Admissions Policy = “4”; 52.5%) and then at liberal/traditional institutions (Admissions Policy = “2”; 41.9%). The odd-numbered categories of the dependent variable (“Open” and “Selective” institutions) have the least accuracy in the multinomial model. A similar pattern is observed in the ordered logistic regression model, although the accuracy rate is slightly different than that of the multinomial model: 71.2% accuracy for predicting students to not attend college; 48.6%

accuracy for predicting students to attend a liberal postsecondary institution; and 30.3% accuracy for predicting attendance at a highly selective institution. Taken together, the multinomial model predicts the top two most selective levels of admissions policy with 32.6% accuracy whereas the ordered model predicts the same levels with 24.9% accuracy (both greater than the 11.2% accuracy of the CLRM for the two most selective categories).

Table 34

Predictive Accuracy of Models by Admissions Policy Category

	CLRM	Multinomial Logistic	Ordered Logistic	
Admissions Policy	Correctly Predicted			Observed
No College	652 (34.2%)	1467 (77.0%)	1356 (71.2%)	1904
Open	711 (58.5%)	163 (13.4%)	159 (13.1%)	1215
Liberal/Traditional	530 (49.6%)	448 (41.9%)	520 (48.6%)	1069
Selective	112 (18.2%)	122 (19.8%)	132 (21.5%)	615
Highly Selective	1 (0.25%)	208 (52.4%)	120 (30.2%)	397

Table 35 shows the degree to which the three models' predictions of admissions policy category approximate the actual data. The CLRM under-predicts the amount of students not attending college (927 predicted compared to 1,904 actual), whereas the multinomial and ordered logistic models over-predict students to not attend college (2,823 and 2,483, respectively). However, the CLRM over-predicts the next-highest selectivity level, open admissions, by 18.4%, and the logistic regression models greatly under-predict this category, resulting in similar accuracy across models for the lowest two levels (all close to the observed rate of 60.0% attendance). All models under-predict attendance at the top two levels of selectivity, although the CLRM is the farthest off in its predictions.

The models differ on relative position of the dependent variable category, when ordered descendingly by number of cases predicted, as well. The ordered logistic regression most-closely matches the actual order of categories, only transposing the second- and third-most numerous categories. The multinomial logistic regression only correctly predicted the admissions policy category with the greatest observations in the actual data, no college attended. In contrast, the CLRM correctly predicted the fourth and fifth most frequently-observed categories, selective and highly selective institutions.

Table 35

Serial Position of Admissions Policy Frequency for Observed and Predicted Data

Admissions Policy	Observed Rank[a] (Percent of *N*)	Predicted Rank (Percent of *N*)		
		CLRM	Multinomial Logistic	Ordered Logistic
No College	1 (36.6)	3 (17.8)	**1** (54.3)	**1** (47.8)
Open	2 (23.4)	1 (41.8)	4 (7.4)	3 (9.8)
Liberal/Traditional	3 (20.6)	2 (31.8)	2 (23.5)	2 (30.7)
Selective	4 (11.8)	**4** (8.5)	5 (6.5)	**4** (7.8)
Highly Selective	5 (7.6)	**5** (0.02)	3 (8.3)	**5** (4.0)

Note. N = 5,200. Values in bold represent serial positions that agree with the observed data. [a] Dependent variable category with the greatest count = 1; smallest count = 5.

On the basis of predictive ability, the classical linear regression model falls short of the two logistic regression models. Although Somers' D_{yx} for the CLRM suggests that the association between predicted and actual outcomes in the CLRM is as strong as it is in the ordered logistic regression model, the CLRM predicts with only 38.6% accuracy as opposed to the ordered regression's 44.0% accuracy. Furthermore, the CLRM over-predicts the lowest three values of the dependent variable and is far less accurate at predicting the top two values of the dependent variable than are the multinomial and ordered logistic regressions (11.2% accuracy compared to 32.6% and 24.9% accuracies, respectively).

Between the multinomial and ordered logistic regression models, the choice of the better predictor is less obvious. Their overall predictive accuracy is within 2.3% of each other, with the multinomial model having the slight advantage, and their measures of association are within .02 of each other, with the ordered model having the advantage on this criterion. Both the multinomial and the ordered logistic models' most-accurately predicted category is non-college attendance, the lowest level of admissions policy. The multinomial model predicts attendance at highly selective institutions more accurately than does the ordered logistic model, but the ordered logistic model is better at predicting attendance at selective institutions. Finally, the ordered logistic model's frequency distribution more-closely represents that of the actual data. This discrepancy between the multinomial model's predictions and the observed data could be due to the fact that the sample used in the current study is unbalanced, meaning that the cases are not distributed evenly across all five outcomes, resulting in ambiguous predictions for many individuals (Porter, 1999).

Behavior of Predictors

Although the coefficients in linear and logistic regression are on different scales and cannot be compared directly, predictors can be examined for differential behavior, in terms of significance, direction, and standardized impact, among methods. In the current study, choice of statistical method has little effect on the behavior of the predictors. The same predictors have significant effects of the same direction on the dependent variable (or the log odds of the dependent variable) in the CLRM and in ordered logistic regression. When effects of the predictor variables are standardized, the relative importance of each remains essentially the same for the CLRM and the ordered logistic regressions. A one standard deviation increase in ACT® score has the greatest impact on the change in standard deviation of the dependent or the log odds of the dependent, followed by grade point average and income. Math Course-Taking Patterns 5 and 2 (Algebra I, Algebra II, Geometry, Trigonometry, and Calculus or another advanced math course or any combination of four or more years of math) and number of siblings all have the least amount of standardized impact.

Comparisons on the behavior of predictors in the CLRM and ordered logistic regressions with multinomial logistic regression are more complicated due to the different modeling procedure in the latter. Four different logit coefficients are estimated for each predictor variable, depending on the level of dependent variable being modeled. As was done when comparing equations within regression method, only one category of dependent variable, “Selective” institution (reference category: “Highly Selective”), was selected for simplicity. The multinomial method produces different results regarding significant effects on the log odds of the dependent variable than do the CLRM and ordered logistic methods. For example, income is not a significant predictor of odds of attendance at a selective versus highly selective institution, nor is number of siblings. Taking a college preparatory track of courses is not significantly related to the odds of attending a selective versus highly selective institution (although it is significantly negatively related to the odds of not attending any postsecondary institution).

Direction of relationship between the significant predictors and the log odds of the dependent variable is substantively consistent with the CLRM and logistic methods, although the actual signs of the coefficients are reversed in the multinomial model due to the reverse ordering in the comparison of dependent variable categories (categories increase in selectivity in the CLRM and logistic regression, but the comparison in the multinomial model is between the odds of attending an institution of lower versus higher selectivity). Along these same lines, the behavior of the categories of the math course-taking pattern variable appears different in the multinomial logistic regression than in the other two methods but is actually substantively the same. Math Course-Taking Patterns 1, 3, and 4 are significant in the multinomial model whereas patterns two and five are significant in the CLRM and logistic models. However, it makes sense that the course-taking patterns associated with higher selectivity levels would not be associated with attendance at an institution of lower selectivity, which is the comparison modeled in the present multinomial logistic regression.

The relative weight of the standardized predictors in the multinomial logistic regression mirrors that of the other two models, with the exception that income is non-significant in the multinomial regression. Table 36 shows the behavior of predictors across models in terms of significance and impact of a one standard deviation change on the change in standard deviation of either the dependent variable (classical linear regression) or the log odds of the dependent variable (logistic regression models).

