DOCUMENT RESUME

ED 402 860 HE 029 787

AUTHOR	Vernon, James R.
TITLE	The Role of Judgment in Admissions.
REPORT NO	RGSD-129
PUB DATE	96
NOTE	122p.; Ph.D. Dissertation, RAND Graduate School.
AVAILABLE FROM	RAND, 1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138; Internet: order@rand.org
PUB TYPE	Dissertations/Theses - Doctoral Dissertations (041)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Academic Achievement; *Admissions Officers; *College Admission; *Doctoral Dissertations; Enrollment Management; Evaluation Methods; Graduate Study; Higher Education
IDENTIFIERS	Rand Graduate School of Policy Studies CA

ABSTRACT

This dissertation explores issues involved in higher education admissions processes. It analyzes admissions and subsequent performance at the RAND Graduate School. Academic literature informs the presentation of statistical relationships among admissions criteria, admissions committee ratings and performance measures. Transformation techniques provide the foundation for a discussion of the merits of alternative ways of thinking about selection, performance measurement, and prediction. While the RAND Graduate School admissions committee implicitly gives great importance to Graduate Record Examination (GRE) scores, different performance measures correlate most strongly with different selection criteria, complicating the establishment of screening rules. The empirical results show the statistical significance of several selection criteria in predicting a variety of measures of student performance and compare the significance of those criteria to the significance of quantified committee-member ratings in predicting the same performance measures. This report shows several techniques for transforming measures of selection criteria, ratings and performance, in order to discuss the appropriateness of relative and absolute measures. (Contains 113 references.) (MAH)

DISSERTATION

RAND

The Role of Judgment in Admissions

James R. Vernon

RAND Graduate School

The RAND Graduate School dissertation series reproduces dissertations that have been approved by the student's dissertation committee.

RAND is a nonprofit institution that helps improve public policy through research and analysis. RAND's publications do not necessarily reflect the opinions or policies of its research sponsors.

Published 1996 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
RAND URL: http://www.rand.org/
To order RAND documents or to obtain additional information, contact Distribution Services: Telephone: (310) 451-7002; Fax: (310) 451-6915; Internet: order@rand.org

DISSERTATION

RAND

The Role of Judgment in Admissions

James R. Vernon

RAND Graduate School

The original version of this study was prepared as a dissertation in July 1996 in partial fulfillment of the requirements of the doctoral degree in public policy analysis at the RAND Graduate School. The faculty committee that supervised and approved the dissertation consisted of James N. Dertouzos (Chair), William H. Rogers, and Stephen P. Klein.

THE ROLE OF JUDGMENT IN ADMISSIONS

James R. Vernon[1]

[1] This report fulfills the RAND Graduate School dissertation requirement for the doctoral program in public policy analysis for the author, a Graduate Fellow on leave of absence.

The Role of Judgment in Admissions
James R. Vernon

This report explores issues involved in higher education admissions processes. It analyzes admissions and subsequent student performance at the RAND Graduate School. The academic literature on this problem informs the presentation of statistical relationships among admissions criteria, admissions committee ratings and performance measures. Transformation techniques provide the foundation for a discussion of the merits of alternative ways of thinking about selection, performance measurement and prediction.

This research supports several conclusions:

Admissions committees confront conflicting objectives. Committee decisions may show inconsistency across applicants, partly because selection combines individual judgments through persuasion and consensus-building. Furthermore, quantified measures of committee members' ratings may not fully reflect their true judgments about the desirability of applicants.

Accepted applicants typically receive more ratings than rejected applicants. Moreover, committee members sometimes discuss applicants with each other and see other members' ratings before making their own judgments. Nevertheless, including the evaluation of an additional judge in average ratings increases the correlation between average ratings and both the admission decision and some performance measures.

While the RGS admissions committee implicitly gives great importance to GRE scores, different performance measures correlate most strongly with different selection criteria. This complicates the establishment of screening rules.

Transforming ratings and grades increases the correlation between them, and it increases the explained variance in admissions decisions. Thus, both admissions and later performance evaluation depend at least partially on a relative rating of students. Such relativism implies that admissions decisions, which depend at least in part on prediction of future performance, affect how that future performance will be measured.

The empirical results show the statistical significance of several selection criteria in predicting a variety of measures of student performance. They also compare the significance of these criteria to the significance of quantified committee-member ratings in predicting the same performance measures. Furthermore, this report shows several techniques for transforming measures of selection criteria, ratings and performance, in order to discuss the appropriateness of relative and absolute measures.

Although this dissertation focuses on data from RGS, it presents a framework for thinking about problems associated with admissions that general audiences may find useful or enlightening.

To Shanna and Monica, for their patience and faith

To Russ and Kitty, for their inspiration

PREFACE

This dissertation resulted from a debate among members of the admissions committee at the RAND Graduate School. Committee members view the process as very time consuming and lack consensus about the appropriate criteria for admission. Some members emphasize Graduate Record Examination (GRE) scores when evaluating each applicant for admission. Others emphasize different attributes, such as previous grades. Committee members not only disagree over which factors to emphasize in selecting applicants, but they also disagree about what those factors mean. Some members also believe that the admissions process operates inefficiently.

The research for this dissertation began as an attempt to find empirical relationships between a few selection criteria, such as GRE scores and previous grades, and admitted applicants' subsequent grades at RGS, with the hope of finding useful screening criteria. This empirical analysis, however, raised as many questions as it answered. This research broadened to consider a multitude of empirical models. These reflect subtleties of the admissions process as well as complexities in performance measurement. In addition, some of these statistical models led to a consideration of philosophical issues in educational selection. The empirical results show the statistical significance of several selection criteria in predicting a variety of measures of student performance. They also compare the significance of these criteria to the significance of quantified committee-member ratings of applicants in predicting the same performance measures. Furthermore, this report evaluates several techniques for transforming measures of selection criteria, ratings and performance, in order to discuss the appropriateness of *relative* and *absolute* measures in the context of admissions committee decision making.

Although the data for this dissertation come only from RGS admissions records and student transcripts, the implications of this research should interest a wide audience. Decisions such as hiring or promotion, for example, resemble the admissions decision insofar as scarce resources force the selection of fewer individuals than the number of applicants for a position. Predicting individual performance usually plays a significant role in such decision processes. While describing the admissions process at RGS, this dissertation presents a framework for thinking about the variety of problems associated with the process.

ACKNOWLEDGMENTS

The Sloan Foundation provided some funding for the data collection phase of this project. I gratefully appreciate their support. I thank Charles Wolf, Jr., Dean of the RAND Graduate School, for encouraging me to undertake this effort and for providing additional funding toward the end of my research. Jim Dertouzos approached me with the initial request for this study and provided useful guidance as the chairperson of my thesis committee. Bill Rogers patiently explained methodology for survival analysis and suggested numerous improvements to my data analysis. In addition, Jim and Bill taught me economics and statistics at RGS. Steve Klein's expertise in this field of research contributed to the framing of my analysis in the proper perspective. I have learned a great deal from my committee. I thank them for their efforts.

Ellen Reinisch Harrison helped me with several programming difficulties and provided an excellent reference on survival analysis. I thank her and promise to return the book someday. Many of my colleagues eagerly contributed suggestions for useful predictors of performance and provocative ideas about the admissions process. I thank them all collectively; they are too numerous to mention separately.

The staff of the RAND library deserve special praise for excellent work. Joan Schlimgen and Roberta Shanman found references I had not even asked for and enhanced my literature search with their creativity. Barbara Neff and Michael Misumi tracked down obscure articles and out-of-print publications with the utmost patience. Not wishing to omit anyone, but without space the thank them individually, I collectively thank everyone in the library who has assisted me in this and other research.

Contents

SUMMARY

This report explores issues involved in the selection process in higher education. The problem requires an understanding of the objectives of selection, which must relate to the objectives of an institution. Questions of what to teach also influence who to select, and vice versa. Considering these questions fosters greater appreciation for the selection task.

This paper considers that task by analyzing the admissions process and subsequent student performance at the RAND Graduate School (RGS). Predicting how well each applicant will perform represents just one of the difficult issues which admissions committees consider. To put this matter in perspective, this paper also discusses other issues in selection (such as predicting later life performance) as well as broader institutional objectives. The literature on selection and performance prediction informs the presentation of statistical relationships among various selection criteria, admissions committee ratings and measures of performance.

This dissertation also examines the analytical methods used to estimate such statistical relationships. Much of the literature on selection and performance prediction reflects a straightforward approach to measurement of selection criteria, ratings and performance. This research highlights transformation techniques that may provide additional insights into the relationships among these measures. Rather than treat such measures as strict absolutes, this work attempts to show the impact of treating them in a relative context. It does so by considering various measures against relevant group averages and standard deviations. These transformation techniques provide the foundation for discussing the merits of alternative ways of thinking about selection, performance measurement and prediction.

This analysis employs a broad array of statistical models covering admissions committee ratings, the admissions decision and a variety of subsequent student performance measures. The integration of these models provides a way to consider specific issues in the more general context of the academic environment. This context enables direct comparison of how specific factors (for example, GRE scores) influence committee ratings against how those same specific factors influence a particular performance measure (for example, attrition). Although previous research

explores many of the specific issues covered in this report, this dissertation provides data from a single institutional environment to consider them all simultaneously.

This research supports several conclusions. These conclusions apply not only to specific concerns at RGS but also to admissions and performance measurement policies at other institutions.

Admissions committee ratings may not fully reflect decision makers' true judgments about the desirability of each applicant. The imperfect fit between overall committee ratings and actual committee admissions decisions at RGS demonstrates this point. This cannot reflect mere uncertainty. If ratings represented decision makers' best assessments of the desirability of each candidate, then actual decisions should follow those ratings (however uncertain) quite closely. However, actual admissions decisions show that the admissions committee makes judgments that do not strictly follow their ratings of individual candidates. Such judgments involve more than mere predictions of performance. This helps explain the imperfect correlations among selection criteria, ratings and performance measures.

GRE Quantitative score has broad significance in predicting RGS outcomes, from ratings to admissions decisions to several performance measures. However, *different performance measures correlate most strongly with different selection criteria and have varying predictability.* Moreover, ratings have predictive value for RGS outcomes even after adjusting for criterion measures. Including the evaluation of an additional judge in average ratings increases the correlation between average ratings and both the admission decision and some performance measures, at least for the first few judges. This reinforces the role of judgment in admissions, independent of objectively measurable selection criteria.

Transformation techniques demonstrate that treating ratings and performance measures as strict absolutes overlooks an important aspect of the issues involved in selection. Transforming ratings increases their ability to explain variance in admissions decisions. The improvement comes from grouping the ratings of each individual admissions committee member. Normalizing each rating against these group distributions transforms them to relative measures. (This normalization uses both a standard normal distribution as well as the distribution of GRE scores in each group as the distribution for the transformed measures.) *In*

addition, transforming grades increases the predictability of GPA. These transformations group the grades of each individual course. Normalizing each grade against these course distributions transforms them to relative measures. (This normalization also uses both a standard normal distribution as well as the distribution of GRE scores in each course as the distribution for the transformed measures.) Thus, both the admissions decision and later performance evaluation depend at least partially on a relative rating of students. Traditional analyses of ratings and performance measures as strict absolutes do not account for this complexity. This analysis seems to reflect a more realistic model of the processes involved.

These considerations complicate the establishment of screening rules. *Institutions can address these issues by clarifying objectives, rationalizing decision processes and enhancing evaluation policies.* Clarifying objectives and linking them to specific performance measures may enhance decision makers' understanding of their task and the limitations of their predictions. Even without strict screening rules, the RGS admissions committee can use the empirical results in this research to help lower its workload. Ratings of candidates and evaluation of students along separate dimensions can at least make uncertainty more explicit, and perhaps lead to more accurate (although narrower) predictions.

I. INTRODUCTION

Spending for higher education in the U.S. surpasses $175 billion annually.[1] The number of graduate students exceeds two million, and the number of undergraduates exceeds 12 million.[2] These students attend over 3,600 separate institutions.[3] How to select these students has received much attention, as debates about affirmative action and truth in testing show.[4] This dissertation illustrates a methodology for evaluating some of the choices implicit in the admissions decision.

Policy Issues in the Selection of Graduate Students

When the number of applicants exceeds its capacity, an institution must decide which individuals will most likely best serve its goals. Although quite common, this problem remains complex. Institutions face a multitude of possible objectives. Some of these objectives may conflict with each other, and the implicit or explicit relative importance of each may vary.

In addition to uncertainty over objectives, decision makers must consider a variety of selection criteria and determine their relative importance. Typically, such selection criteria correspond to institutional objectives.[5] In other words, once decision makers determine an objective, they look for performance measures that suit that objective. They then look for ways to predict that performance measure. Institutions consider many criteria for their usefulness in predicting some aspect(s) of future student behavior, such as academic performance or achievements later in life. In contrast, some schools also use criteria, such as race or gender, for their usefulness in predicting the characteristics of future *groups* of students and their impact on society, as distinct from the prediction of future *individual* behavior.

Regardless of the emphasis on various objectives and criteria, admissions committees usually confront explicitly the prediction of academic performance. This prediction uses limited information and cannot eliminate uncertainty. (Even for simple performance measures such as

---

1 Statistical Abstract of the U.S., U.S. Department of Commerce, Washington, DC, 1994, p. 151.

2 Ibid., p. 152.

3 Ibid., p. 151.

4 See, for example, Ralph Nader, *The Reign of ETS: The Coporation That Makes Up Minds*, Allen Nairn and Associates, Washington, DC, 1980; and Robert Klitgaard, *Choosing Elites*, Basic Books, New York, NY, 1985.

5 This analysis does not consider principal-agent models or other impediments to rational decision making.

attrition, prediction involves uncertainty.) Nevertheless, the importance of performance among institutional objectives makes predicting it (as accurately as possible) a desirable part of the selection process. This report considers the problem of predicting performance in the context of selection objectives, the literature on selection and prediction, and the admissions process at the RAND Graduate School (RGS), a doctoral program in public policy analysis.

Selection criteria vary among institutions and among decision makers within institutions. Analysis of applicant characteristics, ratings and admissions at RGS illustrates a methodology for understanding more clearly the selection process and performance prediction. Placing selection criteria and processes in this context fosters greater appreciation of the difficulty in determining (empirically or rationally) selection objectives and their roots in broader institutional objectives.

The usefulness of these criteria in predicting academic performance gives admissions committees a measure of the uncertainty they face and how much their selection criteria can reduce that uncertainty. Decision makers can use empirical models of performance prediction to develop screening criteria or ranking mechanisms to improve the selection process by making better predictions or reducing the need for decision making resources. One of these resources, the human effort of individual admissions committee members causes some to wonder what they contribute to the selection process. Empirical models of selection and performance prediction shed some light on the role of judgment in admissions committee decisions.

The measurement of performance, selection criteria and subjective ratings influences the formulation of models for both selection and performance prediction. Decision makers need to choose performance measures with an understanding of how their choice affects the predictive value of various selection criteria and of their subjective ratings. Furthermore, combining individual judgments, for both admissions committee decisions and composite performance measures, raises the question of whether decision makers have consistent standards of judgment. To address some of these measurement issues, this analysis employs a variety of transformations for both subjective ratings and performance measures. Normalizing both ratings and performance measures against group distributions transforms them to relative measures. This analysis considers relative measures within groups such as candidates rated by an individual admissions committee member as well as students evaluated by an individual instructor. These

transformations enable a comparison of *absolute* and *relative* ratings and performance measures. In other words, the statistical models used in this analysis can account somewhat for possibly divergent standards of measurement held by individual decision makers for both admissions and performance measurement.

This comparison of absolute and relative measurement suggests a conceptual approach to the selection process that accounts for the relationships among institutional objectives, selection criteria, subjective ratings, admissions decisions, student performance and faculty evaluations. By exploring the broad context for these issues, this dissertation creates a framework for thinking about selection and prediction, illustrating methodologies for analyzing these problems. These methodologies can help decision makers better understand the nature of the problems they face and allow them to make more informed choices among objectives, selection criteria, screening and ranking mechanisms, performance measures and resource uses.

Organization of This Report

Section II considers institutional objectives, which form the background for a discussion of selection criteria. Section III describes the RGS admissions process and relevant aspects of other selection processes. Section IV reviews the academic literature on selection and performance prediction to explore concepts and methodologies and to frame expectations for the empirical analysis in the present study. Section V describes the data used in that empirical analysis. Section VI shows the relationship between criterion measures and committee ratings, after which Section VII presents the results of analyzing admissions decisions, using both criterion measures and ratings as predictors. Sections VIII and IX show the results of predicting various performance measures at RGS. Section X discusses the conclusions supported by this research and suggests a conceptual approach to thinking about selection and prediction.

II. OBJECTIVES OF SELECTION

Admissions decisions reflect value judgments. Even formulaic selections using only objective factors such as test scores require definition of what type of candidate a school wants. Decision makers approach the selection process with ideas and preconceptions that may only vaguely correspond to explicit institutional objectives. They may have biases or prejudices that affect their evaluation of applicants. They may evaluate applicants in an ad hoc fashion, without a clear sense of why they choose certain applicants over others. Despite these uncertainties and the inevitability of value judgments, admissions decisions also reflect, however loosely, some combination of selection criteria, which in turn reflect the broader goals of the institution. This section discusses such issues to give a richer context for the somewhat narrow issue of predicting performance.

Institutional Objectives

Schools can have many objectives. Some have ancient historical antecedents. These broad objectives shape specific selection objectives and the actual criteria it uses to evaluate applicants. Regardless of their appropriateness or relative importance, considering just the following representative objectives underscores the complexities of the selection problem:

- Contribution to knowledge
- Social value added
- Professional training
- Institutional welfare (including consideration of legacies, diversity, endowment funds, etc.)

Contribution to knowledge probably comes closest, among these objectives, to the classical ideal of education. This ideal, embodied in Plato's Academy, defined education as the process of inquiry and rational discourse. Although modern education emphasizes other objectives, the classical ideal still dominates many discussions of the objectives of educational institutions.[6] Unfortunately, this objective provides little guidance for selection, since it presumes that anyone

[6] See, for example, Allan Bloom, The Closing of the American Mind: How Higher Education Has Failed Democracy and Impoverished the Souls of Today's Students, Simon and Schuster, New York, NY, 1987; Robert Paul Wolff, The Ideal of the University, Beacon Press, Boston, MA, 1969; Mortimer Adler, Ten Philosophical Mistakes, Macmillan Publishing Company, New York, NY, 1985; and Steven Cahn (editor), Classics of Western Philosophy, Hackett Publishing Company, Indianapolis, IN, 1977.

can make a positive contribution to knowledge. The ancient Greeks did not seem to worry about limited resources in academic selection,[7] perhaps because wealth and privilege defined a small enough class of *applicants* to dispense with further resource allocation decisions. Nevertheless, modern scholars and decision makers do have limited resources. These concerns lead to institutional objectives that consider other factors besides contribution to knowledge.

With the need to select fewer students than the number of applicants, some admissions decision makers adopt the objective of maximizing the social value added by the applicants. This includes all the contributions made in individuals' later lives.[8] Even with perfect prescience, social value added has several possible implications for specific selection objectives and criteria.

One interpretation of social value added considers the contribution that each individual would make if admitted to the institution. Under this notion, an institution would maximize social value added by accepting those applicants with the greatest *expected contributions*. However, this interpretation ignores the contributions that each individual would make if rejected. Interpreting social value added to mean the marginal increase in the social contribution of each individual if admitted leads to a different selection objective: accepting those applicants with the greatest *expected increase* in their social contribution. In addition, both total and marginal social contribution also invite consideration of opportunity costs. In other words, an institution might consider whether a prospective applicant's social value added would increase by attending some other institution, or none at all.

Other considerations complicate social value added still further. First, an institution may define social value in a variety of ways. One definition rests on the classical conception of education, equating value with knowledge. Even with this reformulated objective, however, institutions would confront the problem of defining knowledge and judging the relative value of different kinds and amounts of knowledge in order to clarify their selection objectives. Broader definitions of social value, including economic or psychological well-being, moral virtue or other normative factors, make the connection between social value added and selection objectives and criteria more difficult to perceive. Admissions committees may not articulate or even understand

---

7 Cahn (editor), op. cit.

8 Robert Klitgaard, op. cit.

this connection very clearly. But explicit reference to social value added and other objectives can help them understand the sources of their selection objectives and criteria.

A second complexity of social value added results from differences between decision makers. Even if individual decision makers could articulate their institutional objectives and their influence on selection objectives and criteria, admissions committees would have to forge these differing objectives into a unified objective function for the institution in order to specify a set of selection objectives and criteria that serve the institutional objectives rationally. In practice, individual committee members maintain different institutional objectives, selection objectives and criteria. Committee decisions represent compromises among various members, and the points of compromise may vary with each applicant the committee considers. Such compromises occur not only between different conceptions of social value added, but also between different conceptions of other institutional and selection objectives.

Two hypotheses about educational objectives illustrate further the complexity of the selection problem and the difficulty in justifying particular selection objectives and criteria. The human capital hypothesis, a variant of the social value added concept, views education as an investment in productive capacity (human capital).[9] The screening hypothesis, developed as a critique of the human capital approach, posits a different institutional objective: identifying those individuals who already possess the productive capacity sought by employers.[10] Each hypothesis has implications for selection objectives.