Table 36

Behavior of Predictors across Classical Linear, Multinomial Logistic, and Ordered Logistic Regression Models

	Equation C3		Equation M3[a]		Equation O3	
Predictor Variable	Sig. (p)	β^b	Sig. (p)	β^c	Sig. (p)	β^c
Income[d]	<.0001	.110	n.s.		<.0001	.278
Siblings	.001	-.035	n.s.		<.0001	-.111
Black	<.0001	.068	<.0001	-.400	<.0001	.170
Hispanic	n.s.		n.s.		n.s.	
ACT® Score	<.0001	.302	<.0001	-.840	<.0001	.699
GPA	<.0001	.178	<.0001	-.538	<.0001	.473
Col. Prep. Track	<.0001	.075	n.s.		<.0001	.186
Adv. Courses	n.s.		n.s.		n.s.	
For. Lang.	<.0001	.080	.004	-.248	<.0001	.168
Mat. Crse. Pat. 1	n.s.		.011	.354	n.s.	
Mat. Crse. Pat. 2	.003	.050	n.s.		.004	.121
Mat. Crse. Pat. 3	n.s.		.002	.959	n.s.	

Mat. Crse. Pat. 4	n.s.		.021	.493	n.s.	
Mat. Crse. Pat. 5	.001	.053	n.s.		.001	.127
Tot. EC Partic.	<.0001	.068	.006	-.200	<.001	.166

Note. N = 5,200. [a] Highly selective admissions policy is the dependent variable reference category. [b] β is the standardized beta coefficient which represents the s.d. change in admissions policy due to a one s.d. change in a predictor. [c] β is the standardized beta coefficient which represents the s.d. change in the log odds of admissions policy due to a one s.d. change in a predictor. [d] Measured in units of $1,000.

Utility of Model

When determining the appropriateness of a model, statistical analyses are not the only criteria on which to rely. The usefulness of a model for actually explaining student behavior is a critical practical aspect to consider (Dey & Astin, 1993; Porter, 1999). Although such explanatory power cannot be measured, it is an important factor when interpreting results of a model (R. Toutkoushian, personal communication, Fall 2005). A model of student college choice should be able to be meaningfully interpreted by interested parties in both secondary and higher education.

The CLRM has an advantage in its linearity: this type of relationship between predictor and outcome variables is more intuitive than relationships to log odds. In a linear model, a one-unit increase in a significant predictor variable, such as ACT® score, is associated with a fixed increase in the value of the dependent variable (.077 in the current study) across the entire range of the predictor. The disadvantage of the CLRM in the present situation, however, is that the dependent variable is not measured on a continuous scale, making these predictable increases mostly nonsensical.

For example, assume a student with an ACT® score of 18 whose predicted dependent variable value is .712, based on Equation C3 and using his values on the nine other predictor variables. If he raised his ACT® score to 19 (a one-unit increase), his predicted dependent variable value would be .789. However, the observed dependent variable is measured on a discrete scale, so a value with decimal points cannot be interpreted meaningfully without conversion to the observed scale. Information is lost in the manipulation of rounding: both of the above predicted values round to “1.” Thus, small changes in the predicted continuous dependent variable will be lost when rounded to the nearest whole value in the CLRM. The beta values in the CLRM Equation C3 are all less than .35, so large increases in the predictor variables would be needed in order to interpret any noticeable differences in the discrete dependent variable.

On the other hand, although interpretations of the logistic models make sense in terms of the measurement level of the dependent variable, log odds (and even odds ratios) are often less-easily understood. Complicating this is the fact that the multinomial logistic regression model estimates four different coefficients for each variable (representing in the current study the J-1 response categories of the dependent variable). Predicting log odds of the dependent variable in the ordered logistic regression model is not straightforward either, as the equation (and thus threshold value) used depends on the category of dependent variable for which the cumulative odds are desired.

The richness of information provided by both logistic models allows exploration of a multitude of very specific student patterns. However, this information might prove to be too overwhelming for high school students and counselors, especially in the multinomial model. A student who wishes to know how improving his ACT® score will affect the selectivity of postsecondary institution he is likely to attend would receive four different answers in relation to the reference category of highly selective institution. The knowledge that his odds of attending a selective, liberal/traditional, or open institution or no college at all compared to attending a highly selective institution are decreased by a factor of either .85, .75, .71, or .72, respectively, for each one-point increase in ACT® score will probably not be very useful to the student.

If a student did have a particular interest in knowing about his odds of attending one type of institution versus attending a highly selective institution, the multinomial model could provide this information. The more likely scenario, though, would be that trends in (rather than specific) odds ratios might be interpreted – in the present example, a counselor might advise this student that increasing his ACT® score will decrease the odds that he will attend an institution of lower rather than higher selectivity. This same general information could be obtained more parsimoniously and more specifically from the ordered logistic regression, however, as the ordered model estimates one coefficient for each predictor: the effect on the log odds of being at or below a given category of the dependent variable. For example, interpreting from the ordered logistic model, a one-point increase in ACT® score improves the odds of attending a postsecondary institution of the next selectivity level by a factor of 1.15, holding all other variables constant.

Conclusion

When choosing among the competing classical linear, multinomial logistic, and ordered logistic regression models for explaining selectivity of postsecondary institution attended in the current study, assumptions, overall significance of the model, goodness-of-fit, effect size, predictive ability, and utility of the models were considered. On the basis of overall model evaluation, a superior model cannot be determined because all three regressions (Equation C3, Equation M3, and Equation O3) represented a statistically significant improvement over intercept-only models. Effect size descriptive statistics cannot be compared across linear and logistic regression methods; however, within the logistic models, the effect size measure McFadden's R_L^2 indicates that the multinomial logistic model has a greater strength of association than does the ordered logistic model.

Although empirical evidence did not wholly discredit the CLRM on the basis of violation of assumptions, the theoretical violation of the linearity assumption – due to the discrete and bounded nature of the observed dependent variable – can be associated with the violation of the assumptions of normally distributed errors and homoskedasticity. Threshold values indicated that there may be equal interval sizes between admissions policy groups, making use of this dependent variable with linear regression tenable, but continuous predictions of the dependent variable value do not make sense when interpreting with the observed categorical outcomes in the current study. These shortcomings of the CLRM, Equation C3, might be overlooked if the model was excellent at explaining the variance in the dependent variable and in accurately predicting admissions policy category, but with the present data, this is not the case.

The observed R^2 in the present study, .391, is actually comparable or superior to those found in previous studies of selectivity of institution attended conceptualized as a continuous dependent variable (Davies & Guppy, 1997; Hearn, 1988, 1991; McDonough & Antonio, 1996). Hearn (1988) used cost of institution as the dependent variable and reported an R^2 of .15 for his full model, and in 1991 he used average SAT score as the dependent variable and found an R^2 of .27 in the full model. Davies & Guppy (1997) replicated Hearn's 1991 study and obtained an R^2 of .114 with seven background and two academic variables in their model. McDonough and Antonio (1996), whose methodology most-closely matched that of the current study, used an ordered categorical dependent variable with the CLRM and observed a range of R^2 from .43 to .52, depending on the race of the student. However, Long (1997) argues that there is not convincing evidence that the method of selecting a model based on a measure of fit such as R^2 has any desirability other than that the model will simply have a larger value of that measure. Other factors such as predictive ability must be taken into account, and in that regard, Equation C3 falls short.

The CLRM in the present study correctly predicts 38.6% of the cases, which is slightly more than what would be expected by chance, when chance is defined as using the most numerous category to predict all cases (36.6% would be correctly classified in this case). Garson (2012a) suggests as an improvement criterion that a model do 25% better at predicting than by chance. The critical value for the current study would be 45.8%, and Equation C3 misses this mark by 9.2%. Additionally, the CLRM only predicts with 11.2% accuracy in the top two levels of admissions policy ("Selective" and "Highly Selective"). When explaining the influence of predictors on selectivity of postsecondary institution attended, particularly with emphasis on how students can increase their chances of attending selective institutions, accuracy in the higher levels of selectivity is desirable.