Theories of human capital and educational screening have focused attention on the relationship between education and economic success. These theories implicitly assume that personal attributes and achievements can predict or explain some of the variation in future academic or later life performance. Each hypothesis prescribes a different role for education (institutional objectives), but both suggest that selection criteria should predict future behavior.

Theories of human capital assert that differential capacity among individuals manifests itself in differential economic or market value. According to this view, education increases productive capacity, or human capital, and the benefits of education accrue through higher future

---

9 Elchanon Cohn, The Economics of Education, Ballinger Publishing Company, Cambridge, MA, 1979.)

10 Ivar Berg, Education and Jobs: The Great Training Robbery, Praeger Publishers, New York, NY, 1970.

earnings.[11] Economists generally measure such benefits using the capitalized earnings approach, essentially a net present value calculation of an individual's future income stream, adjusted for future costs (including the cost of education) and expected life span. Proponents of the human capital hypothesis "accept the [Adam] Smithian position that an educated man [or woman] may be compared to an expensive machine."[12] Under a conception of social value that stresses maximum contribution of each individual, the human capital hypothesis would lead to the selection objective of accepting those applicants with the greatest *expected future earnings*. Emphasizing marginal contribution would lead to the selection objective of accepting those applicants with the greatest *expected increase in capitalized earnings*. In either case, the human capital hypothesis presumes that the educational process affects productive capacity.

In contrast, under the screening, or credentialism hypothesis, education merely selects and identifies those individuals who already possess the attributes that contribute to market value, including productive capacity.[13] (The screening hypothesis treats the act of completing school as a key signal to prospective employers. Thus, such employers would not ordinarily select these individuals in advance of their completion of schooling.) Despite this philosophical disagreement with the human capital hypothesis, the educational screening hypothesis also suggests a selection objective of accepting those applicants with the greatest expected future earnings. (Since the screening hypothesis rejects the proposition that education affects productive capacity, a marginal approach to expected future earnings does not fit with this approach.) Although both the human capital and screening hypotheses suggest that predictors of future performance make useful selection criteria, not all interpretations of educational objectives lead to the same comfort with predictors.

While economic theory has generally considered the causal relationship between education and education as uni-directional, some scholars have argued that the causal connection may run the other way. In other words, increasing educational attainment may result from greater economic success. In a review of the development of higher education in the United

---

11 Cohn, op. cit.

12 R. Blandy, "Marshall on Human Capital: A Note", *Journal of Political Economy*, Volume 75, Number 1, pp. 874-875, December, 1975.

13 Berg, op. cit.

States, Edward Purcell notes that "the rapidly accumulating wealth and the new technical demands" of industrial society "helped spur the great age of university expansion."[14] This perspective stresses the importance of early life factors on the decision to pursue education beyond minimum legal requirements. As Eric Hanushek succinctly puts the issue, "family background has a pervasive and powerful impact on student achievement; higher socioeconomic status is systematically related to higher achievement."[15]

The notion of education as an effect of economic success rather than as a cause adds another complication to the relationship among institutional objectives, selection objectives and selection criteria. If predictors of some measures of social value added such as educational attainment depend in part on economic status or other measures correlated with educational attainment, then using such predictors as selection criteria suffers from circular reasoning. This does not mean that such predictors have no value as selection criteria, only that their reflection of selection objectives and foundation in institutional objectives may reinforce existing societal conditions. However, even more narrow conceptions of the objectives of education cannot avoid similar value judgments in their implications for selection objectives and criteria.

Many institutions, particularly graduate schools, have explicit missions to provide professional training. Unlike more general objectives, training often follows well-established curricula, and independent groups may set standards for licensing or certification of practitioners. An institution may define its objectives by referring to these curricula and standards. However, limited resources will still require schools to decide which applicants will best achieve these standards, or to apply additional standards. This introduces the same value judgments implicit in the objectives of contribution to knowledge or social value added. By referring to professional standards, however, institutions may clarify the connection between broader educational objectives and specific selection objectives, such as accepting those applicants with the greatest expected achievement in professional certifications. These selection objectives in turn may clarify the need for selection criteria that serve as useful predictors of such achievements.

---

[14] Edward Purcell, The Crisis of Democratic Theory: Scientific Naturalism and the Problem of Value, The University Press of Kentucky, Lexington, KY, pp. 6-7, 1973.

[15] Eric Hanushek, Achievement and Race: An Analysis of the Educational Production Process, Heath Books, Lexington, MA, 1972.

Faced with the need to make value judgments, many decision makers rely explicitly on the objective of maximizing the institution's well-being, as distinct from any benefits to the larger society around the institution. This institutional objective may manifest itself in the selection objective of accepting those applicants whom faculty members would most like to teach.[16] More generally, this leads to selection objectives of accepting those applicants who have the greatest chance of making the institution *a better place,* however admissions decision makers may choose to define *better*. In addition to obvious value judgments, this also implies that selection criteria should serve as useful predictors of such behavior. However, even if individual decision makers held precise conceptions of the future behavior and characteristics of applicants that would maximize their sense of institutional welfare, the need to combine these into a single objective function for the institution would again complicate the relationship among institutional objectives, selection objectives and specific selection criteria.

In practice, admissions committees have multiple institutional objectives that reflect some combination of the objectives previously discussed, and probably other objectives that reflect both individual and group biases and preconceptions. Many institutional objectives suggest selection objectives and criteria that rely on predictive characteristics. Even a narrow focus on prediction, however, involves complicated, uncertain relationships among selection objectives and criteria, at least partly due to the complexity and uncertainty of institutional objectives.

Selection Objectives and Criteria

The multiple institutional objectives held by individual admissions decision makers and by an admissions committee as a group lead to many possible selection objectives and criteria. While selection objectives often relate to expectations about applicants' behavior and characteristics over the rest of their lives, selection criteria usually focus on predictors of more immediate and familiar objects, such as grades or other performance measures used by the institution itself. Understanding how specific institutional objectives can lead to different selection objectives and criteria provides context for an empirical analysis of selection criteria, admissions committee ratings and performance measures.

---

[16] Alternatively, it might lead to a selection objective of accepting those students with the greatest potential contribution to endowment funds, or it could reflect a social value added consideration such as diversity.

Most institutional objectives reinforce selection objectives that emphasize expectations about academic achievements within an institution. Although admissions committees may consider the long-run future in institutional objectives such as contribution to knowledge or social value added, they often focus on behavior or characteristics that they can observe while applicants remain students within their institution. However, such simplifications still leave plenty of ambiguity in selection objectives and criteria.

A selection objective as narrow as accepting those applicants with the highest expected academic performance can have several very different meanings, for example:

- Choosing applicants in descending order of expected performance
- Choosing applicants to maximize the chance of at least one *great success*
- Choosing applicants to minimize the chance that any individual will fail

At the level of the individual applicant, these different interpretations of the simple objective of accepting applicants for their expected academic performance imply the selection of different individuals. Furthermore, recognition of the effect that individuals have on each other may lead to selection of applicants not just for their own expected academic performance, but for their expected influence on the performance of other applicants. Thus, even narrow interpretations of institutional objectives such as contribution to knowledge or social value added that focus on academic performance can imply many different selection objectives and criteria.

In addition, institutional objectives more broadly interpreted often imply selection objectives and criteria that have little or no connection with individual academic performance. Social value added objectives, for example, may suggest the selection of individuals based on expectations of leadership qualities that may not fully develop until much later in life. Other later life contributions relevant to social value added objectives, equally difficult to measure or predict, might include innovations or inventions, other creative endeavors, service to industry or government, or just generally having the moral character that institutions may seek to foster.

Some institutional objectives also support selection objectives and criteria that consider behavior and characteristics, other than scholastic accomplishments, while an individual remains a student at the institution. Choosing applicants for their non-academic contributions to an institution or to ensure a "happy bottom quartile"[17] obviously serves the objective of institutional

welfare or quality of life. But such selection objectives may also serve some social value added objectives or even the classical ideal of contribution to knowledge, by helping the institution function better as a whole. In other words, a *happy bottom quartile* may also inspire or encourage a *smarter top quartile*. Regardless of what institutional objectives they might serve, such selection objectives and criteria clearly add complexity to the selection problem.

Other factors complicate admissions decisions for reasons that reflect societal concerns, such as fairness. Whether or not institutions formally adopt objectives of ensuring a more equitable distribution of goods among various social groups, possibly including affirmative action to remediate existing or past inequalities, individual admissions decision makers often take into consideration an applicant's social background, including ethnicity and gender.

For a variety of reasons, then, selection objectives and criteria for an academic institution involve much more than predictions of each applicant's future academic performance. Individual admissions decision makers each consider multiple objectives and criteria. The process of combining these individual objectives and criteria into a single set of institutional objectives, selection objectives and selection criteria may make it impossible to understand or ascertain an admissions committee's objective function. As Section VII will show, even an analysis of the relationship between admissions committee ratings and actual decisions leaves some unexplained variance. Furthermore, the relationship between ratings and measurable selection criteria also has considerable unexplained variance. As both committee ratings and actual decisions provide evidence of objectives, those later analyses will enrich the preceding discussion. An appreciation for the complexity of objectives provides useful background for discussing the empirical analysis of both selection processes and relationships among selection criteria, admissions committee ratings and academic performance measures. The next section shows how RGS addresses these complex issues in its admissions process.

---

[17] Robert Klitgaard, op. cit., discusses these selection objectives in more detail.

III. THE SELECTION PROCESS

Other studies have examined the procedural context of admissions decisions, providing details of how admissions committees function.[18] RGS and its admissions process differ in several ways from more traditional schools. Some of the characteristics of RGS and its selection process may contribute to a better understanding of the analysis of RGS data. This section describes RGS as it operated in the 1980s. Although RGS has changed since then, this description has relevance for the analysis presented in later sections. The concluding section (as well as occasional footnotes) discusses changes in the process, some of which reflect the policy implications of this analysis.

In 1970, the RAND Corporation, a non-profit organization dedicated to public policy-oriented research, established a doctoral program in public policy analysis to teach students the application of scientific methods to problems of government. Although the RGS curriculum has evolved, the core curriculum at RGS has always included microeconomics, statistics and data analysis, and other social science courses, including organizational behavior. RGS draws its faculty from the professional research staff of the RAND Corporation.

Some students also take courses for RGS credit at other universities, and RGS has had two joint programs - in health policy studies and Soviet studies - with the University of California at Los Angeles. (Like the Soviet Union, the Soviet studies program no longer operates.) In addition to core courses and other courses teaching theoretical perspectives on subjects relevant to public policy, RGS offers workshops to study the application of these methods to areas of public policy, usually where RAND staff members have special expertise.

RGS requires students to work on RAND projects as on-the-job training. This work supports the fellowship funding that RGS provides its students. Students must also complete 21 quarter-length courses, including required core courses and at least two workshops. Before taking qualifying exams, students must complete 17 of these courses, with no more than two grades of C or lower. Qualifying exams cover the core areas of economics, quantitative analysis,

[18] Robert Klitgaard, op. cit.

and social sciences, as well as integration of these disciplines into policy analysis. (Students must pass all fields.) Approval of a dissertation in policy research by a committee of three members (including at least one member of the RAND Corporation staff, excluding the student's project leader) chosen in consultation with the Dean, completes the RGS Ph.D. requirements.

As in most graduate education programs, RGS admissions follow an annual cycle. Applicants submit, usually before a March 1 deadline, each of the following items:

- A formal letter requesting admission and stating the applicant's purpose for studying at RGS
- Reports of test scores from the Graduate Record Examination, GRE (scores from the Graduate Management Admission Test, GMAT; the Law School Aptitude Test, LSAT; or the Medical College Admissions Test, MCAT, may substitute for the GRE)
- A current resume
- Three letters of recommendation (from teachers, employers or others)
- Copies of previous academic transcripts
- A sample of the applicant's written work

The administrative staff of RGS maintains records of all applicants and distributes information folders for each applicant to various members of the admissions committee.

Unlike many other academic institutions, RGS uses faculty members and occasionally other members of the RAND staff, rather than professional admissions officers for its admissions committee. The Dean of the RAND Graduate School, who has chaired the admissions committee each year[19] since its inception, chooses about four or five committee members[20] to assist in the selection process each year. Because the faculty of RGS changes over time, most admissions committee members serve for just a few years.[21]

RGS administrative staff assign applications to admissions committee members somewhat arbitrarily. At least one member reviews each applicant. However, RGS does not require a specified number of members to review each applicant. After they initially distribute applications to committee members, the staff periodically ask members to return any applications they have finished reviewing, and often request the member to review additional applications.[22]

---

19 The Dean no longer chairs the admissions committee. See Section X.

20 Not all committee members play similar roles on the committee. See Section X.

21 More recently, admissions committee membership seems more stable. See Section X.

22 The current process uses more systematic assignments. See Section X.

Several pressures affect committee members in the selection process. First, they have full-time jobs as members of the RAND staff. Their research may preoccupy them while they review applications. In addition, some committee members teach courses at other schools or work as independent consultants as well. Second, although RGS staff sometimes pressure committee members to review applications quickly, they often penalize the fast reviewers by giving them additional information folders to review.[23] Lastly, committee members may impose pressures on themselves to review as many applications as they feel able to, in order to have more influence in the selection process. Thus, each committee member may review few or many applications, thoroughly or not, depending on the influence of these pressures.

Admissions committee members also have different approaches to reviewing applications. Some prefer to evaluate applications one at a time; others like to review them in batches of three, four or more, returning to information folders to compare points of interest among applicants. Some committee members begin their review by reading an applicant's resume, to form a mental image of the applicant; others begin by looking over an applicant's academic transcript(s). Some members admit to varying their approach with each application.

Just as their approaches to reviewing applications vary, different committee members also tend to emphasize different factors in their evaluations. When pressed, committee members do not describe their implicit weightings of various selection criteria with much certainty. Most acknowledge that their decision processes involve a degree of subjectivity. Although they differ in the relative importance they may attach to each selection criteria, committee members all seem to find relevance in most application requirements:

- The letter of application gives members a chance to evaluate the applicant's motivation, writing style and general objectives in applying to RGS.
- The applicant's resume allows members to consider the schools attended by the applicant, his or her major field of study and previous work experience, as well as other experiences that the applicant deems important.
- Transcripts show members the applicant's grades in specific courses, as well as trends that may indicate special potential.
- Test scores provide members a means of comparing all applicants on common scales for both quantitative and verbal aptitude.

---

[23] Ibid.

- The writing sample lets members evaluate an applicant's ability to communicate in the applicant's field of expertise, and possibly to judge the applicant's achievements.
- Letters of recommendation indicate how those who know the applicant personally evaluate her or him, and although committee members often discount positive recommendations, they often feel that a negative recommendation indicates a serious weakness.
- Committee members also consider possible matches for on-the-job training, as well as potential sources for dissertation funding for each candidate.[24]

Precisely how each admissions committee member combines each of these criteria, possibly with other factors such as ethnicity or gender, to form an evaluation of each candidate, depends on the subtle interplay of the member's interpretation of the institutional objectives of RGS, and remains a mystery. Nevertheless, each committee member summarizes his or her evaluation of an applicant with a letter rating:

- A indicates the most favorable evaluation, a recommendation of definite admission.
- B indicates a recommendation of probable admission.
- C indicates a recommendation of possible admission.
- D indicates a recommendation of probable rejection.
- E indicates a recommendation of definite rejection, the least favorable evaluation.
- Occasionally, committee members will use a rating of Z to indicate some special consideration for discussion by the whole admissions committee or lack of confidence in the member's ability to evaluate that particular applicant.

A few weeks after the application deadline, after at least one member has reviewed each application (usually, several members have reviewed each applicant)[25], the admissions committee meets as a group to discuss the applicants. To organize the committee's discussions, RGS staff gather the information folders and tally the committee members' ratings. They group the applicants according to the most favorable evaluation received by each. This grouping breaks the applicant pool into an A list, a B list, etc. Each list displays applicant names, test scores, and ratings, if any, received from each committee member. Often, these lists will indicate whether an applicant's field of interest will most likely concentrate on national security or domestic policy issues.[26]

---

[24] Despite their obvious importance, however, these factors did not impact any empirical analysis. (Both factors present a relatively difficult problem in determining quantitative measures. The few categorical measures used did not produce any statistically significant results.)

[25] In the current process, two members review each initial applicant. Additional members rate only those passed on by the initial reviewers. See Section X.

[26] See note above regarding the importance of on-the-job training and dissertation funding.

The admissions committee discusses each applicant, one at a time, beginning with the A list. This discussion usually involves a review of the applicant's file, including a summary rating by the whole committee. Without formal voting, the committee tries to reach a consensus decision to accept or reject each applicant. The meeting may last an entire day or even continue for a second day, but the committee usually reaches a tentative decision on the majority of applicants at this meeting. The committee sometimes decides to offer admission to a candidate at this initial meeting, but usually it decides to invite those candidates it considers likely admissions for a day of interviews at RAND. For some remaining applicants, the committee may decide to have more members review the information folder before bringing the applicant to RAND for interviews. (The additional reviewers sometimes recommend rejection. Thus, not all additional reviews result in interviews or admissions offers.) Occasionally, the committee may consult with a RAND staff member with personal knowledge of an applicant or special expertise in the applicant's field of interest.

The admissions committee's decision process becomes less formal after the committee's initial meeting. Interviewing candidates occurs over a period of several weeks, depending on the schedules of both candidates and interviewers. The interview process itself usually takes a full day for each candidate. Typical interviews last from 45 minutes to an hour. Interviewers include admissions committee members, other faculty members and RAND staff members with research interests similar to those of a particular candidate. Informally, interviewers often discuss candidates with each other, as well.

The Dean[27] will solicit feedback from interviewers of each applicant (also chosen by the Dean[28]) as well as from committee members reviewing applications after the first meeting. In addition, the Dean almost always interviews any applicant who visits RAND. Without convening any formal meeting of the admissions committee, the Dean usually makes the final decision to accept or reject each remaining applicant, after seeking a consensus of those who have interviewed an applicant or reviewed the applicant's information folder, or both.

This study of the RGS admissions process reveals several important characteristics:

---

27 See notes above describing changes in the process.

28 Ibid.

- Admissions committee members consider a variety of objectives in making their selection recommendations. They do not explicitly predict academic performance in their summary ratings, and they (obviously) do not specify a particular measure of academic performance for their selection objectives.
- Members use a variety of selection criteria, without clarifying the emphasis they put on each.
- Members usually will have an opportunity to evaluate accepted applicants' academic performance at RGS. This might influence their measurement of such performance, as a potential confirmation or validation of their earlier judgment. (Such measurement could also simply reflect their ability to anticipate and predict performance. To the extent that performance measurement reflects subjective judgment, however, prior evaluation might have some influence.)
- More admissions committee members rate applicants with higher ratings.
- Members know the ratings given to some of the applicants by other members before they form their own ratings.
- The Dean exerts unspecified influence on the selection process, with final decision authority for applicants at the margin of the committee's collective judgment.

Although admissions committee members may have other objectives besides academic performance when selecting applicants, they also clearly believe in the importance of academic performance, and they wonder whether or not they can predict it, and how well they can do so. A review of the literature regarding performance prediction can set some general expectations, as well as provide useful guidance for the empirical analysis of RGS data on selection criteria, admissions committee ratings and performance measures.

IV. PREDICTING ACADEMIC PERFORMANCE

Despite the importance of other factors besides performance in admissions objectives, predicting academic performance remains a widely-discussed topic for both admissions committees and researchers in educational testing and measurement. A review of previous studies regarding the prediction of academic performance will introduce the concepts involved, highlight some of the problems encountered along with methodologies for resolving them, and frame expectations for an empirical analysis of RGS data. The empirical analysis of RGS data relies on several concepts from educational psychology and measurement:

- Test validation
- Reliability
- Measurement scales and transformations
- Clinical vs. statistical prediction

Validation involves an attempt to show that some measure, often a standardized test, serves as a *good predictor* of ability or achievement in some field of endeavor. Reliability measures the consistency of individual test scores (or other measures) from one form (or instance) to another.[29] As such, it affects validation directly, and changing measurement scales can improve reliability in some cases. (Without reliability, individual performance or test score measures would show excessive fluctuation, making relative standing highly dependent on the particular instance of measurement.[30]) Comparing clinical and statistical prediction addresses the issue of whether *objective* measures, such as test scores, offer better predictions than *subjective* measures, such as ratings by human judges.

Test Validation

Since the use of standardized tests (such as the GRE) in admissions processes has become practically universal,[31] attempts to show the appropriateness of these tests have become increasingly important. Validation, as Lee Cronbach has put it,

---

29 Lawrence W. Hecht and Donald E. Powers, *Graduate Management Admissions Test: Technical Report on Test Development and Score Interpretation for GMAT Users*, Graduate Management Admissions Council, Los Angeles, CA, 1986, p. 15. "The logic of reliability is perhaps best understood ... as the correlation coefficient between scores on two forms of the same test."

30 Ibid.

was once a priestly mystery, a ritual performed behind the scenes, with the professional elite as witness and judge. Today it is a public spectacle combining the attractions of chess and mud wrestling. Disputes about the appropriateness of tests impose a large responsibility on validators.[32]

This perspective on validity emphasizes that researchers treat validity as something possessed in greater or lesser degree by many possible measures. The general concept of validity, moreover, applies to other selection measures, such as previous grades.