Although the CLRM excels in its intuitive, linear interpretation of actual dependent variable values, Equation C3's predictions are wrong 63.4% of the time, making it not a very useful model in practice. Furthermore, the interpretation of parameter estimates as effects on actual category of admissions policy (rather than on the log odds of attendance in a category compared to another category/ies), while easy to understand, might give a misleading impression of certainty when in fact there is much error in the prediction. The terminology associated with logistic regression itself conveys a sense of chance by referring to the odds of one outcome versus another or to the likelihood of an event's occurrence, perhaps because these terms (specifically "odds") are often associated with gambling which by definition implies uncertainty. The interpretation of logistic models, while unwieldy, more accurately conveys the actual modeling situation in which effects of variables and predictions made regarding the student college choice process carry some uncertainty. Because of its shortcomings on many of the criteria for evaluating the superiority of competing models, Equation C3 (the CLRM) can be eliminated as an appropriate model for the present student college choice data.

When the assumed ordering of a categorical dependent variable is realistic, ordered logistic regression is an improvement over the multinomial logistic model because it incorporates the information about ordering of the dependent variable, resulting in simpler interpretations and greater power (Agresti, 1996; Zumbo & Ochieng, 2002). Tests for significance of the thresholds in the ordered models can provide a check for this assumption (Hilmer, 2001). The estimated thresholds for Equation O3 are significant, implying a natural ordering of the admissions policy dependent variable. In order to use the proportional odds model, the assumption of parallel regressions must hold (Long, 1997). The test of parallel lines for a reduced, less-sensitive sample (n = 200) was non-significant in Equation O3. On the basis of these results, the ordered logistic model cannot be statistically rejected as the appropriate specification for the current data.

Benefits of the multinomial logistic regression model in the current study are the fact that it accurately predicts at a higher rate than the ordered logistic model and that it offers a wealth of very specific information regarding effects of predictor variables on many combinations of outcome variable categories. The multinomial logistic model exceeds the critical prediction rate of 45.8 % by 0.5%, but the ordered logistic regression misses it by 1.8%. Predictions of selectivity level of postsecondary institution attended may be correct more times when using the multinomial rather than the ordered logistic model, especially in the higher levels of selectivity, but this gain in accuracy is perhaps not worth the loss of interpretative efficiency.

If the ultimate goal of a model of student college choice behavior is to predict the level of selectivity of the college a student will attend based on his or her academic and ascriptive characteristics, then Equation M3 in the current study would be the best choice due to its superiority in predictive accuracy. In practice, however, this goal would not be realistic due to the nature of the data. The identities of the institutions in the current study are unknown, and the admissions policy (selectivity level) is defined by the institutions themselves. Students would not be able to compare actual institutions in which they are interested in attending to the category of admissions policy predicted by either Equation M3 or Equation O3. For example, a student might be predicted to attend a highly selective university, but without knowledge of which institutions are considered highly selective, the information provided by the prediction is incomplete. A general idea of a student's postsecondary options is gained by both models – students can roughly estimate the selectivity level of an institution of interest by finding out how many students are admitted from the top 10%, 25%, 50%, and 75% of the high school graduating class (as admissions policy is defined in the current study) – but the statistical advantage the multinomial model has over the ordered model in terms of predictive accuracy does not make much practical difference.

Because of the lack of ability to meaningfully interpret predicted selectivity level, interpretation in terms of the effects of predictor variables on (log) odds of the dependent variable might be more useful. The fact that the multinomial model estimates four different coefficients for each predictor variable makes this model quite cumbersome to interpret. Such detailed information might be useful for tailoring results to individuals with particular interests, such as the effect of taking three or more years of foreign language on the (log) odds of attending a selective versus a highly selective institution, but it is unlikely that such a narrow impact of a model for student college choice would have widespread practical use.

The logical ordering of the categorical dependent variable also makes it difficult to choose a meaningful reference category for the multinomial logistic model. In the interest of student college choice research, the current study seeks factors that increase the chances of attendance at selective and highly selective institutions. Implied in the statement of this research goal is a natural ordering of the dependent variable, and since this assumption was supported by significant threshold values, it is safe to choose Equation O3 as the most statistically sound and practically useful model of the effects of student characteristics on selectivity of postsecondary institution attended.

Student College Choice Findings

Results of the current study are particular to the sample of 5,200 high school students with complete data. The presence of a large amount of missing data prevents valid inferences about the impacts on selectivity of college attended from being made. However, many student college choice studies (with the exception of Perna and Titus, 2005, who used data imputation) do not explicitly discuss the treatment of missing data or only use cases with complete data (see for example Alexander et al., 1987; Davies & Guppy, 1997; Hearn, 1984, 1988, 1991; Hilmer, 2001). As an informative measure, therefore, findings regarding the impact of ascriptive and academic variables from the present study are discussed in comparison to existing research.

Regarding the effects of student background variables on selectivity of postsecondary institution attended, annual family income and number of siblings are significant in the full ordered logistic regression model. Hearn (1984; 1991) and Davies and Guppy (1997) found that in full models with both ascriptive and academic characteristics, family income (or a combination variable of income and parental education) is significantly related to attendance at selective institutions. Hilmer (2001), using ordered probit regression, found a positive and relatively large impact of family income on selectivity of institution attended. Hilmer (2001) did not examine number of siblings, but this variable had a significant but small impact in Hearn (1984) and Davies and Guppy (1997).

The results of the current study are largely in agreement with these previous findings: income and number of siblings are statistically significantly related to admissions policy regardless of which variables are in the model. Hilmer (2001) and

Hearn (1991) used an ordinal variable with seven discrete categories to represent family income level; thus, a one-unit change has a different impact on that scale than it would on the scale in the current study. When standardized beta coefficients are examined, however, the impact of income in the present study is similar to that in existing research.

Davies and Guppy (1997) found that student race was not significantly related to attendance at selective institutions, even when academic variables are controlled for. However, Hearn's 1991 work exposed a different pattern. Using average institution SAT score as the dependent variable, Hearn (1991) found that being Black or Hispanic had a significant negative effect on attendance at selective institutions when only student background characteristics were modeled. However, once academic characteristics were taken into account, this effect for Hispanic students disappeared, but Black students still attended institutions of lower selectivity than White students. Hearn's (1991) finding regarding Hispanic students is similar to that in the present study, but the finding about Black students is in contrast to the present study in which being Black has a significant positive relationship with selectivity of institution only once academic variables are included ($e^B = 2.16$ in Equation O2 and $e^B = 1.78$ in Equation O3). However, the present finding mirrors that of Hilmer's (2001) in which an ordered probit regression discovered a significant positive influence of being Black in a model with ascriptive and academic variables. In the current study, the relationship between being of Hispanic ethnicity and attendance at selective institutions changes as more academic variables are added: the significant negative effect becomes significant and positive once continuous academic variables are included, but in the full model, there is no effect of being Hispanic on selectivity of institution attended. Hilmer (2001) also obtained a non-significant relationship between being Hispanic and selectivity of institution attended in a similar full model.

In accordance with previous research (Davies & Guppy, 1997; Hearn, 1991; Hilmer, 2001; McDonough & Antonio, 1996; McDonough et al., 1997), the current study concludes that the most influential factors for determining selectivity of postsecondary institution attended are academic characteristics – specifically, ACT® score and grade point average. Taking a college preparatory track and/or more than three years of foreign language are also significantly related to admissions policy, as they are in previous research. The novel variable representing whether a student has taken any advanced courses was not significant in the ordered logistic regression model.