Psychologists validate tests or other predictors in three fundamental ways:

- Construct validation
- Content validation
- Criterion-referenced validation

Construct validation questions whether or not a test measures the attribute(s) its creators claim to measure. Content validation considers the representativeness of test questions for the universe to which the test will apply. In other words, do the test questions, "truly sample the universe of tasks or situations they (supposedly) represent?"[33] Lastly, criterion-referenced validation concerns how test scores and other measures correlate with some outcome of interest. In the context of text validation, researchers typically refer to the tests (or other measures) as *predictors* and to the outcome of interest as a *criterion.* (Later sections will refer to test scores and other data considered by admissions committees as criterion measures, since they represent *admissions* criteria.) The same concept has equal relevance for other selection measures. For the purpose of enlightening the empirical analysis of RGS data, criterion-referenced validation offers more promise than construct or content validation.

Validation research has provided an interesting perspective on educational objectives. These may more properly belong in philosophical discussions of institutional objectives[34] or in an examination of the selection process,[35] but their impact on empirical validity research warrants further comment. As Rodney Hartnett and Warren Willingham have noted,

---

31 Ralph Nader, The Reign of ETS: The Corporation That Makes Up Minds, Allan Nairn and Associates, Washington, DC, 1980.

32 Lee J. Cronbach, "Five Perspectives on Validity Argument," in Howard Wainer and Henry I. Braun, Test Validity, Lawrence Ehrlbaum and Associates, Hillsdale, NJ, 1988.

33 Lee J. Cronbach, Essentials of Psychological Testing, Harper and Row, New York, NY, 1970, p. 124.

34 See Section II.

35 See Section III.

> the GRE Board cannot be fully responsible for the validity of the examinations it sponsors without also concerning itself with the validity of the criterion measures.[36]

Although they focus on improving the predictive validity of the GRE, Hartnett and Willingham emphasize the many different attributes that constitute *success* in graduate education, such as grades, examination scores, dissertation quality, departmental evaluations, degree attainment, and other evidence of acquired skill or knowledge.

Even ignoring that these attributes almost certainly fail to cover all the institutional and selection objectives of an academic institution, they demonstrate inherent difficulties in validation of any single predictor. As Cronbach notes,

> tests that predict one outcome will often not ... predict another, and a prediction formula that maximizes one outcome may reject persons who would be outstanding by another.[37]

Thus, while attention to criterion measures may improve the validity of individual predictors, multiple objectives may restrict the validity of any single predictor, particularly if the different criteria corresponding to those objectives have low intercorrelations. Moreover, when some of the educational objectives remain unclear or even unknown, the available criterion measures may not adequately or fairly reflect their successful achievement. But researchers will have few guides to choosing new criterion measures that do so. These considerations suggest that observed validity coefficients for available predictors and criteria illuminate only part of the selection problem, and that admissions decision makers have a major role to play in validation research, by articulating their institutional and selection objectives more clearly.

Reliability

Closely related to validity, reliability considers how well a test consistently measures whatever it purports to measure. It indicates what proportion of the variation in the results of testing a group of individuals comes from the systematic sources of variation that the test attempts to capture, and what proportion comes from non-systematic sources that fall into the class of errors in measurement.[38] Often expressed as a correlation coefficient for two sets of

---

36 Rodney T. Hartnett and Warren W. Willingham, The Criterion Problem: What Measure of Success in Graduate Education?, Educational Testing Service, Princeton, NJ, 1979.

37 Lee J. Cronbach, "Test Validation," op. cit.

38 Linda Conrad, et. al. (editors), Graduate Record Examinations Technical Manual, Educational Testing Service, Princeton, NJ, 1977.

similar measurements[39], reliability affects not only the interpretation of test measurements, but also the criterion-referenced validity of a test. By introducing errors of measurement, imperfect reliability attenuates the *true* correlation between the attributes purportedly measured by the test and the criterion of interest. Obviously, a test with no reliability, where all variation in scoring comes from random error, will, on average have no correlation with any criterion. This helps to explain the concern of ETS with the reliability of its tests, such as the GRE.[40]

Like validity, moreover, reliability applies to other selection measures besides standardized tests. Perhaps admissions officers use standardized test scores so widely because they cannot rely on third parties such as ETS to ensure the reliability of other predictors, such as previous grades. Nevertheless, admissions decision makers often use a variety of measures in making their selections. Reliability of each of these predictors thus becomes a concern for validity researchers. In addition, criterion measures such as grades may suffer from the same reliability problems as predictors, with the same attenuating effects on validity coefficients. Cronbach recognizes this problem in his discussion of outcome measures in education.

> If teachers use different bases for judgment and some are more generous than others, throwing grades from several algebra teachers into a single distribution merely piles one source of error on another. If an investigator is so unwise as to pool classes taught from different materials and by different methods, it becomes even harder to determine what if anything the pretests predict.[41]

Such reliability problems lead to a consideration of appropriate scales of measurement.

Measurement Scales and Transformations

Issues of appropriate measurement scales arise for both predictors and criteria. Such issues arise in part because mental attributes, unlike physical traits, often have neither an intuitive *zero point* nor agreed-upon *units* with widely understood meaning.[42] While most people can readily understand and agree upon the meaning of zero length, for instance, or one item costing

---

39 Note that the two sets of measurements can have large or small variances without necessarily affecting reliability. A high correlation between the two measurements indicates that the test (or other performance measure) still consistently measures the relative standing of individuals.

40 Ibid.

41 Cronbach, "Test Validation," op. cit.

42 William H. Angoff, "Scales, Norms and Equivalent Scores," in Robert L. Thorndike (editor), Educational Measurement, American Council on Education, Washington, DC, 1971.

two dollars more than another, William Angoff points out the difficulty of understanding the meaning of "absence of a mental ability."[43] Since both predictors and criteria in the empirical analysis of RGS data involve measurement of similarly non-intuitive attributes, a discussion of several measurement issues will prove useful.

The most straightforward approach to measurement, and the most common in the literature reviewed for this study, simply uses scores or measures as they appear in test results, transcripts, etc. Some measures, such as letter grades or ratings of A, B, C, and so on, require translation to a numeric scale in order to perform a variety of calculations, such as correlation or linear regression. Fortunately, scales such as the traditional 4-point scale for grades have gained widespread acceptance. Although such scales require assumptions about both ordinality and cardinality of letter grades, such assumptions appear explicitly in grade-point averages.

Three types of linear transformations can enrich the analysis of RGS data in useful ways. The first type of transformation converts any measure to a scale with a mean of zero and a standard deviation of one. Often referred to as z-scoring, this transformation works by subtracting the average of the raw scores and then dividing by the standard deviation of the raw scores. Many researchers prefer to choose a scale without a plausible zero value, such as the IQ scale, with mean 100 and standard deviation 10; or the T-scale, with mean 50 and standard deviation 10.[44] T-scales and similar transformations apply well to concepts like intelligence, the intuitive notion of "zero" intelligence seems confusing and unrealistic. An advantage of the z-score lies in the fact that a measure on that scale instantly conveys the distance from the raw mean, of a given score, in units of the raw standard deviation.

A second type of transformation converts any measure to a scale with the same mean and standard deviation as some other measure of interest. This allows re-scaled measures to reflect known differences in the characteristics of groups. This transformation works by multiplying the z-score by the desired standard deviation and then adding the desired mean. The advantage of this transformation lies in the ability to put two or more groups of measures on a common scale without forcing the two groups to have identical distributions, as the z-score and other similar

---

[43] Ibid.; however, the author has found the concept familiar during this research.

[44] Ibid.

transformations do. Each of the groups must have in common some measure other than the measure subjected to the transformation.

The third type of transformation explicitly links differential grading standards to the underlying abilities of groups of students taking different courses. According to Elliott and Strenta, "where the talent concentrates, instructors will partially adapt to it, and the grading will be more rigorous."[45] They adjusted GPA by subtracting an index value from every raw grade before averaging. This index involved two components. First, the authors adjusted for differences between departments by comparing, for each pair of departments, the average grades received in the two departments by students who took courses in both. "The average of all such pairwise differences for any department was its index value."[46] Thus, if a department produced, on average, higher grades than other departments, for students taking courses in that department and others, then it received a positive index; if it produced lower grades, on average, then it received a negative index. They computed a similar index for within-department differences. However, instead of pairwise comparisons for each course, they compared grades given to students in one course to all other grades given to those same students in other courses within the same department. Their methodology combined all courses that had fewer than 10 students who had taken at least one other course within the department into a single course for the purposes of computing a within-department index for those grades. They added the between-department index to the within-department index to create a single index for each course. They subtracted this index value from each course grade before calculating their index-adjusted GPA.

Apart from its effect on observed validity coefficients, the index adjustment used by Elliott and Strenta provides some useful guidance for the analysis of RGS admissions and performance data. Even within a single institution, or within a specific department, differing grading standards can reduce reliability of criterion measures. Adjustments to improve reliability can use simple transformations, if the data include observations for a subset of students under more than one grading regimen.[47] The analysis presented later in this report uses the Elliott and

---

[45] Rogers Elliott and A. Christopher Strenta, "Effects of Improving Reliability of the GPA on Prediction Generally and on Comparative Predictions for Gender and Race Particularly," *Journal of Educational Measurement*, Volume 25, Number 4, Winter, 1984, pp. 333-347.

[46] Ibid., p. 335.

[47] For a discussion of transformations to improve reliability when pooling data across institutions, see Henry

Strenta transformation to adjust for (possibly) differing grading standards within RGS. In addition, this research employs several transformations that imply alternatives to Elliott and Strenta's hypothesis. Section V provides examples of these transformations and discusses their implicit hypotheses about measurement standards. The guiding principle of that analysis, however, remains consistent with the objectives of the foregoing research: to improve the reliability of measures by adjusting for different standards.

Regardless of measurement scale, committee ratings represent predictions of how well the applicants will satisfy their institutional and selection objectives. Comparing clinical and statistical prediction illuminates more of the complexity of the concepts involved in this research. It also illustrates some of the issues related to measurement scales in particular.

Clinical vs. Statistical Prediction

Clinical psychologists, in the years since World War II, have investigated the extent to which human judgment can contribute to the accuracy of predictions of a number of outcome variables, such as therapy response and criminal behavior, in addition to academic performance.[48] Researchers measure the contribution of human judgment by comparing the predictive validity of such judgments with the predictive validity of purely statistical models, such as linear regression equations involving various predictor variables.

Some researchers have disputed the use of strictly linear models for statistical prediction. For example, Ronald Mitchelson and Don Hoy argue that linear formulae assume that various selection measures compensate for each other. More specifically, a sufficiently high score in one selection measure can produce a high computed prediction, despite a very low score on another selection measure.[49] If successful achievement in graduate education requires sufficiently high

---

I. Braun and Ted H. Szatrowski, "The Scale-Linkage Algorithm: Construction of a Universal Criterion Scale for Families of Institutions," *Journal of Educational Statistics*, Volume 9, Number 4, Winter, 1984, pp. 311-330.

[48] Robyn M. Dawes and Bernard Corrigan, "Linear Models in Decision Making," *Psychological Bulletin*, Volume 81, Number 2, pp. 95-106, February, 1974. See also, P.E. Meehl, Clinical vs. Statistical Prediction: A Theoretical Analysis and Review of the Literature, University of Minnesota Press, Minneapolis, MN; Steve Chan, "Expert Judgments Under Uncertainty: Some Evidence and Suggestions," *Social Science Quarterly*, Volume 63, Number 3, pp. 428-444, 1982; and John R. Hills, "Use of Measurement in Selection and Placement," in Robert L. Thorndike (editor), op. cit.

[49] Ronald L. Mitchelson and Don R. Hoy, "Problems in Predicting Graduate Student Success," *Journal of Geography*, Volume 83, Number 2, pp. 54-57, March-April, 1984.

scores on all predictors, then compensatory (linear, or additive) selection models will not predict success as well as non-compensatory models. Mitchelson and Hoy suggest one type of non-compensatory model: a conjunctive model. This conjunctive model involves multiplying statistical z-score transformations of measures of applicant predictors, in this case, GRE scores and undergraduate grade-point averages.[50] This multiplicative model assures a higher prediction or overall judgment for an applicant with above-average scores (no matter how mediocre) on both selection measures than for an applicant with a high score on one measure but a below-average score on the other measure. Multiplicative models can lead to additional problems, depending in part on the measurement scales used.[51]

While some non-linear models of decision making can avoid these problems, most experts have concluded that linear models lead to predictions that adequately reflect human decisions, even when the underlying decision process involves some degree of non-linearity.[52] Some research addresses specifically the problem of representing these decision processes. Researchers have not only compared the predictive validity of statistical models with the predictive validity of human judgments, but they have also used models to predict actual human judgments as well. Psychologists refer to these models as *paramorphic* representations of such judgments. This term stresses that actual decision processes may not involve weighing various predictor measures, but that explicit schemes such as linear combinations can nevertheless simulate the decision processes. The comparison of clinical and statistical prediction can use either the best-fit combination of predictor variables or the best fit paramorphic representation of admissions decisions for the statistical side of the comparison. The clinical side of the comparison assumes that admissions committee ratings represent the committee's best prediction of how well each applicant will perform if admitted.

---

50 Ibid.

51 For example, multiplicative models give a higher overall judgment to an individual with below-average measures on both characteristics than they do for an indivdual with one above-average characteristic and on below average. Although changing scales (to, for instance T-scales) can overcome this simple problem, it can also lead to relative positioning of individuals that does not differ substantially from a strictly linear combination.

52 Robyn M. Dawes and Bernard Corrigan, op. cit., and Robyn M. Dawes, "A Case Study of Graduate Admissions: Applications of Three Principles of Human Decision Making," *American Psychologist*, Volume 26, Number 2, pp. 180-188, February, 1971.

Researchers generally use the comparison of clinical and statistical predictions to argue about the efficacy of clinical procedures. In the case of admissions decisions, statistical models with higher predictive validity than admissions committee ratings, for some performance criterion of interest such as GPA, could replace the admissions committee for the purpose of making selection decisions. If the performance measure reflects the admissions committee's objective function, then selections made by the model with the highest correlation with the performance measure will maximize the committee's utility.

Several considerations limit the application of this utility-maximizing principle. Decision makers hold multiple objectives as individuals, and the committee process combines these objectives in mysterious ways. If the criteria employed in the comparison of clinical and statistical prediction do not capture all of these objectives, then higher predictive validity for the statistical model does not necessarily imply a higher level of utility for the decision makers using these models. (Clinical judgments might, in such a case, predict some measure of the *true* objective better than statistical models.) Researchers have tended to ignore this difficulty by assuming that available criterion measures represent institutional and selection objectives, at least when comparing clinical and statistical prediction. More general research on validity, on the other hand, has called attention to the criterion problem as it affects the magnitude of validity coefficients.[53]

Even if available criterion measures fairly represent institutional and selection objectives, researchers generally restrict statistical vs. clinical comparisons in additional ways. First, psychologists commonly evaluate the comparative validity of statistical models only after cross-validation. Cross-validation uses one sample from the population of interest to construct the linear combination of predictor variables (either the best fit predictor of the criterion or the best fit paramorphic representation of the decision makers' judgment) and a separate sample for computing the correlation between the statistical model and the criterion. This effectively doubles the required sample size for such comparisons. In many cases, total available sample sizes make cross-validation infeasible. Indeed, the empirical analysis for this study does not employ cross-validation, as the total sample size falls below 200 for all criterion measures.

---

[53] Rodney T. Hartnett and Warren W. Willingham, op. cit.

In addition to cross-validation, most researchers require comparisons of statistical and clinical predictions made on the basis of the same codable input. This does not require codification of all the variables used by human judges. However, if judges have access to information in addition to the variables used in the statistical model that affects the criterion measure, then the judges' predictions might very well have superior correlations with the criterion.[54] In the case of the RGS analysis, admissions committee members had access to slightly more information than the variables coded for the analysis (for example, letters of recommendation and course-by-course grades).[55]

Admissions committee members at RGS also directly affect later criterion measures for at least some of the applicants they evaluate during the admissions process. Not only do admissions committee members come from the RGS faculty, where they tend to teach core courses, but they also serve on the qualifying examinations committees and on dissertation committees.[56] These effects may bias the criterion measures in such a way as to increase the correlations between admissions committee ratings and criterion measures. (They may also simply reflect the superior ability of experienced admissions committee members to anticipate performance.)[57]

In contrast to incremental information, "some information typically requested in the admissions process may serve primarily as a source of error variance."[58] This problem helps explain the appeal of statistical models. If information other than GRE scores and undergraduate GPA[59] provides no explanatory power for criterion measures, then statistical models can not only capture all of the relevant information for decisions, they can also prevent human error in the

---

54 Robyn Dawes, op. cit.

55 In addition, admissions committee members could, in reality, use a more flexible, perhaps even non-linear model in their deliberations. As previoius discussions show, using non-linear statistical models usually does not improve predictability. See Robyn M. Dawes and Bernard Corrigan, op. cit., and Robyn Dawes, op. cit.

56 For example, two members of the committee supervising this dissertation, including the chair, have served on numerous admissions committees, taught the author's economics and statistics courses, and judged the author's qualifying examination performance.

57 Section III discusses these issues as well.

58 Marc J. Wallace and Donald P. Schwab, "A Cross-Validated Comparison of Five Models Used to Predict Graduate Admissions Committee Decisions," *Journal of Applied Psychology*, Volume 61, Number 5, pp. 559-563, 1976.

59 Ibid.

consistent application of decision principles. Furthermore, statistical models that simulate *expert judgments* may cost less than the experts' time and effort required in the decision process.

This focus on the costs involved in using human judgment has led some researchers in comparing clinical and statistical predictions to focus on the incremental contribution to validity of the clinical judges' observations.[60] This incremental analysis usually involves *statistical synthesis*, which treats human judgments as additional explanatory variables in the statistical predictive model. (By contrast, *clinical synthesis* treats the statistical prediction as an additional variable for judges to consider.) Although they do not strictly follow the ground rules of comparing clinical and statistical prediction,[61] both types of synthesis lead to assessments of the efficacy of clinical prediction. Since clinical synthesis requires decision makers to participate in experimental conditions and the development of statistical models before assessing clinical prediction with a fresh set of applicants, statistical synthesis offers more promise for the empirical analysis of RGS data. Before presenting the analysis, this section concludes with a review of empirical findings to provide a context for the correlations calculated from RGS data.

Results of Previous Empirical Research

The empirical research covered by this review falls into two groups:

- Validation studies of various predictors
- Comparisons of clinical vs. statistical prediction

The review of validation research focuses on the use of the Graduate Record Examination (GRE), in light of that test's relevance to the case study of RGS admissions and performance data. Much of the GRE validation literature covers additional predictors, particularly undergraduate grades. The literature on clinical vs. statistical prediction covers a wide variety of contexts, but this review discusses only a few illustrative examples. These examples convey the nearly universal superiority of statistical prediction[62] in settings most relevant to this dissertation.

---

60 John R. Hills, op. cit.

61 Robyn M. Dawes, op. cit.

62 Statistical models predict the outcome criteria used in such studies better than do clinical models. This does not imply that the outcome criteria represent the true institutional or selection objectives in each instance. However, the lack of evidence for the superiority of clinical prediction implies that either such potential objectives remain unknown or have little or no practical outcome criterion measurement.

Validation Studies of Various Predictors

Several studies give excellent summaries of previous empirical results in predicting academic performance. In addition to ETS,[64] Warren Willingham,[65] Gerald Lannholm,[66] and Rick Ingram[67] deserve special mention. Tables IV-1 and IV-2 reproduce summaries of validity studies presented by ETS.[68] Validity coefficients presented in this section represent correlation coefficients, as opposed to r-square values. These tables show median validity coefficients for various GRE scores and undergraduate grades across a variety of fields and several criterion measures.

Table IV-1[1]
Median Validity Coefficients for Various Predictors of Success in Nine Fields of Graduate Study

Predictor	Field of Study								
	Biology	Chem.	Educ.	Engin.	English	Math	Physics	Psych.	Soc. Sci.
GRE Verbal	0.18	0.22	0.36	0.29	0.21	0.30	0.02	0.19	0.32
	(7)	(14)	(15)	(11)	(6)	(6)	(6)	(23)	(11)
GRE Quantitative	0.27	0.28	0.28	0.31	0.06	0.27	0.21	0.23	0.32
	(8)	(13)	(14)	(10)	(6)	(6)	(6)	(22)	(10)
GRE Advanced	0.26	0.39	0.24	0.44	0.43	0.44	0.38	0.24	0.46
	(5)	(9)	(6)	(7)	(3)	(5)	(5)	(17)	(5)
GPA (Undergraduate)	0.13	0.27	0.30	0.18	0.22	0.19	0.31	0.16	0.37
	(2)	(7)	(5)	(4)	(4)	(4)	(4)	(15)	(6)
GRE and GPA	0.35	0.42	0.42	0.47	0.56	0.41	0.45	0.32	0.40
	(3)	(6)	(7)	(4)	(2)	(3)	(2)	(4)	(5)

[1] Source: Linda Conrad, et. al. (editors), Graduate Record Examinations Technical Manual, Educational Testing Service, Princeton, NJ, 1977.