Perna and Titus (2004, 2005), studying a categorical dependent variable of postsecondary institution type, found increasing significant effects as the highest level of math course taken increased from Algebra I with Geometry to Algebra II to an advanced course beyond Algebra II. Similarly, the current study found the math course-taking patterns of any combination of four or more years or the specific combination of Algebra I, Algebra II, Geometry, Trigonometry, and Calculus/ another advanced math course to be significantly positively related to attendance at selective institution and also more influential than ACT® test score. The approximately continuous total extracurricular participation variable was statistically significantly positively related to admissions policy in the current study, a finding which contradicts the previous study which defined extracurricular participation in the same manner (Hilmer, 2001; the extracurricular variable was significantly negatively related to selectivity).

The current study adds to the existing methodological literature on student college choice, specifically on selectivity of postsecondary institution attended. By beginning with a population of test-takers who were not already pre-disposed to attend college, a more representative sample is attempted. Furthermore, more insight into the decision process is gained by studying students who both attend and do not attend college rather than studying only college matriculants. The use of ordered logistic regression captures the ordered nature of the dependent variable, institutional admissions policy, and the findings largely agree with those of a previous study which matched a similar ordered dependent variable with an ordered regression method (Hilmer, 2001). Using an ordered categorical dependent variable, rather than a continuous SAT score variable (Davies & Guppy, 1997; Hearn, 1991), results in a positive relationship between being Black and attending institutions of higher selectivity when controlling for academic characteristics.

As the economic and social benefits of attending a highly selective college are great (Ehrenberg, 2003; Hearn, 1991), it is to a high school student's benefit to have accurate information on the factors that influence attendance at such institutions. Unfortunately, the current study cannot provide such conclusions due to the large amount of missing data which resulted in a non-random sample. There is some indication, though, that at least for the 5,200 students in Illinois and Colorado with complete data, ACT® score exerts the largest influence on selectivity of postsecondary institution attended. Also influential are several course-taking patterns. Students' chances of attending institutions of higher selectivity may be increased by enrolling in a college preparatory track, taking three or more years of foreign language, taking at least four years of math, and earning high grades in these and all other courses. Participating in extracurricular activities may contribute to attendance at selective institutions and is recommended.

The present study informs the methodology of future studies of student college choice by comparing results of three possible regression methods. Researchers that have complete data, and thus a more random sampling of students, can use the methodological implications from this study to examine influences on college choice and draw practical conclusions for students. The current dataset offers the opportunity to explore many patterns of student college choice. The ACT® Student Profile Section, combined with the data obtained from high school transcripts and from the Institutional Data Questionnaire, provides a wealth of information on student high school experiences. For example, math course-taking patterns have not been examined for their influence on selectivity of institution attended. As noted, the novel variable representing whether a student took any advanced or honors courses was not significant in the ordered logistic model. Perhaps an approximately-continuous variable representing a count of all the advanced or honors courses a student has taken, similar to the total extracurricular participation variable, is a candidate for inclusion in future studies. It could also be the case that the advanced courses variable is closely related to the college preparatory track variable, making multicollinearity a problem. Future regressions might omit one of these variables.

Using ordered logistic regression, impacts of specific extracurricular activities and course patterns (re-defining the advanced courses, math course-taking, and/or foreign language variables; adding a science or social science variable; including business or vocational rather than college preparatory track) can be examined in a model which includes ascriptive characteristics and the continuous academic predictors (ACT® score and GPA).

Limitations

Although the initial dataset represented the population of 183,275 high school seniors in Illinois and Colorado in 2005, due to the large presence of missing data, a non-random sample was used for the present study. The sample of 5,200 students was taken after the listwise deletion of missing data; therefore, this sample is not truly random, making discussion of the results of hypothesis tests of statistical significance of the predictors inapplicable. Additionally, inferences regarding effects of predictor variables on selectivity of postsecondary institution attended cannot be drawn because the results are particular to the specific sample used in this study.

Generalizability is of course affected by the non-random sample, but the deletion of cases with missing data is common in the student choice literature (Alexander et al., 1987; Hearn, 1988, 1991; Hilmer, 2001), and missing data are a problem whenever self-reports (surveys, questionnaires, etc.) are used. Where the current study excels is in the definition of the initial population as including all high school students, not just students who elect to take a college entrance exam or who ultimately matriculate to a postsecondary institution. However, due to problems with missing data, the population to which the results of this study are generalizable can be described as high school students from Midwestern states who choose to respond to a questionnaire about background and academic characteristics and whose high schools provide ACT®, Inc., with transcripts. A description of this sub-set of students compared to students who provided complete data is given in Chapter 3.

The assumption of independence of predictor variables and the error term requires that unobserved variables are not responsible for selection into the sample, are not correlated with included variables, and should not be causes of the outcome variable. Random sampling of the entire population of interest is desired for this assumption to be met. My data have addressed the concern that an omitted variable might be responsible for selection into the sample on the basis of taking a college entrance exam by sampling from states in which the ACT® test is required. However, the primary issue with selectivity bias that may lead to endogeneity in the current study is the presence of a large amount of missing data. It is not known whether an unobserved variable may influence both selection into my sample on the basis of complete data and the dependent variable of selectivity of postsecondary institution attended, and this can be a problem for both classical linear and logistic regressions.

Although the inclusion of both college attenders and non-attenders in the current study addresses the problem of sample selection and has been done in previous research (Hilmer, 2001; Perna & Titus, 2004, 2005), it is likely that different factors influence the decision process for attending or not attending college than for attending various selectivity levels of postsecondary institution. In fact, Hilmer (2001) found that a bivariate probit with sample selection may be appropriate for modeling the two-stage student college choice decision process of first whether to attend a postsecondary institution and then to which type of institution to matriculate. Future studies might examine such a sample selection model with the present data for any differential methodological or student college choice findings.

References

ACT®, Inc. (2009). *Understand your score*. Retrieved May 22, 2009, from http://www.actstudent.org/scores/understand/index.html#multchoice

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley & Sons, Inc.

Akerhielm, K., Berger, J., Hooker, M., & Wise, D. (1998). *Factors related to college enrollment: Final report*. Princeton, NJ: Mathtech, Inc. (ERIC Document Reproduction Service No. ED421053)

Alexander, K. L., Holupka, S., & Pallas, A. M. (1987). Social background and academic determinants of two-year versus four-year college attendance: Evidence from two cohorts a decade apart. *American Journal of Education, 96*(1), 56-80.

Allen, J., & Le, H. (2008). An additional measure of overall effect size for logistic regression models. *Journal of Educational and Behavioral Statistics, 33*(4), 416-441.

Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin, 58*(4), 305-316.

Arnold, K. D. (2002). Getting to the top: What role do elite colleges play? *About Campus, 7*(5), 4-12.

Becker, W. E., Jr. (1983). Economic education research: Part III, statistical estimation methods. *Research in Economic Education, 14*(3), 4-15.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care, 9*(1), 112-118. Retrieved July 8, 2009, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1065119#id778952

Blalock, H. M., Jr. (1979). *Social statistics*. New York: Mc-Graw-Hill.

Borgatta, E. F., & Bohrnstedt, G. W. (1980). Level of measurement once over again. *Sociological Methods and Research, 9*(2), 147-160.

Brant, R. (2004). *Logistic regression: Inference.* Retrieved July 8, 2009, from http://www.stat.ubc.ca/~rollin/teach/643w04/lec/node54.html

Brewer, D. J., Eide, E. R., & Ehrenberg, R. G. (1999). Does it pay to attend an elite private college? Cross-cohort evidence on the effects of college type on earnings. *The Journal of Human Resources, 34*(1), 104-123.

Chapman, D. W. (1981). A model of student college choice. *Journal of Higher Education, 52*(5), 490-505.

Christensen, S., Melder, J., & Weisbrod, B. A. (1975). Factors affecting college attendance. *The Journal of Human Resources, 10*(2), 174-188.

Clarke, M. (2007). The impact of higher education rankings on student access, choice, and opportunity. *Higher Education in Europe, 32*(1), 59-70.

Council for Aid to Education. (1997). *Breaking the social contract: The fiscal crisis in higher education.* Santa Monica, CA: Rand Corporation, Institution on Education and Training. (ERIC Document Reproduction Service No. ED414806)

Dale, S. B., & Krueger, A. B. (2002). Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics, 117*(4), 1491-1527.

Davies, S., & Guppy, N. (1997). Fields of study, college selectivity, and student inequalities in higher education. *Social Forces, 75*(4), 1417-1438.

DesJardins, S. L., Dundar, H., & Hendel, D. D. (1999). Modeling the college application decision process in a land-grant university. *Economics of Education Review, 18*, 117-132.

Dey, E. L., & Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education, 34*(5), 569-581.

Digest of Education Statistics. (2006). *Table 189. Enrollment rates of 18- to 24- year olds in degree-granting institutions by sex and race/ethnicity: 1967 through 2005*. Washington, DC: National Center for Education Statistics. Retrieved May 15, 2009, from http://nces.ed.gov/programs/digest/d06/tables/dt06_189.asp

Do, C. (2004). The effects of local colleges on the quality of college attended. *Economics of Education Review, 23,* 249-257.

Dunn, L. F. (1993). Category versus continuous survey responses in economic modeling: Monte Carlo and empirical evidence. *The Review of Economics and Statistics, 75*(1), 188-193.

Ehrenberg, R. G. (2003). *Econometric studies of higher education.* Ithaca, NY: Cornell University, Cornell Higher Education Research Institute. (ERIC Document Reproduction Service No. ED482390)

Fitzgerald, R. A. (2000). *College quality and the earnings of recent college graduates* (Research and Development Report, NCES-2000-043). Washington, DC: National Center for Education Statistics. (ERIC Document Reproduction Service No. ED445625)

Frone, M. R. (1997). *Regression models for discrete and limited dependent variables*

(1997 Research Methods Forum No. 2). Retrieved July 7, 2009, from http://division.aomonline.org/rm/1997_forum_regression_models.html

Fuller, W. C., Manski, C. F., & Wise, D. A. (1982). New evidence of the economic determinants of postsecondary schooling choices. *The Journal of Human Resources, 17*(4), 477-498.

Gaito, J. (1960). Scale classification and statistics. *Psychological Review, 67*(4), 277-278.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*(3), 564-567.

Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research, 45*(1), 43-57.

Garson, G. D. (2010). Measures of association. In *Statnotes: Topics in multivariate analysis.* Retrieved October 19, 2011, from http://faculty.chass.ncsu.edu/garson/PA765/association.htm

Garson, G. D. (2011a). Ordinal regression. In *Statnotes: Topics in multivariate analysis*. Retrieved October 19, 2011, from http://faculty.chass.ncsu.edu/garson /PA765/ordinalreg.htm

Garson, G. D. (2011b). Testing of assumptions. In *Statnotes: Topics in multivariate analysis.* Retrieved January 28, 2012, from http://faculty.chass.ncsu.edu/garson/PA765/assumpt.htm

Garson, G. D. (2012a). Logistic regression. In *Statnotes: Topics in multivariate analysis*. Retrieved February 16, 2012, from http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm

Garson, G. D. (2012b). Multiple regression. In *Statnotes: Topics in multivariate analysis*. Retrieved February 16, 2012, from http://faculty.chass.ncsu.edu/garson/PA765/regress.htm

Greene, W. H. (2000). *Econometric analysis* (4th ed). Upper Saddle River, NJ: Prentice-Hall.

Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). Burr Ridge, IL: McGraw Hill.

Hand, C. A., & Prather, J. E. (1987, April). *Predicting institutional choice: Patterns of enrollment in the higher education student market*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*(1), 105-131.

Hearn, J. C. (1984). The relative roles of academic, ascribed, and socioeconomic characteristics in college destinations. *Sociology of Education, 57*(1), 22-30.

Hearn, J. C. (1988). Attendance at higher-cost colleges: Ascribed, socioeconomic, and academic influences on student enrollment patterns. *Economics of Education Review, 7*(1), 65-76.

Hearn, J. C. (1991, July). Academic and nonacademic influences on the college destinations of 1980 high school graduates. *Sociology of Education, 64*, 158-171.

Heller, D. E. (2003). *Review of NCES research on financial aid and college participation*. Washington, DC: U.S. Department of Education, Advisory Committee on Student Financial Assistance. Retrieved February 4, 2009, from http://www.ed.gov/about/ bdscomm/list/acsfa/Heller_BeckerFinalReports.pdf

Hilmer, M. J. (2001). A comparison of alternate specifications of the college attendance equation with an extension to two-stage selectivity-correction models. *Economics of Education Review, 20*, 263-278.

Hossler, D., Braxton, J., & Coopersmith, G. (1989). Understanding student college choice. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 5, pp. 231-288). New York: Agathon Press.

Hossler, D., & Gallagher, K. S. (1987). Studying student college choice: A three-phase model and the implications for policymakers. *College and University, 62*(3), 207-221.

Hossler, D., & Stage, F. K. (1992). Family and high school experience influences on the postsecondary educational plans of ninth-grade students. *American Educational Research Journal, 29*(2), 425-451.

Hu, S., & Hossler, D. (1998, November). *The linkage of student price sensitivity with preferences to postsecondary institutions*. Paper presented at the annual meeting of the Association for the Study of Higher Education, Miami, FL.

James, E., Alsalam, N., Conaty, J. C., & To, D. (1989). College quality and future earnings: Where should you send your child to college? *The American Economic Review, 79*(2), 247-252.

Kaufman, J., & Gabler, J. (2004). Cultural capital and the extracurricular activities of girls and boys in the college attainment process. *Poetics, 32*, 145-168.

Kennedy, P. (1998). *A guide to econometrics* (4th ed.). Cambridge, MA: The MIT Press.

Kinzie, J., Palmer, M., Hayek, J., Hossler, D., Jacob, S. A., & Cummings, H. (2004). *Fifty years of college choice: Social, political, and institutional influences on the decision-making process* (Lumina New Agenda Series, Vol. 5, No. 3).

Indianapolis, IN: Lumina Foundation for Education. (ERIC Document Reproduction Service No. ED484237)

Kirk, R. E. (1999). *Statistics: An introduction* (4th ed). Ft. Worth, TX: Harcourt Brace.

Kodde, D. A., & Ritzen, J. M. M. (1988). Direct and indirect effects of parental education level on the demand for higher education. *The Journal of Human Resources, 23*(3), 356-371.

Kotler, P. (1981). Strategic planning for higher education. *Journal of Higher Education, 52*(5), 470-489.

Kotler, P., & Fox, K. F. A. (1985). *Strategic marketing for educational institutions*. Englewood Cliffs, NJ: Prentice Hall.

Kreiberg, D. (n.d.). *Diagnostic testing in Eviews*. Retrieved July 7, 2009, from http://www.bi.no/BibliotekFiles/_nedlastingsfiler/eviews/Diagnostic%20testing.pdf

Labovitz, S. (1967). Some observations on measurement and statistics. *Social Forces, 46*(2), 151-160.

Lane, D. (2003). *Variables.* Retrieved July 7, 2009, from http://cnx.org/content/m10802/latest/

Litten, L. H. (1982). Different strokes in the applicant pool: Some refinements in a model of student college choice. *Journal of Higher Education, 53*(4), 383-402.