[64] Linda Conrad, et. al. (editors), op. cit.

[65] Warren W. Willingham, "Predicting Success in Graduate Education," *Science*, Volume 183, January, 1974, pp. 273-278.

[66] Gerald V. Lannholm, Review of Studies Employing GRE Scores in Predicting Success in Graduate Study, 1952-1967, Educational Testing Service, Princeton, NJ, 1968.

[67] Rick E. Ingram, "The GRE in the Graduate Admissions Process: Is How it is Used Justified by the Evidence of its Validity?" *Professional Psychology: Research and Practice*, Volume 14, Number 6, 1983, pp. 711-714.)

[68] Linda Conrad, et. al. (editors), op. cit.

Table IV-2 also covers letters of recommendation as a predictor. Although they mask the true extent of variability among the empirical results, they do give some idea what to expect from the analysis of RGS data. For instance, across all academic fields, the median validity coefficient for at least one of the three GRE scores falls in the .20 to .40 range.[69]

Table IV-2[1]
Median Validity Coefficients for Various Predictors and Criteria of Success

Predictor	Measure of Success				
	GPA	Faculty Rating	Department Exam	Attain Ph.D.	Time to Completion
GRE Verbal	0.31	0.42	0.18	0.16	0.26
[2]	(27)	(5)	(47)	(18)	(46)
GRE Quantitative	0.27	0.27	0.26	0.25	0.23
	(25)	(5)	(47)	(18)	(43)
GRE Advanced	0.30	0.48	0.35	0.34	0.30
	(9)	(2)	(40)	(18)	(25)
GRE Composite	0.41		0.31	0.35	0.33
	(8)		(33)	(18)	(30)
GPA (Undergraduate)	0.37		0.14	0.23	0.31
	(15)		(30)	(9)	(26)
Recommendations			0.18	0.23	
			(15)	(9)	
GRE and GPA			0.40	0.40	0.45
			(16)	(9)	(24)

[1] Source: Linda Conrad, et. al. (editors), Graduate Record Examinations Technical Manual, Educational Testing Service, Princeton, NJ, 1977.

[2] Number of studies in parentheses

Furthermore, at least one of the three scores has a higher validity coefficient than undergraduate GPA in all fields. Combining undergraduate GPA and one of the GRE scores

[69] Validity coefficients presented in this dissertation do not generally adjust for restriction of range, or sample censoring. As Section VI shows, the RGS data do not suffer greatly from this problem. In addition, potential adjustments require fairly stringent assumptions about the underlying data. For further discussion, see Ross M. Stolzenburg and Daniel A. Relles, Calculation and Practical Application of GMAT Predictive Validity Measures, Graduate Management Admissions Council, Los Angeles, CA, 1985, and V. Srinivasan and Alan G. Weinstein, "Effects of Curtailment on an Admissions Model for a Graduate Management Program," *Journal of Applied Psychology*, Volume 58, Number 3, pp. 339-346, 1973.)

produces somewhat higher validity coefficients than does any of the GRE scores alone. GRE Advanced Test scores seem to have higher validity than either of the other two GRE scores. Unfortunately for this study, the GRE does not include an advanced test in public policy analysis, nor does RGS require advanced test results for application. Ignoring the results for GRE Advanced scores and a few outliers,[70] Table IV-1 indicates that typical validity coefficients range from about .20 to .30 for GRE scores and from about .15 to .30 for undergraduate GPA. Table IV-2 also indicates that Department Examsgenerally have less predictability than other criteria. Faculty Ratings seem slightly easier to predict than GPA (although it includes a small number of studies). In addition, Undergraduate GPA has more validity in predicting similar criteria (GPA and Faculty Ratings) than in predicting Completion and Time to Completion.

Table IV-3 lists the results of many of the validity studies reviewed by the authors cited above as well as others discovered during the course of this research. This table gives an indication of the variability in the magnitude of validity coefficients as well as in the types of criterion measures used. A glance at Table IV-3 suggests that validity coefficients outside the typical ranges cited above would not, per se, constitute startling results for the analysis of RGS admissions and performance data. Furthermore, the best predictor variables might well differ, depending on the success criterion chosen. This underscores the importance of Cronbach's observation, cited earlier in this section: "tests that predict one outcome will often not ... predict another."[71] Sections VIII and IX explore this observation in greater detail, considering various models to predict academic performance at RGS.

---

70 While it may seem unfair to treat medians as outliers, the relevance of Quantitative scores to English and Verbal scores to Physics seems questionable.)

71 Lee J. Cronbach, op. cit.

Table IV-3
Summary of GRE Validity Research

Year	Author(s)	Success Measure(s)	N	GRE Verbal	GRE Quant.	GRE Total	GPA Undergrad
'67	Alexakos	GPA	46	.34			.31
'75	Bean	GPA	91	.31	.10		
		Comprehensive Exam		.19	-.13		
'58	Benson	Obtained Ph.D.	56	[1]	[2]		
'60	Besco	Faculty Rating	82	.23	.27		
			26	.47			
			16	.51			
			40	.32			
			20	.47	.57		
			42		.30		
			44		.38		
			24		.56		
'63	Borg	GPA	175	.36	.37		
'84	Bornheimer	GPA	43	.39	.18	.31	
		Faculty Rating		.59	.43	.57	
'57	Capps & DeCosta	GPA	41			.34	.42
'55	Conway	GPA	NA	.72	.72		
'71	Dawes	Faculty Rating	86			.11	.21
'79	Dole & Baggaley	Faculty Rating	61	.35	.24		
				.22	.15		
'69	Ewen	Obtained Ph.D.	31	.17	.22		
		Percent A Grades		-.01	.17		
'74	Federic & Schuerger	GPA	47	.30	.01		
		Faculty Rating		.24	-.09		
'58	Florida St. University	GPA	7 to 96			.12 to .65	
'53	Gorman	GPA	NA	.23			
'70	Hackman et. al.	Faculty Rating	42	.22	.15		.28
				.20	.23		.05
				.21	.29		-.08
				.19	.32		-.22
'63	Harvey	GPA	89	.36	.30		.30
		Faculty Rating		.53 to .66			.28 to .37
'78	Hirschberg & Itkin	# First Author	87	-.22			
		# Author		-.32			

Table IV-3
Summary of GRE Validity Research

Year	Author(s)	Success Measure(s)	N	GRE Verbal	GRE Quant.	GRE Total	GPA Undergrad
'62	Johnson & Thompson	GPA	NA	.72 .45 .36	.72 .69		.63 .38 .46
'60	Law	Comprehensive Exam	46	.72			
'60	Lorge	Comprehensive Exam	165	.63			
'60	Maberly	GPA	104 148	.13 .23	.10 .10		
'65	Madaus & Walsh	GPA	NA	.44 .40 .47 .26	 .33 .46		
'69	Mehrabian	GPA Faculty Rating	266	.17 -.14	.27 .15		.10 .08
'60	Michael	Comprehensive Exam	41	.42	.55		
'86	Narvancik & Golsan	GPA	619	.34	.32	.38	.27
'68	Newman	GPA	66	.08	.21		
'55	Olsen	GPA Faculty Rating	43	.37 .39	.51 .52		
'63	Pitcher & Harvey	Faculty Rating	221 195	.33 .17	.38 .14		
'64	Robertson & Hall	Faculty Rating Peer Rating	73	.22 -.14	.18 -.24		
'61	Robertson & Nielson	Faculty Rating	50	.27	.20		
'57	Robinson	GPA	285				.38
'71	Roscoe & Houston	GPA Faculty Rating Obtained Ph.D. Faculty Rating	252 30	.32 .26 .21 .38	.21 .17 .28 .27		
'59	Rupiper	Obtained Ph.D.	25	[3]			
'61	Sistrunk	Comprehensive Exam	57	.32	.24		
'61	Sleeper	GPA	24	.31	.49	.54	
'70	Stordahl	GPA	120	.29	.07		.37

Table IV-3
Summary of GRE Validity Research

Year	Author(s)	Success Measure(s)	N	GRE Verbal	GRE Quant.	GRE Total	GPA Undergrad
'67	Stricker &	Comprehensive Exam	37	.20	.26		
	Huber		37	.40	-.15		
'85	Thornall &	GPA	462	.49	.30	.44	
	McCoy		27	.45	.35	.45	
			35	.42	.22	.36	
			58	.47	.37	.48	
'52	Wallace	Faculty Rating	100	.35			.57
'69	Wiggins et. al.	GPA	46	.20	.21		
			58	-.13	-.10		

[1] ANOVA, $p < .05$

[2] ANOVA, $p < .05$

[3] ANOVA, $p. < .05$

Comparisons of Clinical vs. Statistical Prediction

Previous discussions in this section examined the rationale for comparing clinical and statistical prediction. A brief review of previous comparisons might discourage the effort in the case study of RGS admissions and performance data: researchers have virtually unanimously concluded that statistical prediction always has more accuracy than clinical prediction, or at least never less accurate. Thus, according to John Hills,

> ... it seems clear that mechanical statistical combination is practically never improved upon by modification by clinical judgment. That is, given a specific set of data and a criterion, combining these data by means of a statistical procedure, such as a multiple regression equation, always yields as accurate predictions, and in some studies more accurate predictions, as does letting a clinician or judge examine those same data and make predictions from them. In fact ... the clinicians do not improve on the statistical predictions even when they are given the statistical predictions as part of the data they may consider.[72]

The contexts in which researchers have found the superiority of statistical prediction include psychological diagnosis, physical experimentation with shapes and shading, sports,

[72] John R. Hills, op. cit.

medicine and, of course, academic performance. Discussing a variety of such examples, Robyn Dawes and Bernard Corrigan observe that " ... no examples ... (within the standard limitations) ... have purported to show the superiority of clinical judgment."[73]

The consensus on the superiority of statistical prediction would seem to suggest that the only fruitful area for empirical analysis of RGS data would lie in determining the optimal statistical predictive model (or models, depending on the performance criteria used). However, Section X will briefly address clinical vs. statistical prediction in the case study of RGS data anyway, after Sections VIII and IX examine the prediction of various performance. This comparison cannot adhere to the classical ground rules discussed previously in this section because, technically, admissions committee members have access to data not coded as input to the statistical models (primarily, writing samples, recommendations and transcript details such as specific grades at specific institutions, grade trends, as well as occasional personal interviews).

Although this section provides a context for expectations regarding the empirical analysis of RGS data, several considerations might diminish the relevance of this review. First, previous validation studies do not explicity address public policy as an academic field. Second, institutional and selection objectives at RGS probably differ at least slightly from those at other institutions. Third, the selection and evaluation process at RGS also differ from those at other institutions. Thus, empirical analysis of RGS admissions and performance data do not fit the exact context of this review. Section X will consider these issues in light of the actual facts presented in Sections V through IX. Before turning to the empirical results of the analysis of RGS admissions and performance data, Section V presents an overview of the data, with descriptive statistics and an explanation of certain transformations used in the analysis.

---

[73] Robyn M. Dawes and Bernard Corrigan, op. cit.

V. FRAMEWORK FOR RGS DATA ANALYSIS

The data analyzed for this study cover applicants to the RAND Graduate School (RGS) for the years 1970-1988. These data fall into three categories:

- Criterion measures
- Ratings
- Performance measures

Criterion measures consist of all codable data available from each applicant's information folder. Ratings summarize evaluations given by admissions committee members during the admissions process. Performance measures include all grades received by each accepted applicant as well as several other measures of progress during each student's matriculation. This section gives descriptive statistics for each of these types of variable and concludes with a discussion and explanation of the types of transformations used for the analysis presented in the following sections.

Criterion Measures

As discussed in Section III, RGS collects a variety of data from each applicant before making its admissions decisions. Although some of this data does not readily enable encoding, Table V-1 gives means and standard deviations for as many variables as possible. Table V-1 also gives sample sizes for each variable, to indicate the extent of the missing data problem. In all, RGS records indicate that 575 people applied for admission during the years 1970-1988. However, the only variables in Table V-1 that have a complete sample of 575 relate to the attainment of prior educational degrees and whether or not the applicant had a quantitative field of study, as well as ethnic identity. While most of the data appear self-explanatory, some discussion of two types of variables may clarify their use.

For the remainder of this report, all binary variables will assign a value of 1 to Yes and 0 to No. Thus, the mean for the variable Female in Table V-1 indicates that females comprise 22% of all applicants.

Table V-1
Predictors For All Applicants

Predictor	Sample Size	Mean	Standard Deviation
GRE Total Score	459	1,204	198
GRE Quantitative Score	460	620	115
GRE Verbal Score	459	584	127
Undergraduate GPA	187	3.2	0.5
Last Institution GPA	211	3.5	0.5
Rating of Undergraduate Institution	543	3.1	0.8
Rating of Last Institution	569	3.2	0.7
BA Degree Earned	575	63%	48%
BS Degree Earned	575	34%	47%
MA Degree Earned	575	37%	48%
MS Degree Earned	575	29%	45%
Professional Degree Earned	575	25%	43%
Ph.D. Degree Earned	575	6%	24%
Quantitative Degree Earned	575	41%	49%
Age	547	30.8	6.7
Ethnic Minority	575	16%	37%
Female	574	22%	42%
Married	204	44%	40%
RAND Employee When Applying	570	7%	25%
U.S. Citizen	575	83%	38%

Notes:
Degree Earned variables may overlap (i.e., candidates may have multiple degrees).
Professional Degree Earned includes law, medicine, business, public health and public policy.
Quantitative Degree Earned includes mathematics, engineering and applied sciences, physical sciences and computer-related fields.

For all variables based on a standard 4-point grading or rating scale, this analysis uses the following convention:

- A = 4.0
- B = 3.0
- C = 2.0
- D = 1.0

Plus and minus grades add or detract .3 from the numerical equivalent of the letter grade. E (rarely used) and F both equate to 0.0. (Z translates to missing data.) For colleges and graduate schools, this research converted published rankings to the standard 4-point scale, as with GPA. For undergraduate schools, Barron's Guide to Colleges provides rankings of selectivity. The translation used for this study equated the most selective or top ranking with 4.3, the next with 4.0, and so on. For graduate schools, RGS compiled a list of top schools in various fields. The translation for top-ranked graduate programs proceeded according to the following rule:

- Top 3 = 4.3
- Next 3 = 4.0
- Next 4 = 3.7
- Next 10 = 3.3.

Graduate programs not on the RGS list of top graduate programs each received a rank of 3.0 (B). In practice, this minimum did not create a problem, since most students came from fairly selective undergraduate institutions.

Some of the variables listed in Table V-1 suffer from reliability problems. Binary variables such as Age and Female have extremely high reliability. However, as the last section indicated, differential grading standards make grade-point average an inherently unreliable measure. Unfortunately, the applicant pool comes from 364 different institutions (not counting different graduate programs within many of those). The scale-linkage algorithm developed by Braun and Szatrowski[74] thus has no practical application to this situation, since it would require knowing which applicants had been accepted to each of the 364 different institutions. It would produce linkages based on too few observations common to each pair of schools. Regarding Elliott and Strenta,[75] the data for these criterion measures would not provide any information to create comparative index values for schools or other common factors among applicants. However, later sections will show that their approach applies well to the analysis of admissions committee ratings, and can obviously assist the analysis of RGS grades. The analysis for this study also uses several simpler z-score and re-scaling transformations of criterion measures. Because the transformations for criterion measures resemble closely the transformations of

[74] Henry I. Braun and Ted H. Szatrowski, op. cit.

[75] Rogers Elliott and A. Christopher Strenta, op. cit.

ratings and performance measures, this section discusses transformations of all three types of variables together, after giving descriptive measures of ratings and performance.

Ratings

As discussed in Section III, admissions committee members assign letter grades to each applicant they consider. While the committee does not formally combine these ratings in determining who to admit, the committee does discuss each applicant and his or her ratings as a group. For this reason, as well as convenience, this study tends to use average ratings for each applicant. The 524 applicants who received at least one rating had average ratings of 2.6, or between a C+ and a B-. The same sample had a standard deviation of 1.0, or a full letter grade.

Because RGS does not have a specific requirement for the number of ratings given to any one applicant (and partly because the data might omit some ratings), the number of recorded ratings received by the 575 applicants ranges from 0 to 8. Figure V-1 shows the frequency distribution of the number of ratings. The typical RGS applicant receives 2 or 3 ratings. More importantly, though, the proportion of applicants accepted differs with the number of ratings.

Figure V-1
Distribution of Number of Judges

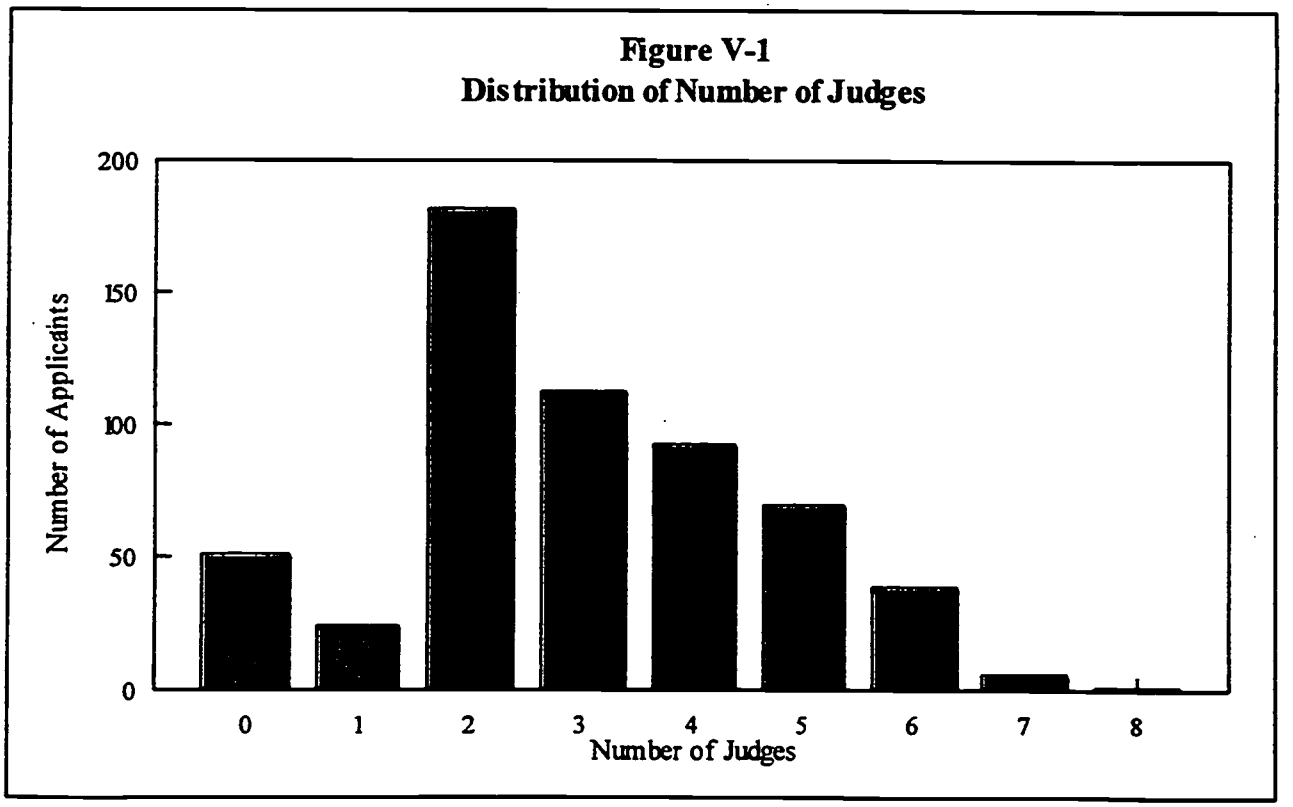


Table V-2 shows the average rating and the proportion admitted for applicants by the number of admissions committee members who rated each applicant. The high admissions rate for applicants with no ratings suggests a missing data problem. Indeed, most of these admitted applicants entered RGS among the first three cohorts, in 1970, 1971 and 1972. Furthermore, even after excluding applicants from these early cohorts, the admissions rate among applicants with no rating (60%) remains well above the rate for other applicants, except for those with four or more ratings (see Figure V-2).