Liu, X. (2007, April). *Ordinal regression analysis: Fitting the proportional odds model using Stata, SAS and SPSS*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED501264)

Logistic regression. (2002). San Francisco: San Francisco State University. Retrieved March 15, 2011, from http://userwww.sfsu.edu/~efc/classes/biol710/logistic/ logisticreg.htm

Long, J. S. (1997). *Regression models for categorical and limited dependent variables* (Advanced Quantitative Techniques in the Social Sciences Vol. 7). Thousand Oaks, CA: Sage Publications.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8*, 750-751.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge, England: Cambridge University Press.

Mahoney, J. L., Cairns, B. D., & Farmer, T. W. (2003). Promoting interpersonal competence and educational success through extracurricular activity participation. *Journal of Educational Psychology, 95*(2), 409-418.

Marcus-Roberts, H. M., & Roberts, F. S. (1987). Meaningless statistics. *Journal of Educational Statistics, 12*(4), 383-394.

Marsh, H. W. (1992). Extracurricular activities: Beneficial extension of the traditional curriculum or a subversion of academic goals? *Journal of Educational Psychology, 84*(4), 553-562.

McDonough, P. M., & Antonio, A. L. (1996, April). *Ethnic and racial differences in selectivity of college choice.* Paper presented at the annual meeting of the American Educational Research Association, New York.

McDonough, P. M., Antonio, A. L., & Horvat, E. M. (1997, October-November). *College* *choice as capital conversion and investment: A new model*. Paper presented at the annual meeting of the Association for the Study of Higher Education, Memphis, TN.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology, 4*, 103-120.

McPherson, M. S., & Schapiro, M. O. (1994). *College choice and family income: Changes over time in the higher education destinations of students from different income backgrounds*. (ERIC Document Reproduction Service No. ED380024)

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*(1), 17-24.

Michell, J. (2008a). Is psychometrics pathological science? *Measurement, 6*(1/2), 7-24.

Michell, J. (2008b). Rejoinder. *Measurement, 6*(1/2)*,* 125-133.

Monks, J. (2000). The returns to individual and college characteristics: Evidence from the National Longitudinal Survey of Youth. *Economics of Education Review, 19*, 279-289.

Mueller, R. O. (1988). The impact of college selectivity on income for men and women. *Research in Higher Education, 29*(2), 175-191.

National Student Clearinghouse. (2001). *About the National Student Clearinghouse*. Herndon, VA: Author. Retrieved November 15, 2011, from http://www.studentclearinghouse.org/about/pdfs/Clearinghouse_profile.pdf

O'Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. Y. J. (2008). Multilevel logistic models for dichotomous and ordinal data. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 199-242). Charlotte, NC: Information Age Publishing.

Owings, J., Madigan, T., & Daniel, B. (1998, November). *Who goes to America's highly ranked "national" universities?* (Statistics in Brief, NCES 98-095). Washington, DC: National Center for Education Statistics. Retrieved July 8, 2008, from http://nces.ed.gov/pubs98/98095.pdf

Owings, J., McMillen, M., Burkett, J., & Daniel, B. (1995). *Making the cut: Who meets highly selective college entrance criteria?* (Statistics in Brief, NCES 95-732). Washington, DC: National Center for Education Statistics. Retrieved February 11, 2009, from http://nces.ed.gov/pubs95/95732.pdf

Paulsen, M. B. (1990). *College choice: Understanding student enrollment behavior* (ASHE-ERIC Higher Education Report No. 6). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement. (ERIC Document Reproduction Service No. ED333855)

Peng, C. Y. J., Harwell, M., Liou, S. M., & Ehman, L. (2006). Advances in missing data methods and implications for educational research. In S. Sawilowsky (Ed.), *Real data analysis* (pp. 31-78). Greenwich, CT: Information Age Publishing, Inc.

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3-14.

Peng, C. Y. J., & Naegle Nichols, R. (2003). Using multinomial logistic models to predict adolescent behavioral risk. *Journal of Modern Applied Statistical Methods, 2*(1), 177-188.

Peng, C. Y. J., & So, T. S. H. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics, 1*(1), 31-70.

Peng, C. Y. J., So, T. S. H., Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Research in Higher Education, 43*(3), 259-293.

Perna, L. W. (2000). Differences in the decision to attend college among African Americans, Hispanics, and Whites. *The Journal of Higher Education, 71*(2), 117-141.

Perna, L. W., & Titus, M. A. (2004). Understanding differences in the choice of college attended: The role of state public policies. *The Review of Higher Education, 27*(4), 501-525.

Perna, L. W., & Titus, M. A. (2005). The relationship between parental involvement as social capital and college enrollment: An examination of racial/ethnic group differences. *The Journal of Higher Education, 76*(5), 485-518.

Porter, S. R. (1999, May-June). *Viewing one-year retention as a continuum: The use of dichotomous logistic regression, ordered logit and multinomial logit*. Paper presented at the annual meeting of the Association of Institutional Research, Seattle, WA.

Rumberger, R. W., & Thomas, S. L. (1993). The economic returns to college major, quality and performance: A multilevel analysis of recent graduates. *Economics of Education Review, 12*(1), 1-19.

Scurry, J. E. (2003). *Access and achievement building block: Making the case for all to achieve.* Providence, RI: Brown University, A. Alfred Taubman Center for Public Policy and American Institutions. (ERIC Document Reproduction Service No. ED478809)

Senders, V. L. (1958). *Measurement and statistics*. New York: Oxford University Press.

Sharp, V. F. (1979). *Statistics for the social sciences*. Boston: Little, Brown.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Solmon, L. C. (1972). *The definition and impact of college quality*. New York: National Bureau of Economic Research, Inc. (ERIC Document Reproduction Service No. ED065080)

Southerland, J. N. (2006, November). *Formulating a new model of college choice and persistence.* Paper presented at the annual meeting of the Association for the Study of Higher Education, Anaheim, CA.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.

Stoops, N. (2004). *Educational attainment in the United States: 2003* (Current Population Reports No. P20-550). Washington, DC: U.S. Department of Commerce, U.S. Census Bureau. Retrieved January 26, 2009, from http://www.census.gov/prod/2004pubs/p20-550.pdf

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.

Teachman, J. D. (1987). Family background, educational resources, and educational attainment. *American Sociological Review, 52*(4), 548-557.

Thomas, S. L. (2003). Longer-term economic effects of college selectivity and control. *Research in Higher Education, 44*(3), 263-299.

Thomas, S. L., & Zhang, L. (2005). Post-baccalaureate wage growth within four years of graduation: The effects of college quality and college major. *Research in Higher Education, 46*(4), 437-459.

Toutkousian, R. K. (2001). Do parental income and educational attainment affect the initial choices of New Hampshire's college-bound students? *Economics of Education Review, 20*(3), 245-262.

Trusheim, D., & Crouse, J. (1981). Effects of college prestige on men's occupational status and income. *Research in Higher Education, 14*, 283–304.

UCLA: Academic Technology Services, Statistical Consulting Group. (n.d.). *FAQ: What are pseudo R-squareds?* Retrieved May 22, 2009, from http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47*(1), 65-72.

Venti, S. F., & Wise, D. A. (1982). Test scores, educational opportunities, and individual choice. *Journal of Public Economics, 18*, 35-63.

Wetzel, J., O'Toole, D., & Peterson, S. (1998). An analysis of student enrollment demand. *Economics of Education Review, 17*(1), 47-54.

Winship, C., & Mare, R. D. (1984, August). Regression models with ordinal variables. *American Sociological Review, 49*, 512-525.

Wolniak, G. C., & Engberg, M. E. (2007). The effects of high school feeder networks on college enrollment. *The Review of Higher Education, 31*(1), 27-53.

Zumbo, B. D., & Ochieng, C. O. (2002, April). *The effects of various configurations of Likert, ordered categorical, or rating scale data on the ordinal logistic regression pseudo R-squared measure of fit: The case of the cumulative logit model*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology, 34*(4), 390-400.