Table V-2
Ratings and Admissions by Number of Judges

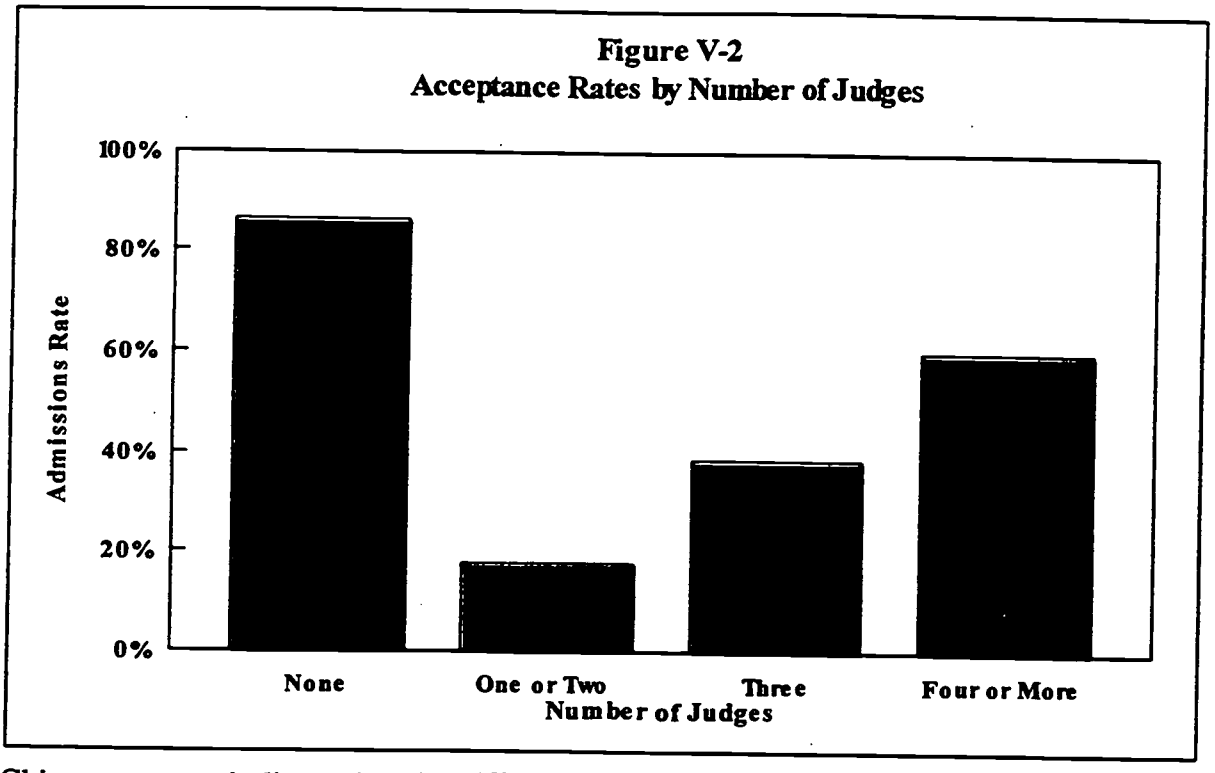
Number of Judges	Applicants	Rating Average	Percent Admitted
0	51	NA	86%
1	24	1.8	21%
2	182	1.9	18%
3	112	2.6	38%
4	92	2.9	60%
5	69	3.0	59%
6	38	3.1	63%
7	6	3.1	67%
8	1	2.8	0%
Total	575	2.6	43%

Because the number of judges has obvious relationships with average rating and proportion admitted, subsequent sections discuss the implications of these relationships in the context of predicting admissions. An analysis of variance in admissions decisions by number of judges indicates four logical groups:[76]

- None
- One or Two
- Three
- Four or More

[76] ANOVA shows that variance in raw ratings also changes significantly with Number of Judges (and between some pairs of individual Judges). However, transformations introduced in this section eliminate such heteroskedasticity. The significance of changes in average ratings and in admissions rates remains even after transformation.

Figure V-2
Acceptance Rates by Number of Judges



Chi-square tests indicate that the differences in admissions rates between these groups remain statistically significant while the differences within each (e.g., between 1 and 2 or between 5 and 6 judges) do not. Figure V-2 shows the differences in admissions rates between these groups. Although the percent admitted falls to 60% among applicants with no ratings after 1972, analysis of variance still suggests significant differences between this group and the others. Grouping these applicants together with applicants who have four or more ratings would reduce the number of groups without mixing fundamentally different admissions rates. This seems illogical, since the groups lie at opposite ends of a spectrum (number of judges) that has a significant relationship to the admissions decision. In order to make reasonable inferences, the data analysis that follows this section occasionally makes use of different samples based on the above grouping of the number of ratings per applicant (both including and excluding applicants before 1973).

In addition to differences by number of judges, ratings also vary somewhat among individual judges. Table V-3 presents sample sizes, means and standard deviations for ratings by individual judges. For anonymity, this study identifies judges by an arbitrary number from 1 to 25. An analysis of variance indicates that the average rating does not differ significantly across judges, though some comparisons suggest differences between specific pairs of members.

Table V-3
Ratings of Individual Judges

Judge	Applicants Rated	Average Rating	Standard Deviation
1	210	2.9	0.8
2	107	2.8	0.8
3	38	3.1	0.7
4	320	2.3	1.0
5	129	2.6	0.9
6	87	2.8	0.9
7	48	2.4	1.0
8	27	2.7	1.1
9	51	2.5	1.1
10	54	2.6	1.0
11	89	2.2	0.9
12	29	2.7	1.1
13	59	2.8	0.8
14	13	2.4	1.2
15	80	3.0	0.8
16	9	3.2	0.9
17	90	2.8	0.7
18	55	2.7	0.9
19	116	2.8	1.1
20	14	3.0	0.7
21	27	3.1	0.9
22	11	2.6	1.0
23	30	2.4	1.0
24	14	2.4	1.4
25	5	3.3	0.5
Total	1,715	2.6	1.0

Average ratings clearly suffer from the same kinds of reliability problems as grade-point averages. Unlike the data on criterion measures, however, the ratings data do offer some hope for transformations that increase reliability. As mentioned previously, the transformation proposed by Elliott and Strenta can apply well to RGS ratings. Table V-3 shows some reason for caution in performing this type of transformation, however: several judges rated fewer than 10 applicants, and a few judges accounted for a disproportionate share of the ratings. Varying judges in different years may also have affected their transformation. The last part of this section discusses the Elliott and Strenta transformation as well as other transformations.

Performance Measures

The RGS data provide a variety of useful performance measures, summarized in Table V-4. In addition to grades, RGS records indicate the results of comprehensive qualifying examinations, and perhaps more importantly, whether or not a student has dropped out of the program. In addition, at any point in time, each student's current status and date of matriculation determine the time to completion, for graduates; the length of time enrolled, for current students; and the time to attrition, for those who dropped out.

Table V-4
Performance Measures for All Enrolled Students

Performance Measure	Number of Students	Mean	Standard Deviation
GPA	161	3.1	0.7
Any C Grade	164	55%	50%
Number of Cs	164	1.2	1.7
Any "Core" C Grades	164	40%	49%
Number of "Core" Cs	164	0.7	1.2
Any First Year C Grades	164	43%	50%
Number of First Year Cs	164	0.7	1.2
Any F Grade	164	34%	47%
Number of Fs	164	0.7	1.4
Any "Core" F Grades	164	24%	43%
Number of Core Fs	164	0.4	0.8
Any First Year F Grades	164	23%	42%
Number of First Year Fs	164	0.4	0.9
Dropped Out	165	29%	46%
Failed Qualifying Exams	165	4%	20%
Distinction on Qualifying Exams:			
Any Distinction	165	14%	35%
General Distinction	165	5%	23%
Economics Distinction	165	10%	30%
Statistics Distinction	165	8%	28%
Social Science Distinction	165	7%	26%
Policy Analysis Distinction	165	8%	27%

Time presents special challenges for analysis. Section IX discusses the techniques used to attack this problem. For the purposes of this section, Table V-4 omits time to completion. Later sections will examine various performance measures in four distinct categories:

- Grades (including problem indicators)
- Attrition (whether or not a student completed the program)
- Qualifying exam results
- Completion time

In addition to overall grade point average, the data on RGS grades yield some other specific GPAs. As Section IV indicated, some researchers focus on first-year grades as the relevant performance measure. (This may reflect better predictability rather than any normative consideration of the merits of first-year vs. later grades.) The RGS data provide both first-year GPA and core course GPA. (Section III discussed the concept of core courses at RGS.) In addition, sorting courses into departments creates additional measures of GPA, although not all students take courses in all departments. Tables V-5 and V-6 show means and standard deviations for these various subsets.

Table V-5
RGS GPA[1] by First-Year and Core Courses

GPA Measure	Number of Students	Mean	Standard Deviation
First Year	164	3.1	0.7
Core	164	3.1	0.8

[1] 868 grades at RGS used a Pass/Fail rating system. This table excludes them.

Rather than showing the number of students, Table V-6 shows the total number of grades in each category. Table V-6 indicates that transformations based on Elliott and Strenta's methodology will tend to boost the reported grades for Soviet Studies, Military Studies, Economics and Quantitative Methods while decreasing those for Health Policy and Social Science courses.

Table V-6
RGS Grades by Department

Department	Number of Grades	Mean	Standard Deviation
Quantitative Methods	438	3.2	0.8
Economics	436	3.1	0.9
Social Science	309	3.4	0.7
Workshops	219	3.3	0.8
Military Studies	99	3.0	1.1
Health Policy	68	3.7	0.3
Soviet Studies	13	2.9	1.3
Total[1]	1,582	3.3	0.8

[1] 868 grades at RGS used a Pass/Fail rating system. This table excludes them.

Transformations

As mentioned in Section IV, the analysis for this study included several transformations of criterion measures, ratings and performance measures, which fall into three distinct categories:

- z-score transformations
- re-scaling transformations
- transformations based on the Elliott-Strenta hypothesis

Each of these transformations implies a hypothesis about the standards of measurement, or about sources of unreliability. The following examples illustrate each type and indicate the assumptions that would make each transformation appropriate.

Z-Score Transformations

This research used three distinct types of z-score transformations. Z-score transformations of variables all have mean 0 and standard deviation 1. The first type of z-score transformation involved converting variables such as each applicant's Undergraduate GPA or GRE scores to z-scores within each *cohort* of applicants (people applying in the same year). Table V-7 illustrates a z-score transformation of hypothetical GRE Verbal scores. Equal raw scores may not remain equal after transformation, depending on the distribution of scores within

cohorts. Thus, in Table V-7, Student 1 has a transformed score of -1.27, and Student 7 has a transformed score of -.63, because Student 7 belongs to a cohort with lower raw scores.

Table V-7
Illustration of Cohort-Based z-Score Transformation

Student	Cohort	Raw GRE Verbal	Cohort Mean	Cohort Standard Deviation	Transform: Cohort z-Score
1	1	600	700	79	-1.27
2	1	650	700	79	-0.63
3	1	700	700	79	0.00
4	1	750	700	79	0.63
5	1	800	700	79	1.27
6	2	550	650	79	-1.27
7	2	600	650	79	-0.63
8	2	650	650	79	0.00
9	2	700	650	79	0.63
10	2	750	650	79	1.27

The second distinct type of z-score transformation used in this research involved converting measures such as ratings or grades to z-scores within each pool of ratings or grades for a particular judge, before taking averages to develop a single measure for each student. For ratings, the z-score transformation used mean and standard deviations of ratings for each individual judge. Table V-8 shows an example of such transformation using hypothetical ratings data. Thus, a rating of 3.0 (B) from Judge 1 in Table V-8 would equate to a transformed rating of 1.54 (about 1.5 standard deviations above average for Judge 1). However, a rating of 3.0 from Judge 2 would convert to a z-score rating of 0.0 (exactly average for Judge 2). For grades, the z-score transformation used mean and standard deviations of grades for each specific course, by year and quarter. These z-score transformations for ratings and grades consider each judge's ratings separately before averaging. Thus, Student 4 in Table V-8 would receive an average z-score rating of about .38 (average of .77 and 0.00). Averaging these transformed ratings actually yields means and standard deviations slightly different from 0 and 1, respectively, but the differences do not reflect any fundamental difference from cohort-based z-score transformations.

Table V-8
Illustration of Judge-Based z-Score Transformation

Student	Judge	Rating	Judge Mean	Judge Standard Deviation	Transform: Judge z-Score
1	1	2.0	2.4	0.39	-1.03
2	1	2.0	2.4	0.39	-1.03
3	1	2.3	2.4	0.39	-0.26
4	1	2.7	2.4	0.39	0.77
5	1	3.0	2.4	0.39	1.54
1	2	2.7	3.0	0.19	-1.58
2	2	3.0	3.0	0.19	0.00
3	2	3.0	3.0	0.19	0.00
4	2	3.0	3.0	0.19	0.00
5	2	3.3	3.0	0.19	1.58

The third type of z-score transformation used in this study first takes averages of raw measures such as ratings or grades and subsequently converts them to z-scores within cohorts. This produces a relative ranking on each measure within each cohort. Table V-9 illustrates this post-hoc z-score transformation for the hypothetical ratings data from Table V-8. These post-hoc z-score transformations differ from a simple average of cohort-based z-score transformations. For example, the simple average of Student 1's z-score transformations from Table V-8 (-1.31) differs from Student 1's post-hoc z-score transformation from Table V-9 (-1.25), primarily because the standard deviations for each cohort pool of ratings differ.

Table V-9
Illustration of Post-Hoc z-Score Transformation

Student	Mean Rating	Cohort Mean	Cohort Standard Deviation	Transform: Post-Hoc z-Score
1	2.35	2.70	0.28	-1.25
2	2.50	2.70	0.28	-0.71
3	2.65	2.70	0.28	-0.18
4	2.85	2.70	0.28	0.54
5	3.15	2.70	0.28	1.61

Each of the z-score transformations just discussed implies a similar hypothesis about the evaluation of individuals. These transformations essentially eliminate all differences between

groups (cohorts, applicant pools by judge, etc.). The resulting variable reflects only *relative* ratings within groups. Z-score transformation assumes that differences between groups do not matter, arbitrarily equating measurement scales across groups. (The three types of z-score transformation imply slightly different objectives. Cohort-based z-scores imply a desire to treat performance and evaluation within cohorts as the most relevant indicators. Judge-based z-scores imply that evaluation within such applicant pools matters more than cohorts. Post-hoc z-scores imply that average performance or evaluation within cohorts has more relevance to objectives.)

Re-Scaling Transformations

As mentioned in Section IV, converting one variable to a scale with the mean and standard deviation of a second variable provides another basis for transforming the data used for this research. This second class of transformation relies on a z-score, but multiplies that z-score by the standard deviation (measured over the same group as the basis for the z-score) and adds the result to the mean of some *other* variable that provides a descriptive measure for the group (for instance GRE scores for the students rated by each admissions committee member). While this transformation converts the distribution of raw variables in a pool to the distribution of, for example, GRE scores within the same pool, *it does not convert the measure of a student with a high raw score to a high GRE score, nor vice versa.* A student with a rating two standard deviations above average for a judge rating students with a low GRE score mean and small GRE score standard deviation, would still receive a transformed rating equivalent to a relatively low GRE score. Conversely, a student with a very low GRE score who received an average rating from a judge rating students with an extremely high group average GRE score, would receive a transformed rating equivalent to that (extremely high group average) GRE score. Table V-10 illustrates this type of transformation on the data from Table V-8 to clarify the methodology.

GRE-score transformation of grades follows a similar logic. As Table V-10 demonstrates, an individual need not have a high GRE Verbal score in order to have a high transformed rating, since this transformation preserves order within pools. Since the GRE-score transformations of variables all have the same mean and standard deviation as each pool's GRE score distribution, this section does not describe the specific transformations in further detail.

Table V-10
Illustration of GRE Verbal Score Re-Scaling Transformation

Student	Judge	Rating	Transform: Judge Rating z-Score	GRE Verbal Score	Judge GRE Mean	Judge GRE Standard Deviation	Transform: Judge-Pool GRE-Scale Rating
1	1	2.0	-1.03	600	600	141	455
2	1	2.0	-1.03	500	600	141	455
3	1	2.3	-0.26	700	600	141	563
4	1	2.7	0.77	800	600	141	709
5	1	3.0	1.54	400	600	141	817
6	2	2.7	-1.58	800	700	71	588
7	2	3.0	0.00	750	700	71	700
8	2	3.0	0.00	700	700	71	700
9	2	3.0	0.00	600	700	71	700
10	2	3.3	1.58	650	700	71	812

As with the z-score transformations, this research uses three distinct types of re-scaling transformations, each analogous to one type of z-score transformation described previously in this section. Thus, a cohort-based re-scaling transformation converts one set of variables (such as age or rating of undergraduate institution) to the mean and standard deviation of some other measure within each cohort of applicants. Judge-based re-scaling transformations convert grades or ratings to the distribution of some other measure for each pool of students rated by an admissions committee member or instructor. Lastly, post-hoc re-scaling transformations convert averages of one set of variables (such as ratings or grades) to the mean and standard deviation of some other variable for each cohort.

The re-scaling transformations just discussed imply a more stringent alternative to the hypothesis implied by z-score transformations. Using GRE scores or some other measure as the basis of transformation imposes a common measurement scale on each group (cohort, course, etc.), regardless of what scale each judge uses. This type of transformation still implies that judges evaluate their subjects on a relative basis. However, these re-scaling transformations compare groups based on their distribution of some common measure like GRE scores. High or low average evaluations by a judge have no effect on the transformed measures. Transformed measures still reflect relative ratings within groups. However, they also reflect differences

between groups. Furthermore, they imply that a relatively high rating in a strong group represents a higher absolute rating of an individual.

Transformations Based on the Elliott-Strenta Hypothesis

As discussed previously the Elliott-Strenta methodology applies to both ratings variables and grades. Since Section IV described the methodology used by Elliott and Strenta, this section only describes a slight modification to improve upon their intent. The algorithm used by Elliott and Strenta appears to over-compensate for differential grading or rating standards. To demonstrate how their method over-compensates, Table V-10 presents a simplified example of two students whose grades in two courses differ. The example ignores the two-tier approach to departments and courses within departments, but the flaw applies to both stages.

Table V-11
Hypothetical Data to Illustrate Elliott-Strenta Transformation Flaw

Student	Course	Grade	E-S Grade	*True* E-S Grade
1	1	4.0	3.0	3.5
2	1	4.0	3.0	3.5
1	2	3.0	4.0	3.5
2	2	3.0	4.0	3.5

According to Elliott and Strenta, since the students in course 1 received As while they received B grades in course 2, transformation should subtract 1 point from course 1 grades and add 1 point to course 2 grades. This simplistic example shows that Elliott and Strenta would simply reverse the grades in this case. Obviously, the spirit of Elliott and Strenta requires that the grades in this case all equate to the average. That is, an A in course 1 should represent the same level of performance as a B in course 2. This change requires a slight modification to the Elliott-Strenta transformation. Instead of pairwise differences (for courses or departments, or judges in the case of ratings), the modified transformation uses differences from the overall average (of courses, departments or judges, as required). With this change in mind, the remainder of this section discusses some of the implications of each type of transformation used in the analysis of RGS data.

The previous section discussed the implicit assumptions used by Elliott and Strenta to create their transformation. In contrast with z-score transformations, the Elliott-Strenta hypothesis allows group differences to affect measurement scales. Unlike re-scaling transformations, however, the Elliott-Strenta hypothesis uses the evaluations themselves to determine differences between groups. This implies that individuals remain consistent across groups. In other words, a difference in the evaluation of Student A in groups 1 and 2 implies that the standards of evaluation differ between groups 1 and 2.

Advantages and Disadvantages of Transformations

The z-score transformation allows analysis to explore whether or not either the admissions decision or performance depends on the *relative* magnitude of predictor variables among pier groups. In other words, the admissions committee might simply choose to admit the top x percent of applicants each year, based on some combination of predictor variables. Similarly, performance measures might also use such a *relative* ranking. While analysis of variance techniques make it possible to explore these possibilities by using dummy variables for all but one cohort, such techniques lose many degrees of freedom. (Small sample size and student to cohort ratios make this a problem for the RGS data analysis.) The GRE-score transformation has advantages similar to the z-score transformation, and can also retain between-cohort differences that may affect admissions rates and performance measures. This transformation introduces two additional problems, however. First, by using a criterion measure as the basis of transformation, this technique may make GRE-related effects appear as effects of other predictor variables. Second, when regression models also include GRE scores, this transformation may increase colinearity. The Elliott-Strenta hypothesis might avoid these disadvantages. However, such transformations do not appear to provide much insight in this case.