Kelly E. Rapp

4590 S. Quail Creek Ave.
Springfield, MO 65810
(417) 890-0492
ktrapp@indiana.edu

EDUCATION

2005 - 2012	Ph.D., Educational Psychology (specializing in Inquiry Methodology) Indiana University, Bloomington, IN Minor: Education Policy Studies Dissertation: *Regression Methods for Categorical Dependent Variables: Effects on a Model of Student College Choice* Advisors: Dr. Ginette Delandshere and Dr. Jonathan Plucker
2001 - 2003	M.S. in Education, Educational Psychology and Research University of Kansas, Lawrence, KS Thesis: *Perceptions of the Scholarship Awarding Process: A Study of High School Counselors and University Scholarship Personnel* Advisor: Dr. Nona Tollefson (Deceased)
1995 - 1999	B.A. in Psychology, with Highest Distinction and with Honors University of Kansas, Lawrence, KS

PROFESSIONAL EXPERIENCE

Teaching

2011-present — **Per-Course Faculty**, Psychological Statistical Methods. Missouri State University, Springfield, MO.

- Acquaint undergraduate students with principles and methods of statistics in the context of scientific behavioral research.

2007-present — **Adjunct Instructor**, Educational Psychology. Ozarks Technical Community College, Springfield, MO.

- Teach online and seated classes to sophomores in the Associates of Arts in Teaching program.
- Developed the online course shell that is used by all educational psychology instructors.
- This course explores intelligence, human development, culture, learning, motivation, research, and assessment, each as it relates to education.

Fall 2006 | **Adjunct Instructor**, Introduction to Psychology. Ozarks Technical Community College, Springfield, MO.

- Introduced undergraduate students to topics in psychology including history and systems, research methods, physiology, human growth and development, sensation and perception, learning, memory, emotion, motivation, personality, psychopathology, and social psychology.

Spring 2006 | **Permanent Substitute Teacher**, Hillcrest High School, Springfield, MO.

- Completed instruction for the remainder of the school year for a secondary math teacher in Algebra I, Algebra II, Business Math, and Math Matters.
- Developed and conducted lesson plans and administered and graded homework and exams for freshmen through senior students.
- Completed end-of-year administrative duties for the absent teacher.

Fall 2005 | **Associate Instructor**, Special Education Research: Approaches and Issues. Indiana University, Bloomington, IN

- Instructed undergraduates in the Teaching All Learning special education teacher preparation program.
- This course emphasizes the importance of research on educational practice, especially as it pertains to special education. Introductory research methods are taught as well.
- This course is cross-listed in the Department of Counseling and Educational Psychology and the Department of Curriculum and Instruction.

Fall 2005 | **Associate Instructor**, Lab for Intermediate Statistics Applied to Education. Indiana University, Bloomington, IN

- Instructed graduate students on how to use SPSS, a statistical analysis computer package.
- Worked closely with the professor of the lecture section, Dr. Chao-Ying Joanne Peng, to integrate the lab and lecture components.
- Tutored students individually on statistical concepts.

2002 - 2003 | **Graduate Teaching Assistant**, Lab for Introduction to Statistical Analysis. University of Kansas, Lawrence, KS

- Was the only masters-level GTA in the department selected to teach graduate students.
- Developed the entire curriculum for this graduate-level computer class, including the syllabus, lesson plans and assignments.

- Instructed students on how to use SPSS, a statistical analysis computer package, in two-hour class periods.
- Implemented the use of a class website for posting course notes and assignments and transmitting other information.

2003 **Teaching Assistant**, Computer Applications for Statistical Analyses. University of Kansas, Lawrence, KS

- Chosen by the instructor to assist weekly with this advanced, graduate-level SPSS laboratory and to grade the class assignments.

2002 - 2003 **Independent Statistical Consultant**
Lawrence and Overland Park, KS

- Consulted with Longmire & Company, Inc., a Kansas City marketing company, on the appropriate statistical analyses for their survey data of prospective college students. Instructed the president of the company on statistical procedures and the use of SPSS.
- Advised doctoral candidates on the appropriate statistical analyses for their dissertations.
- Tutored graduate-level introductory statistics students, both individually and in groups.

Research

Summer 2006 **Summer Intern,** ACT Inc., Education and Workforce Research Services Department. Iowa City, IA.

- One of 10 graduate students selected for this summer program in which I completed a research project under the guidance of professional staff mentors. Project title: *High School Quality, Postsecondary Outcomes, and Student Satisfaction.*
- ACT, Inc. is responsible for research and development of testing programs and other educational services.

2003 - 2005 **Research Associate,** Center for Evaluation and Education Policy. Indiana University, Bloomington, IN

- Working directly with the principal investigators, I led data collection, analysis, and reporting efforts for three projects:
    - Statewide Evaluation of Indiana's Charter Schools
    - Evaluation of Simon Youth Foundation's Educational Resource Centers
    - Georgia Charter Schools Evaluation Project
- I significantly contributed to data collection, analysis, and reporting efforts for various projects:
    - Identification of Students Who Are Limited English Proficient as Gifted, Ohio Department of Education

    - Indiana Early Literacy Grant Program
    - Full Day Kindergarten Initiative
    - Michigan Comprehensive School Reform Parent Involvement Evaluation
- Duties included conducting literature reviews, making site visits, interviewing stakeholders, performing qualitative and quantitative data analysis, and supervising graduate students on the research teams.

2003 **Research Assistant**, Department of Psychology and Research in Education. University of Kansas, Lawrence, KS

- Performed the statistical analyses on a national survey of counseling psychologists led by Dr. James Lichtenberg, Professor, and supported by Division 17 of the American Psychological Association (Society for Counseling Psychology).

2002 **Research Assistant,** Institute for Educational Research and Public Service (IERPS). University of Kansas, Lawrence, KS

- Worked closely with Dr. Nona Tollefson on her grant study on Professional Development Schools, including developing the sampling plan and sampling pool for this survey research.
- Created an extensive database of subjects and subsequently entered and analyzed all data.

2001 - 2002 **Research Assistant**, Center for Educational Testing and Evaluation (CETE). University of Kansas, Lawrence, KS

- Working with the director, Dr. John Poggio, I performed assorted duties critical to the administration of the Kansas statewide assessments as well as other assessments.
- Conducted a literature review on English proficiency testing.

Higher Education Administration

2002 - 2003 **Academic Advisor,** Freshman-Sophomore Advising Center**.** University of Kansas, Lawrence, KS

- Counseled freshmen and sophomores individually and in groups, helping them adjust to university life and navigate the academic system.

1999 - 2002 **Admissions Counselor,** Office of Admissions and Scholarships. University of Kansas, Lawrence, KS

- Admissions representative for Johnson and Douglas Counties (Kansas), the University's largest feeder territories, which included implementing new ideas in planning large recruiting events.

- Admissions representative for Dallas, TX, 2000-2001. Planned the recruitment strategy for this major metropolitan area.
- Gave on-campus presentations to groups as large as 100, as well as met with families on an individual basis.

PUBLICATIONS

Refereed Journals

Harris, B., Plucker, J. A., Rapp, K. E., & Martinez, R. S. (2009, Spring). Identifying gifted and talented English Language Learners: A case study. *Journal for the Education of the Gifted*, *32*(3), 368-393.

Harris, B., Rapp, K. E., Martinez, R. S., & Plucker, J. A. (2007). Identifying English Language Learners for gifted and talented programs: Current practices and recommendations for improvement. *Roeper Review, 29*(5), 26-29.