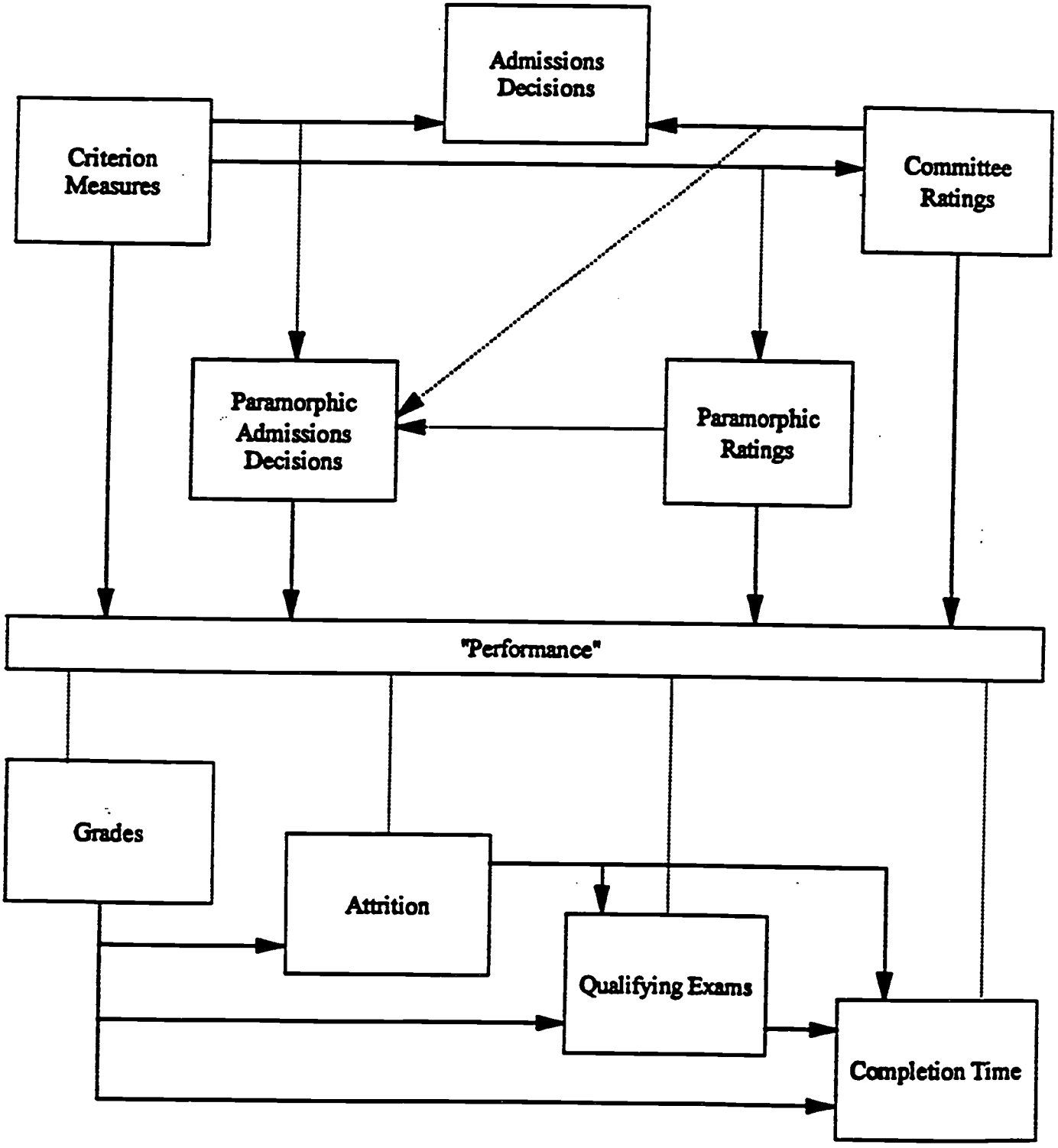
For ease of presentation, the remainder of this report will generally use the GRE Quantitative scale transformations discussed above in order to show the impact of such transformations on models of both the admissions decision and performance measures. Most

results will show the use of raw variables as a basis for comparing transformed variables. The choice of GRE scale transformation has several justifications:

- GRE-scale transformations lead to better-fit models than do other transformations.
- Between-cohort and between-course variation in GRE scores do not have statistical significance. Thus, the transformation does not appear to introduce bias toward GRE scores as predictors. At the same time, this transformation allows the actual differences between judges and cohorts to affect the levels of ratings or grades. (Ratings and grades do not vary between judges or courses in statistically significant amounts.)
- The implicit assumptions for z-score transformations seem too simplistic. Essentially, z-score transformations do not allow differences between judges or cohorts to affect ratings or grades.
- Although z-score transformations require a computationally simpler methodology, ample computing resources make this consideration less important.
- Transformations based on the Elliott-Strenta hypothesis do not provide much insight, as the transformed variables have negligible impact on correlations among criterion measures, ratings and admissions decisions and performance measures.

The next several sections of this report each deal with specific aspects of the general problem of academic performance prediction. Figure V-3 shows how these aspects relate to each other in one framework. Section VI begins with a more detailed look at the differences in both predictor and ratings variables between admitted and rejected applicants.

Figure V-3
Modeling Relationships to Analyze Admissions Decisions and Performance



VI. RATINGS

This section begins by comparing criterion measures for accepted and rejected applicants. While subsequent discussion explores the relationship between these criteria and admissions in greater detail, Table VI-1 gives a broad indication of the differences between the two groups.

Table VI-1
Differences in Criteria: Admitted vs. Rejected Applicants

Criterion	Range Restriction	Difference in Means	t-Statistic	Significant ($p \leq .05$)
GRE Total Score	19%	193	12.1	YES
GRE Quantitative Score	10%	107	11.6	YES
GRE Verbal Score	18%	86	7.7	YES
Undergraduate GPA	0%	0.2	2.7	YES
Last Institution GPA	4%	0.2	2.8	YES
Rating of Undergraduate Institution	0%	0.6	8.5	YES
Rating of Last Institution	0%	0.0	0.5	
BA Degree Earned	0%	7%	1.7	
BS Degree Earned	0%	-5%	-1.3	
MA Degree Earned	0%	-4%	-1.0	
MS Degree Earned	0%	8%	2.2	YES
Professional Degree Earned	0%	0%	0.0	
Ph.D. Degree Earned	0%	-3%	-1.5	
Quantitative Degree Earned	0%	7%	1.7	
Age	11%	-1.6	-2.9	YES
Ethnic Minority	0%	-8%	-2.6	YES
Female	0%	2%	0.5	
Married	0%	-11%	-1.2	
RAND Employee When Applying	0%	13%	5.8	YES
U.S. Citizen	0%	6%	1.8	

Table VI-1 shows the restriction of range as well as results of t-tests of significance on differences in means for criterion measures of accepted and rejected applicants. Restriction of range in each case indicates the percentage of the range among all applicants omitted in the sample of accepted applicants. As an example, GRE Quantitative Score among all applicants has a range from 220 to 800, while the range among accepted applicants starts at 280 and extends to

800. Thus, an analysis of the relationship between GRE Quantitative Score and performance (e.g., GPA) will reflect only 90% of the original range of GRE Quantitative Score among all applicants. No additional restriction of range resulted from admitted applicants who chose not to attend RGS, except for Age, which lost an additional 26% of the range of accepted applicants.

Table VI-1 suggests that several factors may influence admissions decisions at RGS. However, significant differences between accepted and rejected applicants do not mean necessarily that a factor influences admissions decisions. Some factors may show spurious correlations based on their relationships not only with admissions decisions but also with other factors that influence decisions. Multiple regression analysis can show which factors retain a significant relationship to admissions decisions after adjusting for other factors. Before presenting specific results of such analyses, this section explores the relationship between criterion measures and ratings.

This research used standard linear regression techniques to show the relationship between various criterion measures and committee ratings. The general assumption of these models specifies the following functional relationship:

vi-1:	**Average Rating**	=	$\mathbf{f(X_1, X_2, \ldots X_n)}$
where	$X_1, X_2, \ldots X_n$		represents some combination of the criteria coded for the analysis and shown in Table V-1;
	and		
	$f(X_1, X_2, \ldots X_n)$	=	$b_0 + b_1 x X_1 + b_2 x X_2 + \ldots + b_n x X_n$

Equation vi-1 assumes a first-degree polynomial for the functional form of the right-hand side. All of the regression equations presented in this dissertation include only variables with less than or equal to a 5% level of significance. This means that, if the *true* coefficient has a value of zero, the regression coefficient will have an absolute value higher than the observed value no more than 5% of the time due to random chance alone. Regression equations also show coefficient standard errors, so that readers can compute more exact probabilities for each coefficient.

Equations vi-2 and vi-3 show the best-fit regression equations for raw and transformed ratings. The fit between criterion measures and transformed ratings slightly exceeds the fit for raw ratings.

	Equation:		vi-2	vi-3
	Average Rating			
=			-1.6	85
			(0.42)	(49)
+	GRE Quantitative Score	x	0.0034	0.46
			(0.00050)	(0.057)
+	Undergraduate GPA	x	0.56	68
			(0.11)	(13)
+	Number of Judges	x	0.12	13
			(0.043)	(4.9)
	R-Square (adj.):		0.51	0.56

Note: Coefficient standard errors in parentheses
Equation vi-2 uses raw ratings; equation vi-3 uses GRE Quantitative Scale (by Judge) ratings
117 observations

The best-fit regression equation for average raw rating includes three variables: GRE Quantitative Score, Undergraduate GPA and Number of Judges. For transformed ratings, the *same* three variables comprise the best fit regression model. Section IV established expectations that test scores and previous grades would influence ratings, since they commonly have significance in the prediction of performance. However, including number of judges *in addition to* these criterion measures suggests a complication in the analysis. Committee members might influence each other's ratings, since they review better candidates more often. Regardless of the reason for the significance of number of judges in predicting average rating, analyses of the relationships between ratings and admissions and between ratings and performance will need to account for it. Sections VII through IX all consider number of judges as a potential predictor variable for admissions and various performance measures.[77]

[77] Essentially, this analysis treats Number of Judges as an endogenous variable included in regression equations. Analysis of variance reveals that neither ratings nor performance show heteroskedasticity (significant changes in error variance) with respect to Number of Judges, at least after transformation.

Equations vi-2 and vi-3 can yield point estimates and provide heuristic explanations for the relationship between criterion measures and average committee ratings. For example, using average values for GRE Quantitative Score, Undergraduate GPA and Number of Judges, the predicted values from equations vi-2 and vi-3 represent the average ratings expected for students with these average characteristics. The following calculations demonstrate this:

Predicted Ratings for a Typical Candidate

	Equation:	vi-2	vi-3
	Average Rating (Predicted)		
=	b_0 (Constant)	-1.6	85
+	b_1 x GRE Quantitative Score	0.0034 x 650	0.46 x 650
+	b_2 x Undergraduate GPA	0.56 x 3.2	68 x 3.2
+	b_3 x Number of Judges	0.12 x 3.8	13 x 3.8
=		2.9	650

Notes: Equation vi-2 uses raw ratings; equation vi-3 uses GRE Quantitative Scale (by Judge) ratings
Average values of independent variables for regression sample

These calculations do not yield the same average ratings as do the summary tables earlier in Section V. The regression coefficients reflect only the 117 observations without any missing values for any of the variables. Since the averages reported in the summary tables reflect all non-missing values for each variable, these values naturally differ from the values calculated above.[78] However, they do not differ by much: the difference in average raw rating (.3) and average re-scaled rating (35) both lie within a third of one standard deviation of their respective averages. (The coefficients reflect only two significant digits.)

[78] This dissertation does not report results using models that account for missing values by substituting predicted values. Such methods do not affect regression results significantly, and do not change the comparisons made using various transformations. Furthermore, such substitution would make the context of prior validation studies less relevant unless this analysis first compared models using substitution to models without substitution.

Equation vi-2 shows that the committee will give a typical candidate an average rating of 2.9, or almost a B. An increase in GRE Quantitative Score from 650 to 750 would raise the expected average rating to about 3.2, or about a B+. Equation vi-3 provides a less obvious but equivalent interpretation. The committee gives a typical candidate an average rating equivalent to 650, after transforming individual ratings to a scale with the mean and standard deviation of GRE Quantitative Scores for the candidates rated by each judge. Increasing GRE Quantitative Score from 650 to 750 raises the expected average transformed rating to 700 (after rounding).

Thus, the same 100 point rise in GRE Quantitative Score leads to a .34 increase in average raw rating and an increase of 46 in average transformed rating. This increase in average transformed rating represents a greater magnitude relative to the standard deviation of transformed ratings.[79] Of course, neither equation implies causality. Rating by an additional judge would raise the expected average raw rating by .12 and the expected transformed rating by 13; but this does not mean that ratings would increase *because* of the additional ratings.

As mentioned previously, equations vi-2 and vi-3 show that transformed ratings have a better fit with criterion measures than do raw ratings. This research examined several alternative measures of committee ratings as well. In all such regressions, the same three variables remain significant: GRE Quantitative Score, Undergraduate GPA and Number of Judges. Maximum ratings show a closer fit with the criterion measures than do average ratings. For instance, the regression of maximum raw rating on these three variables yields an adjusted r-square of .57. This exceeds the adjusted r-square using average raw ratings (.51) by .06. The regression of maximum transformed ratings (using the distribution of GRE Quantitative Scores by Judge) also yields a higher adjusted r-square than does average transformed rating: .62 versus .56. In general, transformed ratings provide the best fit with criterion measures, but transformed criterion measures *do not* improve the fit. Although maximum ratings provide a better fit with criterion measures than do average ratings, average ratings provide a better fit with both admissions decisions and subsequent performance measures. Therefore, this dissertation focuses on average committee ratings. The next section discusses the relationships between criterion measures and admissions and between ratings and admissions.

---

[79] The difference in r-square values between equations vi-2 and vi-3 indicates this as well. A one standard deviation increase in GRE Quantitative Score leads to a .51 standard deviation increase in average raw rating and a .56 standard deviation increase in average transformed rating.

VII. ADMISSIONS

For the analysis of admissions decisions as dependent variables, this research used logistic regression techniques. Like linear regressions, these models also assume a functional relationship between the dependent variable (i.e., admission) and one or more independent variables. However, in logistic regression, the dependent variable has a binary format, and the functional relationship uses a transformation (called a logit function) of the dependent variable:

$$\text{vii-1a: } \ln[p/(1-p)] = f(X_1, X_2, \ldots X_n)$$

$$\text{vii-1b: } p = \frac{\exp[f(X_1, X_2, \ldots X_n)]}{1 + \exp[f(X_1, X_2, \ldots X_n)]}$$

where	ln[ ]		denotes the natural logarithm;
	p	=	probability of admission;
	$X_1, X_2, \ldots X_n$		represents some combination of the criteria coded for the analysis and shown in Table V-1, as well as committee ratings; and
	exp		denotes exponentiation by e, the base of natural logarithms
	$f(X_1, X_2, \ldots X_n)$	=	$b_0 + b_1 \times X_1 + b_2 \times X_2 + \ldots + b_n \times X_n$

As with the linear regression models discussed in the previous section, equation vii-1a assumes a first-degree polynomial for the functional form of the right-hand side. Using exponentiation to transform equation vii-1a provides a more convenient form for directly computing the expected admissions decision, given specific values for the independent variables. Equation vii-1b makes the influence of each separate independent variable more difficult to interpret than in standard linear regression. However, this form makes it easier to compute the estimated probability of admission from specific values of the independent variables. Equation vii-1b also provides the direct basis for graphing the relationships between the independent variables and the dependent variable. This dissertation presents equations in the form of equation vii-1a for convenience. Occasional graphs use the form of equation vii-1b.

Predicting Admissions Using Criterion Measures

As with average ratings, the best-fit logistic regression equation for the admissions decision includes GRE Quantitative Score and Undergraduate GPA. Using transformations of selection criteria indicates that the same variables comprise the best-fit equation as with the use of raw criterion measures. However, the fit between transformed selection criteria and admissions decisions compares unfavorably with the fit using raw selection criteria. Equation vii-2 shows the best-fit logistic regression equation for the admissions decision. This equation uses the same sample as equations vi-2 and vi-3.

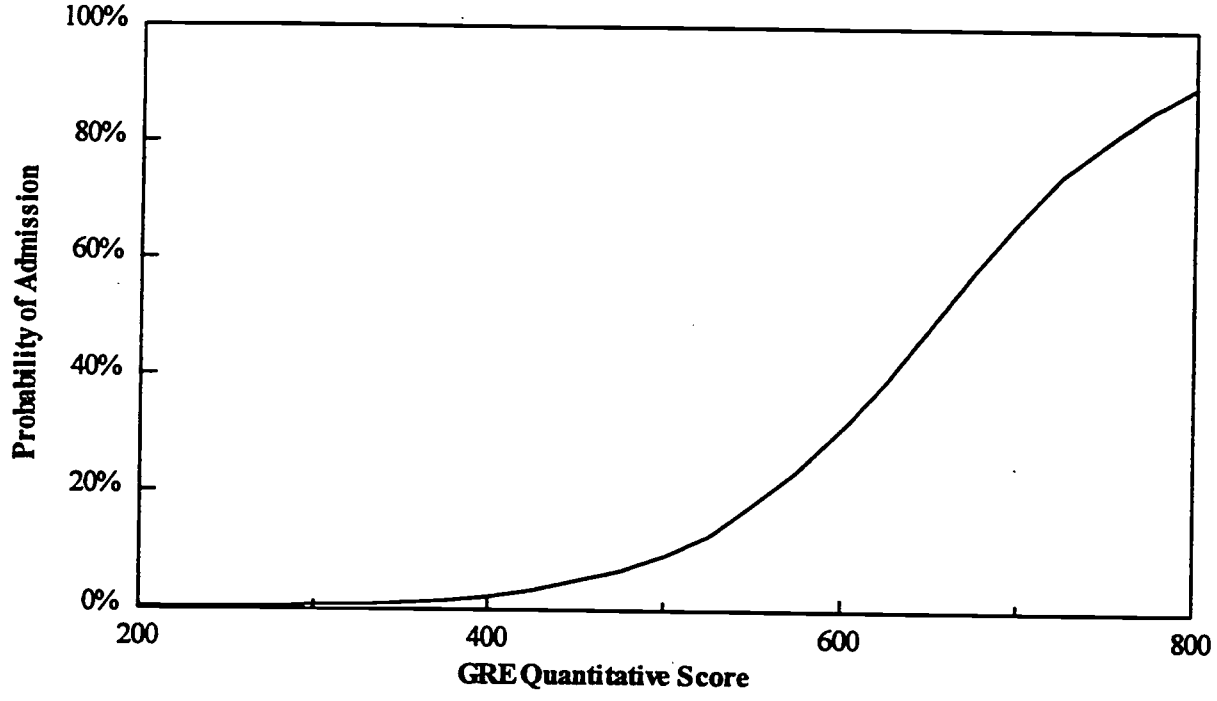
vii-2: ln[p/(1-p)]	= -14		
	(2.8)		
	+ 0.015	x GRE Quantitative Score	
	(0.0030)		
	+ 1.4	x Undergraduate GPA	
	(0.48)		

Notes: p = probability of admission
Coefficient standard errors in parentheses
123 observations
"Psuedo" r-square = .30, calculated as percentage change in log likelihood after estimation, using log likelihood before estimation as a base

As equation vii-2 demonstrates, Number of Judges *does not have statistical significance in predicting admissions* after including GRE Quantitative Score and Undergraduate GPA.

Figure VII-1 shows the relationship between GRE Quantitative score and the probability of admission, given an average value for Undergraduate GPA. Figure VII-1 indicates that an increase in GRE Quantitative score from 550 to 650 will, on average, lead to an increase in probability of admission from 18% to 49%, or 31%. An increase in GRE Quantitative score from 650 to 750 will, on average, lead to an increase in probability of admission from 49% to 81%, or 32%. These examples reflect approximately symmetric changes around the average GRE Quantitative score (about 650 for this sample). Changes of GRE Quantitative score by 100 points at other levels (e.g., from 350 to 450) would lead to smaller changes in the probability of admissions.

Figure VII-1
Admissions as a Logit Function of GRE Quantitative Score



Undergraduate GPA = 3.0

The pseudo r-square gives one indication of the accuracy of predicted probabilities. Since the actual outcomes (admissions) have a binary format, an alternative approach to the predicted probabilities can also use a binary approach. To follow such an approach, Table VII-1 shows a cross-tabulation of actual and predicted admissions using equation vii-2. This table translates predicted probabilities into binary outcomes by ranking the predicted probabilities and transforming into admitted the same percentage as the sample admissions rate.

A random ordering of candidates provides a logical benchmark. On average, such a random selection process will match actual admissions decisions in $[p^2 + (1-p)^2]$ of the sample, where p represents the sample admissions rate. That is, if the sample admissions rate equals .5, then random ordering with a cut-off percentage of .5 will match actual decisions in 50% of the sample, on average. In the case of RGS admissions, the actual admissions rate equals 43%, but 56% for students included in the sample for equation vii-2. Random selection using the methodology just described would match actual decisions in about 51% of the sample.

Table VII-1
Actual vs. Predicted Admissions Using Criterion Measures

		Predicted Admission		
		Yes	No	Total
Actual Admission:	Yes	38	16	54
	No	16	53	69
	Total	54	69	123

Proportion correct = .74

Table VII-2
Admissions Rate by Matriculation Year

Matriculation Year	Number of Applicants	Admissions Rate
73	10	70%
74	24	38%
75	29	38%
76	32	28%
77	34	38%
78	17	47%
79	29	48%
80	28	36%
81	15	60%
82	32	34%
83	42	38%
84	38	34%
85	51	33%
86	35	46%
87	52	35%
88	56	41%
Total	524	39%

Notes: Excludes 1970 to 1972 (no records of rejected applicants, and most applicants had no ratings)

As Table VII-1 shows, using predicted admissions from criterion measures matches actual decisions in proportions well exceeding this benchmark. Using predicted probabilities

from equation vii-2 would improve the percentage of correct predictions to 74%. Since admissions committees in practice consider only one group or cohort of applicants at a time, the logistic regressions just shown can form even more accurate predictions by sorting predicted probabilities within each cohort and then using the actual admissions rate of each cohort as a cut-off point. Table VII-2 shows the variation in admissions decisions by cohort[80], for all applicants with at least one rating. Using these varying percentages as cut-off points, the proportion of correct predictions rises from 74% to 80% (an increase of 8 correct predictions or 6% of the sample). As the next few equations and tables demonstrate, using average committee rating as the independent variable increases the proportion of correct predictions.

Predicting Admissions Using Ratings

Equations vii-3 and vii-4 show the logistic regression equations for the admissions decision using raw and transformed ratings.

Equation:		vii-3	vii-4
ln[p/(1-p)]			
=		-10	-21
		(0.93)	(1.9)
+ Average Rating	x	3.6	0.033
		(0.33)	(0.0029)
"Pseudo" R-Square:		0.54	0.56

Note: p = probability of admission
Coefficient standard errors in parentheses
Equation vii-3 uses raw ratings; equation vii-4 uses GRE Quantitative scale (by Judge) ratings
523 observations: 1 observation deleted for equation vii-3 to match sample for equation vii-4
"pseudo" r-square calculated as percentage change in log likelihood after estimation, using log likelihood before estimation as a base

[80] Table VII-2 suggests the inclusion of Number of Applicants and Cohort (or some time trend indicator) as endogenous variables. However, unlike Number of Judges, these variables do not retain significance after using transformed ratings and/or grades. Thus, such variables would make comparisons between models using raw and transformed measures more difficult. It also seems reasonable to argue that such transformations, along with the inclusion of other significant explanatory variables, adequately reflect the appropriate regression models.

As with the prediction of ratings, the fit using transformed ratings again slightly exceeds the fit using raw ratings.[81] Thus, an increase in average transformed rating leads to a slightly greater increase in probability of admission than the equivalent increase in average raw rating. Increasing average transformed rating (equation vii-4) from 550 to 650 increases the probability of admission by 56% (from 5% to 61%). This magnitude exceeds the impact (31%) of the same increase in GRE Quantitative Score (equation vii-2).

Using the same methodology as in Table VII-1, random selection would match actual decisions in about 52% of the sample, given a sample admissions rate of 39%. Predictions made from equations vii-3 and vii-4 match actual decisions in proportions well exceeding this 52% benchmark. Using cohort-varying percentages as cut-off points, the proportion of correct predictions rises even further for both raw and transformed ratings. For raw ratings, the proportion rises from 87% to 89% (an increase of 12 correct predictions or 2.3% of the sample).. For transformed ratings (GRE Quantitative scale by Judge), the proportion of correct predictions rises from 88% to 90% (an increase of 8 correct predictions or 1.5% of the sample). Thus, when predicted decisions account for cohort-varying admissions rates, the predictions improve, and the impact of transformation diminishes. Transformed ratings provide more accurate predictions in either case.

Predictive Value of Additional Ratings on the Margin

A comparison of several logistic regressions can show the predictive value of adding additional ratings on the margin. Table VII-5 shows both the "pseudo" r-square and the proportion of correct admissions decisions derived from logistic regressions of admissions decisions using the average of one to four randomly selected ratings as the independent variable. The samples used for these regressions include only those applicants with at least four ratings.[82] Table VII-5 shows an apparent inconsistency between the "pseudo" r-square and the proportion of correct

---

81 The sample size for equations vii-3 and vii-4 greatly exceeds that for equation vii-2. The fit between admissions and ratings still exceeds the fit between criterion measures and ratings with identical samples.

82 Section VIII shows a similar analysis of the impact of additional ratings on predicting performance. Note that, for the data analyzed here, the predictive value of additional ratings does not imply that the admissions committee only gives additional ratings to "marginal candidates." As Section V shows, typically higher-rated candidates receive more ratings. Section X discusses changes in the admissions process that address this issue.

predicted admissions. This inconsistency has a simple explanation. Since "pseudo" r-square values measure the differences between actual admissions and predicted probabilities, the "pseudo" r-square increases when the predicted probability moves closer to either zero or one, depending on the actual decision. Observations at the tails of the distribution of ratings will have relatively greater weight in determining the "pseudo" r-square. Thus, if an applicant with a relatively high rating receives a rejection, this will decrease the "pseudo" r-square disproportionally. Including additional ratings in the average for that applicant may move the applicants average rating lower, thereby increasing the "pseudo" r-square. However, unless the average rating changes sufficiently, the predicted decision for the applicant will remain the same. Conversely, applicants near the center of the distribution of ratings could more likely have different predicted admissions decisions as a result of including additional ratings in their average rating. This can lead to the divergence between "pseudo" r-square values and the proportion of correct predictions observed in Table VII-5.

Table VII-5
Predictive Value of Additional Ratings

	Raw Ratings		GRE Quantitative Scale (by Judge) Ratings	
Number of Ratings	"Pseudo" R-Square	Proportion of Correct Predictions Using Cut-Off Probability	"Pseudo" R-Square	Proportion of Correct Predictions Using Cut-Off Probability
1	0.23	76%	0.26	78%
2	0.33	83%	0.30	74%
3	0.34	81%	0.35	80%
4	0.36	83%	0.38	83%

Predicting Admissions With Statistical Synthesis

Including ratings with criterion measures in the same regression equation for admissions provides another view of the relationships between criteria, ratings and admissions. As previous equations and tables have indicated, ratings measures appear to have a closer fit with admissions decisions than do criterion measures. However, an additional criterion measure remains significant in the prediction of admissions even after adjusting for average rating. This variable, Rating of Undergraduate Institution, does not have significance in the best-fit equation for average

rating. Equations vii-5 and vii-6 show the best-fit logistic regression equations for admissions using both ratings and criterion measures as independent variables.

	Equation:		vii-5	vii-6
	ln[p/(1-p)]			
=			-13	-24
			(1.2)	(2.2)
+	Average Rating	x	3.5	0.032
			(0.34)	(0.0031)
+	Rating of Undergraduate Institution	x	0.81	0.79
			(0.20)	(0.20)
	"Pseudo" R-Square:		.57	.58

Note: p= probability of admission
Coefficient standard errors in parentheses
Equation vii-5 uses Raw ratings; equation vii-6 uses GRE Quantitative Scale (by Judge) ratings
491 observations: 1 observation deleted from equation vii-5 to match the sample for equation vii-6
"pseudo" r-square calculated as percentage change in log likelihood after estimation, using log likelihood before estimation as a base

The sample sizes for equations vii-5 and vii-6 again greatly exceed the sample size for equation vii-2. However, even using identical samples, Rating of Undergraduate Institution remains significant in addition to Average Rating. The significance of Rating of Undergraduate Institution in addition to Average Rating suggests that committee ratings do not fully reflect the judgments of committee members, even as regards the desirability of candidates for admission. Alternatively, the significance of Average Rating in addition to Rating of Undergraduate Institution implies that ratings do have something unique to contribute to the explanation of admissions decisions.[83] Section X discusses the implications of this divergence between ratings and decision behavior. The remainder of this section shows the results of using linear combinations of the criterion measures as paramorphic ratings in logistic regressions for the admissions decision.

[83] Some might argue that additional reviewers reduce the variation and likely inconsistencies of alternative selection criteria. Note elsewhere, however, that heteroskedacity does not show up with respect to Number of Judges.

Predicting Admissions With Paramorphic Representations

Paramorphic representations of ratings do not appear to provide a better fit than actual ratings. Equations vii-7 and vii-8 show the results of logistic regressions for the admissions decision using predicted average ratings from equations vi-2 and vi-3 respectively. Transformation has little impact on the fit between paramorphic ratings and the admissions decision.

Equation:		vii-7	vii-8
ln[p/(1-p)]			
=		-8.4	-16
		(1.7)	(3.1)
+ Paramorphic Rating	x	3.0	0.024
		(0.58)	(0.0047)
"Pseudo" R-Square:		.30	.31

Note: p = probability of admission
Coefficient standard errors in parentheses
Paramorphic representation of average ratings from equation vi-2 and vi-3
117 observations: 1 observation deleted from equation vii-5 to match the sample for equation vii-6
"pseudo" r-square calculated as percentage change in log likelihood after estimation, using log likelihood before estimation as a base

Although paramorphic representation does not improve the fit between ratings and admissions, this does not mean that ratings provide superior predictive power. The rationale for using paramorphic ratings (see Section IV) suggests only that paramorphic ratings will predict *performance* better than will actual ratings. The next section examines the prediction of performance using criterion measures, ratings and paramorphic representations of ratings.

VIII. GRADES

This section examines the relationship between criterion measures and/or committee ratings and various measures of grades. This research used both standard linear and logistic regression techniques to analyze various measures of course grades as dependent variables. The general assumptions of these models specify similar functional relationships as in equations vi-1 and vii-1a. However, the dependent variable changes from average rating to average grade in linear regressions. Similarly, the dependent variable changes from probability of admission to probability of some other event (such as occurrence of a C grade) for logistic regression equations. Potential independent variables include all the criterion measures coded for this research, as well as committee ratings. As with the prediction of ratings and admissions decisions, the equations assume a first-degree polynomial for the functional form of the right-hand side.

Predicting Grades With Criterion Measures

The best-fit regression equation (viii-1) for overall raw GPA at RGS using criterion measures includes just two variables: GRE Quantitative Score and Age.

viii-1: GPA	=	2.8	
		(0.47)	
	+	0.0016	x GRE Quantitative Score
		(0.00055)	
	+	-0.023	x Age
		(0.011)	

Notes: GPA calculated with raw grades, excluding pass/fail grades
Coefficient standard errors in parentheses
R-square (adjusted) = .09
99 observations

A 100 point increase in GRE Quantitative Score leads to a .16 increase in raw GPA, according to equation viii-1. By the same equation, a 10 year increase in age leads to a .23 decrease in GPA. For transformed GPA, Age has no statistical significance and the product of Undergraduate GPA

and Rating of Undergraduate Institution does have significance (equation viii-2). As equation viii-2 shows, a 100 point increase in GRE Quantitative Score leads to an increase of 40 in average transformed GPA. This 40 point increase represents a much greater magnitude (in relation to the standard deviation) than the .16 increase in average raw GPA indicated by equation viii-1. The r-square values indicate this relationship as well.

viii-2: GPA	= 330	
	(76)	
	+ 0.40	x GRE Quantitative Score
	(0.093)	
	+ 6.7	x Undergraduate GPA x Rating of Undergraduate Institution
	(3.0)	

Notes: GPA calculated with GRE Quantitative scale (by Course) grades, excluding pass/fail grades
Coefficient standard errors in parentheses
R-square (adjusted) = .26
61 observations

Using transformed grades increases the adjusted r-square from .09 to .26, about a three-fold improvement in the proportion of variance explained. Placing these figures in the context of previous research, both lie within the range of results reported in Section IV. Using the square root of the r-squares as multiple correlation coefficients yields "validity" coefficients of .3 and .51 for equations viii-1 and viii-2, respectively. While .51 represents a relatively high value compared with most previous research, it does not fall outside the range of expectation. More imortantly, using transformed grades effects a noticeable increase in the correlation.

Although the sample differs between equations viii-1 and viii-2, the superior predictability of transformed grades remains when equation viii-1 uses the smaller sample.[84] In equation viii-1, the coefficient for Age has a negative sign. In other words, older students tend to perform less well than younger students, at least when calculating GPA with raw grades. Age loses its statistical significance in predicting transformed grade measures.

[84] Indeed, even using predicted values as substitutes for missing values would not change this result.

Equation viii-2 suggests that explanation of the variance in GPA involves at least some non-compensatory combination of criterion measures. The product of Undergraduate GPA and Rating of Undergraduate Institution also has a rather intuitive appeal for the prediction of performance. Both previous grades and the overall quality of previous schools attended have positive correlations with performance in graduate school. The significance of their *product* indicates that the combination of high previous performance *at* high quality schools leads to high performance in graduate studies. Neither high previous grades nor high quality previous schools by itself can explain high graduate GPA. (In other words, these factors do not compensate for each other in the prediction of GPA at RGS.)

For predicting problem indicators, a different criterion measure has significance in addition to GRE Quantitative Score: MS Degree Earned. Equation viii-3 shows the results of this analysis.[85]

viii-3: ln[p/(1-p)]	=	4.2	
		(1.8)	
	+	-0.0069	x GRE Quantitative Score
		0.0017)	
	+	1.0	x MS Degree Earned
		(0.45)	

Notes: p = probability of at least one grade of C or lower, including F grades in pass/fail courses
Coefficient standard errors in parentheses
"Pseudo" r-square = .09
103 observations

The negative sign of the coefficient for GRE Quantitative Score shows that an increase in GRE score will decrease the probability of this problem indicator. A student with GRE Quantitative Score of 650 has a probability of receiving at least on C grade (or lower) of 43%. A student with GRE Score of 750 (holding other factors constant) has a probability of only 27%. Furthermore, the positive sign of the coefficient for MS Degree Earned indicates that students with Masters of

[85] Note that the logistic regression equation with multiple independent variables implies an interactive effect between the independent variables. The algebra of equations vii-1a and vii-1b requires this. Intuitively, logic also indicates the interactive effect: when one or more independent variables cause the logit function to move close to 0 or 1, the impact of any remaining variables must diminish, since the logit function cannot fall outside these bounds.

Science degrees have a *higher* propensity for problems, measured by the incidence of C grades or lower. In contrast, the negative sign of the coefficient for GRE Quantitative Score in equation viii-3 means that students with higher scores have a lower propensity for problems. The same criterion measures comprise the best-fit equation for the prediction of other problem indicators, such as Number of C Grades, Any F Grade and Number of F Grades. In addition, the predictors for these measures remain the same for core and first-year courses.

Predicting Grades With Ratings

Equations viii-4 and viii-5 show the results of regression analysis for overall GPA at RGS using both raw and transformed grades. Using transformed grades increases the adjusted r-square from .06 to .19, again a three-fold improvement in the proportion of variance explained by ratings. Using transformed ratings has little impact on the predictability of either raw or transformed grades.

	Equation:		viii-4	viii-5
	GPA			
=			2.2	49
			(0.35)	(37)
+	Average Rating	x	0.30	59
			(0.10)	(11)
	R-Square (adj.):		.06	.19

Note: Equation viii-4 uses GPA calculated with raw grades, excluding pass/fail grades; equation viii-5 uses GPA calculated with GRE Quantitative scale (by Course) grades, excluding pass/fail grades
Coefficient standard errors in parentheses
121 observations: 1 observation deleted from equation viii-4 to match the sample for equation viii-5
Raw ratings

Based on equation viii-4, an increase of one point (one letter grade) in raw Average Rating leads to an increase in raw GPA of .3, or about one step (plus or minus) in the letter-based scale, on average. C (2.0) rated students have average GPAs of 2.8 (about a B-). B(3.0) rated students have average GPAs of 3.1 (about a B). A (4.0) rated students have average GPAs of 3.4

(about a B+). Using transformed grades (equation viii-5), C rated students have average GRE Quantitative Scale GPAs of 610. For B rated students, the average rises to 670. For A rated students, the average reaches 730. The magnitude of these increases in transformed GPA greatly exceeds the magnitude of the changes in raw GPA.

For the prediction of problem indicators, ratings do not have statistical significance at the 5% level. This dissertation used logistic regression analysis of the indicator for each of the twelve problem indicators listed in Table V-4 using both raw and transformed ratings as the predictor. The lack of statistical significance persists in all cases.

Predictive Value of Additional Ratings on the Margin

For GPA, a comparison of several linear regression analyses can show the predictive value of additional ratings on the margin. Table VIII-1 shows the adjusted r-square values for regression equations for transformed GPA using the average of one to four randomly selected raw ratings. These regression equations all use the same sample, students with at least four ratings. As Table VIII-1 shows, additional ratings increase the proportion of variance explained.

Table VIII-1
Value of Additional Raw Ratings for Predicting Transformed GPA

Number of Ratings	Adjusted R-Square
1	0.08
2	0.11
3	0.18
4	0.21

Notes: Number of observations = 84

Because Average Rating does not have statistical significance in predicting problem indicators, this section does not discuss the value of additional ratings in predicting them.

Predicting Grades With Statistical Synthesis

Equation viii-6 shows the results of combining criterion measures and ratings in a single regression equation for transformed GPA. This statistical synthesis suggests that ratings contribute to the explanation of variance in grades, even after adjusting for criterion measures. For predicting problem indicators, GRE Quantitative Score has statistical significance while ratings do not. (Thus, this dissertation does not show the regression results.) This indicates that

ratings do not contribute to the prediction of problem indicators, adjusting for criterion measures. Consistent with equation viii-3, MS Degree Earned again leads to a *higher* propensity for problems, even after adjusting for GRE Quantitative Score (with or without ratings in the equation).

viii-6: GPA	=	400	
		(57)	
	+	37	x Average Rating
		(12)	
	+	0.35	x GRE Quantitative Score
		(0.064)	
	+	-2.6	x Age
		(1.2)	

Notes:
GPA calculated with GRE Quantitative scale (by Course) grades, excluding pass/fail grades
Raw ratings
Coefficient standard errors in parentheses
R-square (adjusted) = .40
93 observations

Predicting Grades With Paramorphic Representations

Equation viii-7 shows the prediction of transformed GPA using paramorphic representations of Average Rating.

viii-7: GPA	=	520	
		(53)	
	+	57	x Paramorphic Rating
		(17)	

Notes:
GPA calculated with GRE Quantitative scale (by Course) grades, excluding pass/fail grades
Coefficient standard errors in parentheses
R-square (adjusted) = .15
61 observations
Predicted average raw rating from equation vi-2

Equation viii-8 uses a slightly different paramorphic representation to predict GPA. Instead of using predicted ratings, equation viii-8 uses predicted probabilities of admission,

based on equation vii-2. In this case, paramorphic representations predict grades more accurately than actual ratings.[86] (r-square increases from .19 in equation viii-5 to .27 in equation viii-9). This superiority remains when the two equations use identical samples.

viii-8:	GPA	= 590		
		(17)		
		+ 160	x Paramorphic Admissions	
		(27)	Probability	

Notes:
GPA calculated with GRE Quantitative scale (by Course) grades, excluding pass/fail grades
Coefficient standard errors in parentheses
R-square (adjusted) = .27
98 observations
Predicted admissions probability from equation vii-2

Equation viii-9 shows the results of predicting problem indicators with the same paramorphic representation used in equation viii-8. Not surprisingly, paramorphic ratings do not have statistical significance at the 5% level in problem indicators, since neither raw nor transformed measures have significance either.[87]

viii-9:	ln[p/(1-p)]	= 1.5		
		(0.63)		
		+ -2.9	x Paramorphic Admissions	
		(1.0)	Probability	

Notes:
p = probability of at least one grade of C or lower, including F grades in pass/fail courses
Coefficient standard errors in parentheses
"Pseudo" r-square = .06
99 observations
Predicted probability of admission from equation vii-2

Paramorphic Admissions Probability predicts problems more accurately than actual ratings, which do not even have statistical significance. The superiority of Paramorphic Admissions Probability remains when using identical samples.

---

86 Contrary to expectations based on other research, paramorphic representations have less explanatory power than actual ratings. However, equation viii-8 excludes almost half the observations from the sample used for equation viii-5. Using identical samples, paramorphic representations have a slightly higher r-square. Models with predicted values to account for missing data does not change this result.

87 The superiority of paramorphic representations in predicting various performance indicators simply confirms what previous research would suggest. (See Section IV.)

IX. ALTERNATIVE PERFORMANCE MEASURES

This section examines the relationships between selection criteria and/or committee ratings, and several alternative measures of performance:

- attrition
- qualifying exam results, and
- completion time

Since attrition and qualifying exam results have binary measures, this research used logistic regression techniques to analyze them as dependent variables. The general assumptions of these models specify a functional relationship similar to equation vii-1a and several of the equations in the preceding sections. (In this section, the dependent variable changes to probability of dropping out before qualifying examinations.) As before, potential independent variables include all selection criteria coded for this analysis, as well as ratings and paramorphic representations of committee decisions.

This section also considers GPA and problem indicators as independent variables. (In other words, such intermediate performance measures also serve as potential predictors.) For attrition and qualifying exam results, this section presents summary tables rather than detailed regression equations, which would follow the same pattern as in Section VIII. For completion time, this research used a different methodology, which this section explains after reviewing the analysis of attrition and qualifying exam results.

Predicting Attrition

Table IX-1 summarizes the results of logistic regressions for the probability that a student will leave the RGS program before attempting the qualifying examinations. (Normally, students take their qualifying examinations two years after entering RGS.) The best-fit logistic regression equation for attrition includes just one criterion measure: GRE Quantitative Score. Students with higher test scores have lower propensity to drop out, on average. Students with GRE Quantitative Scores of 500 have a 33% chance of attrition, on average. Students with scores of 600 have a 23% chance, and students with scores of 700 have a 15% probability of attrition, on

average. Even students with scores of 800 have a 10% chance of dropping out. (The predicted probabilities for lower GRE scores may not mean much, since they fall outside the range of scores for admitted students.)

Table IX-1
Predicted Probability of Attrition

GRE Quantitative Score	Probability of Attrition
200	71%
300	59%
400	46%
500	33%
600	23%
700	15%
800	10%

Notes: Only GRE Quantitative Score has significance in multiple regressions using selection criteria and ratings. Paramorphic admissions probability also has significance.

Neither raw nor transformed ratings have statistical significance in predicting attrition. Using raw ratings, more highly rated students have a slightly *higher* propensity for attrition, although this result could happen from random chance alone. At least the sign for the coefficient makes intuitive sense for transformed ratings. Because ratings do not have statistical significance in the prediction of attrition, this section does not consider the predictive value of additional ratings on the margin.