Plucker, J. A., Eckes, S. E., Rapp, K. E., Benton, S. A., & Simmons, A. B. (2007, Fall). Perspectives on the university sponsorship of charter schools: An exploration of one state's decision-making approach. *Charter School Review, 1*(2), 17-30.

Plucker, J. A., Makel, M. C., & Rapp, K. E. (2007). The impact of charter schools on promoting high levels of mathematics achievement. *Journal of School Choice, 1*(4), 63-76.

Rapp, K. E., & Eckes, S. E. (2007). Dispelling the myth of "white flight": An examination of minority enrollment in charter schools. *Educational Policy, 21*(4), 615-661.

Rapp, K. E. (2005, Fall). Merit scholarships: Are high school counselors' perceptions aligned with university practices? *Journal of College Admission, 189*, 14-20.

Distributed Papers

Rapp, K. E., Eckes, S. E., & Plucker, J. A. (2006, Winter). *Cyber charter schools in Indiana: Policy implications of the current statutory language* (Education Policy Briefs, Vol. 4, No. 3). Bloomington, IN: Center for Evaluation and Education Policy.

Spradlin, T., & Rapp, K. (2004, December 10). *Potential charter school issues for the 2005 session.* Memorandum prepared for State Senator Teresa Lubbers. Bloomington, IN: Center for Evaluation and Education Policy.

Thompson (Rapp), K. E. (2002/2003, Winter). Life lessons beyond the curriculum. *The Jayhawk Educator: The University of Kansas School of Education Alumni Magazine*, 17.

Book Chapters

Rapp, K. E. (2008). Public residential high schools for the gifted. In C. M. Callahan & J. A. Plucker (Eds.), *Critical issues and practices in gifted education: What the research says* (pp. 617-628). Austin, TX: Prufrock Press.

Eckes, S., & Rapp, K. (2005). Charter school research: Trends and implications. In E. St. John (Ed.), *Readings on education: Vol. 21. Public policy and equal educational opportunity* (pp.3-36). New York, NY: AMS Press, Inc.

Research Reports

Plucker, J. A., Eckes, S. E., Rapp, K. E., Ravert, R. D., Hansen, J., Trotter, A., & Makel, M. (2005). *Baseline evaluation of Georgia's Charter School Program.* Bloomington, IN: Center for Evaluation and Education Policy.

Plucker, J. A., Martinez, R. S., Harris, B., Rapp, K. (2005). *Identification of students who are Limited English Proficient as gifted.* Bloomington, IN: Center for Evaluation and Education Policy.

Plucker, J. A., Muller, P. A., Rapp, K. E., Jones, M. E., & Ravert, R. D. (2005). *Evaluation of Simon Youth Foundation's Education Resource Centers.* Bloomington, IN: Center for Evaluation and Education Policy.

Plucker, J. A., Eaton, J. J., Rapp, K. E., Lim, W., Nowak, J., Hansen, J. A., & Bartleson, A. (2004). *The effects of full day versus half day kindergarten: Review and analysis of national and Indiana data*. Bloomington, IN: Center for Evaluation and Education Policy.

Plucker, J. A., Simmons, A. B., Eckes, S., Rapp, K. E., Benton, S. A., & Nowak, J. (2004). *University sponsorship of charter schools in Indiana.* Bloomington, IN: Center for Evaluation and Education Policy.

Plucker, J. A., Simmons, A. B., Rapp, K. E., Benton, S. A., Eaton, J. J., Walter, C. N., Bartleson, A. A., & Sotoo, N. (2004). *Indiana's Early Intervention Grant Program site visit report for 2002-2003*. Bloomington, IN: Center for Evaluation and Education Policy.

Rapp, K. E., Plucker, J. A., Simmons, A. B., Sotoo, N., & Pope, A. P. (2004). *Parent involvement and resource leveraging: A case study of three Michigan schools*. Bloomington, IN: Center for Evaluation and Education Policy.

Tollefson, N., Thompson (Rapp), K. E., & Carmen, C. (2002). *A comparison of the classroom instructional practices of teachers in PDS and non-PDS schools*. Lawrence, KS: Institute for Educational Research and Public Service.

CONFERENCE PRESENTATIONS

Gentry, M., Thomas, A. T., Plucker, J. A., Martinez, R. A., Harris, B., Rapp, K., & Owen, S. V. (2006, April). *Underidentification of Minority, Poor, and English Language Learners as Gifted: A Statewide Evaluation Study.* Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, CA.

Rapp, K. E., & Trotter, A. (2005, October). *New Directions in the Charter School Movement and the Need for Alternative Evaluation Approaches*. Roundtable session presented at the joint meeting of the American Evaluation Association (AEA) and Canada Evaluation Society (CES), Toronto, Ontario, Canada.

Harris, B., Plucker, J., Martinez, R., Rapp, K., & Perez, B. (2005, August). *Best Practices in Working with Gifted English Language Learners*. Poster session presented at the annual meeting of the American Psychological Association (APA), Washington, DC.

Eckes, S., & Rapp, K. (2005, April). *Racial integration in U.S. charter schools*. Paper presented at the meeting of the American Education Research Association (AERA), Montreal, Canada.

Plucker, J. A., Eckes, S. E., Rapp, K. E., & Benton, S. A. (2005, April). *Universities as Potential Sponsors of Charter Schools*: *What do Universities Consider When Deciding Whether to Sponsor Charter Schools?* Paper presented at the annual meeting of the American Educational Research Association (AERA), Montreal, Quebec, Canada.

Rapp, K. E. (2005, March). *Merit Scholarships: Are High School Counselors' Perceptions Aligned with University Practices?* Poster session presented at the annual meeting of the American Association for Higher Education (AAHE), Atlanta, GA.

PROFESSIONAL ACTIVITIES

Developer, Educational Psychology (EDU 250) online course shell for use by all EDU 250 instructors. Ozarks Technical Community College, 2010.

Presenter, Indiana Pathways to College Network Access to Success Conference, *Standardized Testing: Pros and Cons*. October 7th, 2005.

Attendee, Scientific Software International (SSI) Training Session on Modern Data Analysis: Hierarchical Data Analysis with HLM 6. Instructed by Steven Raudenbush and Anthony Bryk, September 9th – 11th, 2004.

Recurring guest lecturer on SPSS for Professor Allan Miller's Understanding Research in Education graduate classes (six sessions) and McNair Scholarship students (a Federal TRIO program). University of Kansas, Summer 2002-Summer 2003.

Psychology and Research in Education department chair search committee student representative, by invitation. University of Kansas, 2002–2003.

Psychology and Research in Education student representative to the departmental Meetings. University of Kansas, 2002.

Greater Plains Association of College Admissions Counselors (GPACAC) Student Ambassador Training Workshop Committee Member, 2001.

Greater Plains Association of College Admissions Counselors (GPACAC) New Counselor Workshop Committee Member, 2000.

PROFESSIONAL AFFILIATIONS

Member, American Educational Research Association (AERA), 2005-2009.
Reviewed proposals for Division L-Educational Policy and Politics for the 2007 Annual Meeting.

Reviewed proposals for the Charter School Research and Evaluation Special Interest Group (SIG) for the 2007 Annual Meeting.

Reviewed proposals for the Law and Education Special Interest Group (SIG) for the 2006 Annual Meeting.

Member, American Evaluation Association (AEA), 2005.

Member, Division 15 (Educational Psychology), American Psychological Association (APA), 2006.

HONORS AND AWARDS

Phi Beta Kappa, initiated 1999.

Agnes Wright Strickland Award, 1999. Chancellor's Award given to two University of Kansas graduating seniors annually for demonstrated leadership in manners of University concern.

Mortar Board National Honor Society, Tapping and Initiation Chair, 1998-1999.

Nelson Scholarship, University of Kansas, 1998. Given by the Department of Psychology annually to an outstanding junior major.