Not surprisingly, ratings do not have statistical significance when used in a statistical synthesis to predict attrition. Following the pattern revealed in predicting problem indicators, paramorphic representations of ratings do not have statistical significance in predicting attrition, but paramorphic representations of admissions probabilities do. Students with higher predicted ratings have lower propensity to drop out before qualifying exams, although (again) this result could happen from random chance alone. Using Paramorphic Admission Probability, the pseudo r-square exceeds the pseudo r-square using GRE Quantitative Score. (However, the predicted probabilities from equation vii-2 incorporate more than GRE Quantitative Score. Equation vii-2 included Undergraduate GPA as well. Thus, the paramorphic representation of admissions

probability effectively adds an additional explanatory variable that, considered separately, does not have statistical significance in the prediction of attrition.)

Cross-tabulations of predicted admissions with attrition indicate no significant difference in attrition rates when using predictions based only on ratings. Table IX-2 shows the results of such cross-tabulations.

Table IX-2
Probability of Attrition Using Paramorphic Admissions Decisions

Basis of Predicted Admission	Admitted	Rejected
Raw Criteria	12%	22%
GRE Quantitative Scale (by Cohort) Criteria	14%	23%

Notes: * Differences not statistically significant at the 5% level for ratings or statistical syntheses.

Equation ix-1 shows the results of using raw GPA as the predictor for attrition. Based on this equation, C (2.0) students, have, on average, an 87% probability of attrition, while B (3.0) students have a 23% chance, and A (4.0) students have just a 1% chance, on average. The pseudo r-square using GRE Quantitative scale GPA falls to .18, suggesting that raw grades have greater predictive value in the case of attrition than do transformed grades.

ix-1: $\ln[p/(1-p)] = \underset{(1.5)}{8.1} + \underset{(0.50)}{-3.1} \times GPA$

Notes:
p = probability of dropping out before qualifying examinations
Coefficient standard errors in parentheses
"Pseudo" r-square = .43
159 observations
Comprehensive raw GPA

Problem indicators also predict attrition rather well. Table IX-3 shows cross-tabulations of various problem indicators with probability of attrition. In each case, students with problems in their grades have significantly higher propensity to drop out of RGS. First-year problems identify likely dropouts better than overall problems do, and F grades predict attrition better than

C grades do. Two thirds of students with at least one F grade in their first year at RGS leave the program before taking their qualifying examinations. Of course, grades reflect, in part, an instructor's evaluation of how well a student has mastered subject material, which will inevitably confront the student during qualifying examinations.

Table IX-3
Probability of Attrition Using Problem Indicators

Problem Indicator	Problem	No Problem
Any C Grade	39%	15%
First Year C Grade	45%	15%
Any F Grade	55%	15%
First Year F Grade	67%	17%

GPA and problem indicators predict attrition more accurately than either selection criteria or ratings, regardless of transformations. This suggests an underlying problem in performance prediction. As performance measures fall further away in time from their predictors, they tend to lose predictability. Section X discusses the implications of this phenomenon.

Predicting Qualifying Exam Results

The measurement of qualifying exam results must determine how to treat attrition. Students who drop out before qualifying examinations neither fail nor pass. As previous discussions have shown, attrition has little correlation with course grades. This dissertation treats attrition as equivalent to failing the qualifying exam. (The two events have identical effects on completion of the degree requirements.) This ignores the consideration that a student who drops before taking the exams may indeed have passed. Indeed, the data show that not all dropouts have lower than average grades. However, regression results for qualifying exam results do not change substantially when excluding students who drop out from the analysis. This section includes them to minimize missing data problems.

Tables IX-4 and IX-5 summarize the results of logistic regressions for two measures of qualifying exam results: passage at some time and distinction in at least one subject. The underlying regressions show that students entering RGS at age 40 have a 50% chance of passing their qualifying exams. Students entering at age 30 have a 70% chance of passing. Students

entering at age 20 have an 84% chance of passing. Obviously, this equation has absurd implications for ages outside the normal range of entering students. Newborns do not likely have, on average, a 96% chance of passing their qualifying exams at RGS!

Table IX-4
Predicting Passage of Qualifying Exams

Independent Variables	"Pseudo" R-Square	Sample Size
Age	0.04	158
Average Raw Rating *	0.00	123
Average GRE Quantitative Scale Rating *	0.01	123
Statistical Synthesis (GRE Quantitative Scale Rating* and Age)	0.05	118
Paramorphic Raw Rating *	0.00	60

Notes: * not statistically significant at the 5% level
Only Age has significance in multiple regressions using selection criteria

Table IX-5
Predicting Distinction on Qualifying Exams

Independent Variables	"Pseudo" R-Square	Sample Size
Undergraduate GPA, Rating of Undergraduate Institution, and Age	0.35	98
Average Raw Rating	0.17	123
Average GRE Quantitative Scale Rating	0.19	123
Statistical Synthesis (GRE Quantitative Scale Rating, Undergraduate GPA, Rating of Undergraduate Institution, and Age*)	0.40	71
Paramorphic Raw Rating	0.11	60

Notes: * Age not statistically significant at the 5% level

Using dummy variables can overcome this problem. However, careful use of the results to make inferences only within reasonable age ranges also prevents this. Using dummy variables would also imply some interactive effect between different age categories (see the discussion in the previous section). Furthermore, using dummy variables in the logistic regression loses important continuity. This analysis also examined quadratic terms using Age and other methods to account for the intuitive notion that, at some point, younger students will have more difficulty with the RGS program. (The z-score transformation squared would equate the young with the

old.) However, such analyses did not yield further insights into the impact of Age on qualifying exam performance.

Tables IX-4 and IX-5 show that distinction has greater predictability from selection criteria than does passage, indicated by the pseudo r-square of .35, compared to the pseudo r-square of .04 in the prediction of passage. (These results do not change appreciably when using identical samples.) Using the underlying logistic regression coefficients shows a much greater impact of age, for example, in predicting distinction. Students entering RGS at age 30 have a 70% chance of distinction on their qualifying exams, assuming Undergraduate GPA and Rating of Undergraduate Institution both equal 3.9. (This equates the probability of distinction to the probability of passage from Table IX-4. It does not imply that such levels represent typical values.) Students entering at age 40 with the same levels of the other predictors have a 23% chance of distinction. Students entering at age 20 with those other predictor levels have a 94% chance of distinction. Thus, the same 10-year increment in Age changes the probability of distinction by a greater amount than the change in probability of passage.

Using ratings as predictors of qualifying exam results reveals the same pattern for comparative predictability: distinction has higher predictability than passage. In fact, Average Rating does not have statistical significance in the prediction of passage, regardless of transformation. (The p-value for transformed ratings comes much closer to 5%, but the standard error of the coefficient for Average Rating remains greater in magnitude than the coefficient itself.) The underlying logistic regressions show that students with Average Rating of 2.0 (C) have about a .2% probability of receiving distinction; students with Average Rating of 3.0 (B) have about a 4% chance; and students with Average Rating of 4.0 (A) have about a 40% chance. Consistent with previous research, the fit between criterion measures and qualifying exam results exceeds the fit using ratings. (This holds even when using identical samples.)

Table IX-6 shows the pseudo r-squares for predicting distinction using both raw and transformed ratings. Rating by an additional judge increases the fit between average ratings and qualifying exam results in both cases. Generally, the fit for transformed ratings exceeds the fit using raw ratings. Sampling anomalies probably account for deviations from this pattern.

Table IX-6
Value of Additional Ratings for Predicting Any Distinction

	"Pseudo" R-Square	
Number of Ratings	Raw Ratings	Transformed Ratings
1	0.10	0.17
2	0.13	0.17
3	0.19	0.18
4	0.19	0.23

Including both ratings and criterion measures in the same logistic regression equation to predict qualifying exam results has varying impact depending on the result measured. In the prediction of passage, Average Rating has no statistical significance at the 5% level, while Age retains its significance. This implies that committee ratings do not explain any of the variance in qualifying examination performance, after adjusting for criterion measures. However, in the best-fit prediction for distinction, Average Rating does have statistical significance at the 5% level, while Age loses its significance. This suggests that ratings do contribute to the explanation of variance in qualifying exam results, after adjusting for criterion measures. Clearly, the choice of performance measure has a great impact on the results of statistical synthesis.

The relative predictability of passage and distinction holds when using paramorphic representations of ratings as predictors. Using paramorphic ratings to predict passage has no statistical significance. On the other hand, paramorphic ratings have statistical significance in the prediction of distinction.

Table IX-7 shows cross-tabulations of predicted admissions decisions with probability of passing the qualifying exams. Only predictions using criterion measures alone show significant differences between students with predicted admission and students with predicted rejection. Table IX-8 shows similar cross-tabulations of predicted admissions decisions with probability of distinction on the qualifying exams. In this case, students with predicted admissions have significantly higher probability of success than students with predicted rejection, regardless of whether the prediction uses ratings, criterion measures or both. Transformation also makes little difference in these relationships. Furthermore, the differences narrow when the prediction uses only criterion measures.

Table IX-7
Predicting Passage With Paramorphic Admissions Decisions

	Probability of Passage at Some Time	
Basis of Predicted Admission	Admitted	Rejected
Raw Ratings*	0.75	0.75
GRE Quantitative Scale (by Judge) Ratings*	0.75	0.75
Raw Criteria	0.84	0.78
GRE Quantitative Scale (by Cohort) Criteria	0.82	0.73
Raw Ratings and Criteria*	0.75	0.76
Transformed Ratings and Criteria*	0.75	0.79

Notes: * Differences not statistically significant at the 5% level

Table IX-8
Predicting Distinction With Paramorphic Admissions Decisions

	Probability of Any Distinction	
Basis of Predicted Admission	Admitted	Rejected
Raw Ratings	0.21	0.00
GRE Quantitative Scale (by Judge) Ratings	0.21	0.00
Raw Criteria	0.24	0.11
GRE Quantitative Scale (by Cohort) Criteria	0.21	0.15
Raw Ratings and Criteria	0.22	0.00
Transformed Ratings and Criteria	0.21	0.00

Equations ix-2 and ix-3 show the results of using raw GPA to predict passage and distinction. C (2.0) students have, on average, a 6% probability of passing their qualifying exams at some time. (This treats dropping out as equivalent to failing.) B (3.0) students have a 71% chance of passing, and A (4.0) students have a 99% chance of passing, on average. When using GRE Quantitative scale GPA to predict passage, the pseudo r-square in equation ix-2 falls to .18, again indicating some relevance for absolute standards in grading.

According to equation ix-3, C (2.0) students have, on average, a probability of earning distinction on their qualifying exams that falls below one tenth of one percent. (This treats dropping out as equivalent to not earning distinction.) B (3.0) students have about a 2% chance of distinction, and A (4.0) students have a 50% chance of distinction, on average.

Equation:		ix-2 (Passage)	ix-3 (Distinction)
ln[p/(1p)]			
=		-9.9	-16
		(1.7)	(3.9)
+ GPA	x	3.6	4.0
		(0.57)	(1.1)
"Pseudo" R-Square:		0.46	0.23

Note: Coefficient standard errors in parentheses
Raw GPA
159 observations
"Pseudo" r-square calculated as percentage change in log likelihood after estimation, using log likelihood before estimation as a base

Table IX-9 presents cross-tabulations of various problem indicators with qualifying exam results. In each case, students with problems in their grades have significantly lower probability of both passage and distinction.

Table IX-9
Predicting Qualifying Exam Results With Problem Indicators

	Probability of Passage at Some Time		Probability of Any Distinction	
Problem Indicator	No Problem	Problem	No Problem	Problem
Any C Grade	85%	55%	22%	8%
Any Core C Grade	87%	42%	21%	5%
Any F Grade	83%	38%	17%	8%
Any Core F Grade	80%	32%	16%	8%

Reflecting earlier patterns, GPA and problem indicators generally predict qualifying exam results more accurately than either selection criteria or ratings, regardless of transformations. As these performance measures fall still further away in time from selection criteria and ratings, these predictors tend to lose accuracy (at least relative to measures closer in time, such as GPA and problem indicators Curiously, however, passage seems to have greater predictability than distinction, using these prior performance measures. This suggests that grades

reflect somewhat more closely an instructor's assessment of a student's prospects for passing. Perhaps additional factors intervene in determining who receives distinction. (However, this analysis could not find statistical significance for any such additional factors.) Section X discusses the implications of this phenomenon.

Predicting Completion Time

This dissertation used survival analysis techniques to analyze completion time. These techniques, also known as proportional hazard models, assume that the instantaneous probability of some event has an exponential relationship to some set of independent variables. The proportional hazards models specify the following functional form in equation ix-4a. As with logistic regression, transformation of equation ix-4a provides a more convenient form for interpreting the results of proportional hazards analyses.

$$\text{ix-4a: } h(t) = \frac{p(\text{event occurs between time } t \text{ and } t+d)}{d \times p(\text{event occurs after time } t)} = h_0(t) \times \exp[f(X_1, X_2, \ldots X_n)]$$

$$\text{ix-4b: } \ln[h(t)/h_0(t)] = b_1 \times X_1 + b_2 \times X_2 + \ldots + b_n \times X_n$$

where

p		denotes probability of completion;
d		denotes an increment of time;
$X_1, X_2, \ldots X_n$		represents some combination of the criteria coded for the analysis and shown in Table V-1;
$f(X_1, X_2, \ldots X_n)$	=	$b_1 x X_1 + b_2 x X_2 + \ldots + b_n x X_n$;
$h_0(t)$		represents a baseline hazard function, not requiring estimation; and
exp		denotes exponentiation of e, the base of natural logarithms
ln[ ]		denotes the natural logarithm

Moving the baseline hazard function in equation ix-4a to the left-hand side and taking the natural logarithm of both sides yields equation ix-4b. Statisticians refer to the left-hand side of equation ix-4b as the relative risk, or hazard ratio. This type of analysis does not estimate a

constant term for the linear combination of independent variables. In effect, the baseline hazard function makes any constant term arbitrary.

Only one criterion measure, MS Degree Earned, has statistical significance in predicting the relative risk for completion time. Consistent with the prediction of problem indicators, the negative coefficient of MS Degree Earned in equation ix-5 means that students with MS degrees experience more delay in completing the RGS program than students without such degrees.

ix-5:	$\ln[h(t)/h_0(t)]$	= -0.61	x	MS Degree Earned
		(0.28)		
Notes:	$h(t)/h_0(t)$ = hazard ratio for completion Coefficient standard errors in parentheses 117 observations "Pseudo" R-Square = .01			

The coefficient of MS Degree Earned indicates that students with Master of Science degrees have an instantaneous probability of completion equal to 54% (exponentiation of -.61) of the hazard for students without such degrees, on average. While this may seem like a dramatic impact on completion time, the pseudo r-square indicates that, overall, MS Degree Earned explains only about 1% of the observed variance in time to completion. Moreover, GRE scores have no statistical significance in the prediction of the relative risk, nor does Number of Judges.

Equations ix-6 and ix-7 show the results of using average raw and transformed ratings as the independent variable in the proportional hazards model for completion time. The pseudo r-square exceeds the pseudo r-square for equation ix-5 in both cases, although it remains low, at about .02.[88] The coefficient for Average Rating in equation ix-6 means that, on average, a one point increase in average rating will increase the instantaneous probability of completion by a multiplicative factor of about 2.1 (exponentiation of .71). Thus, students with one letter higher ratings, on average, have twice the hazard of completion at any given time. Equation ix-7 offers a similar interpretation of transformed ratings. On average, students with 100 point higher ratings, on average have about twice the hazard of completion at any given time. In both cases, the proportion of variance explained by ratings remains relatively low, however.

[88] The pseudo r-square remains higher for ratings than for criterion measures when these equations use identical samples

Equation:		ix-6	ix-7
$\ln[h(t)/h_0(t)]$			
= Average Rating	X	0.72	0.0069
		(0.33)	(0.0026)
"Pseudo" R-Square:		.02	.02

Note: $h(t)/h_0(t)$ = hazard ratio for completion
Coefficient standard errors in parentheses
96 observations
Equation ix-6 uses raw ratings; equation ix-7 uses GRE Quantitative Scale (by Judge) ratings

This analysis performed Cox regression analyses using the average of 1 to 4 randomly selected ratings as predictors of the hazard ratio for completion time. None of the ratings measures has statistical significance at the 5% level for the sample of students with at least four ratings. Clearly, sampling has a major impact on the statistical significance of ratings, since average ratings, overall, have statistical significance (equations ix-6 and ix-7). In fact, using just one randomly selected rating as a substitute for the independent variables in equations ix-6 and ix-7 has no statistical significance at the 5% level.

The best-fit equation combining ratings and criterion measures in the proportional hazards model for completion includes Average Rating and Rating of Undergraduate Institution.

ix-8: $\ln[h(t)/h_0(t)]$	=	0.0058	x Average Rating
		(0.0026)	
	+	0.35*	x Rating of Undergraduate Institution
		(0.24)	

Notes: $h(t)/h_0(t)$ = hazard ratio for completion
Coefficient standard errors in parentheses
92 observations
"Pseudo" R-Squate = .03
GRE Quantitative scale (by Judge) ratings
* Not statistically significant at the 5% level

However, as equation ix-8 shows, Rating of Undergraduate Institution lacks statistical significance at the 5% level. (MS Degree Earned and other criterion measures have even lower

p-values than Rating of Undergraduate Institution when combined with Average Rating as independent variables.) This indicates that committee ratings have predictive value for completion time, after adjusting for criterion measures. Equation ix-9 shows the relationship between paramorphic ratings and completion time.

ix-9:	$\ln[h(t)/h_0(t)]$	=	1.1*	x Paramorphic Rating
			(0.72)	

Notes:
$h(t)/h_0(t)$ = hazard ratio for completion
Coefficient standard errors in parentheses
52 observations
"Pseudo" R-Squate = .03
Predicted average raw rating from equation vi-2
* Not statistically significant at the 5% level

Although the pseudo r-square appears superior to the fit using actual ratings, it does not have statistical significance at the 5% level. However, this does not invalidate the proposed superiority of statistical prediction. First, the fit with paramorphic ratings still exceeds the fit with actual ratings. Second, when equations ix-6 and ix-7 use samples with non-missing paramorphic representations, they also lack statistical significance at the 5% level. Once again, sampling has a major impact on the results of analyzing RGS data. Within the smaller samples, paramorphic ratings retain their superior predictive value.

Equation ix-10 shows the results of using a paramorphic representation of admissions decisions as the predictor for completion time. This equation indicates that, on average, students with a predicted admission have three times (exponentiation of 1.1) the instantaneous probability of completion at any time as do students with predicted rejection.

ix-10:	$\ln[h(t)/h_0(t)]$	=	1.1	x Paramorphic Admissions
			(0.52)	Decision

Notes:
$h(t)/h_0(t)$ = hazard ratio for completion
Coefficient standard errors in parentheses
96 observations
"Pseudo" R-Squate = .02
Predicted probability of admission from equation vii-2

Other performance measures also lack much predictive power in the analysis of time to completion. Equation ix-11 shows the results of using raw GPA as the independent variable in

the proportional hazards model for completion time. Consistent with other analyses in this dissertation, the pseudo r-square surpasses the pseudo r-square for Average Rating. However, predictability for completion time remains very low, with a pseudo r-square of just .03.

ix-11:	$\ln[h(t)/h_0(t)]$	=	1.5	x GPA
			(0.43)	

Notes:
- $h(t)/h_0(t)$ = hazard ratio for completion
- Coefficient standard errors in parentheses
- 117 observations
- "Pseudo" R-Squate = .03
- GPA calculated with raw grades over all courses, excluding pass/fail grades

The coefficient for GPA in equation ix-11, 1.5, indicates that a one point increase in raw GPA increases the instantaneous probability of completion by a multiplicative factor of about 4.5 (exponentiation of 1.5). Actual raw GPA ranges from 1.75 to 4.0 in this regression sample. This corresponds to a total difference in hazard rate of a factor of 29, between students with the lowest GPA and students with the highest GPA, on average. Again, however, the magnitude of this increase does not imply better fit.

Table IX-10
Predicting Completion Time With Problem Indicators

Problem Indicator	Relative Risk	Pseudo R-Square
Any C Grade	37%	0.03
Core C Grade	32%	0.03
Any F Grade	39%	0.02
Core F Grade	43%	0.01

Table IX-10 summarizes the results of predicting completion time with various problem indicators. The relative risk for Any C Grade indicates that the instantaneous probability of completion for a student with at least one grade of C or lower falls to 37% of the probability for a similar student with no C grades. A student with at lease one grade of C or lower in a core course has 32% of the hazard level for a student with no such problems (all other factors held constant.) Indicators for C grades have more predictive value than indicators for F grades. In fact, Core F Grade does not have statistical significance at the 5% level.

As this section demonstrates, completion time has lower predictability than other performance measures.[89] This may reflect the complexity of the constructs required to estimate the relationship between predictors and completion time. It may also reflect an attenuation effect of time itself. Even the fastest finishers have taken two years to complete the RGS degree requirements. Just as overall grades have slightly lower predictability than first-year grades, events further removed in time may have lower *inherent* predictability. The concluding section of this dissertation considers the varying predictability of performance dimensions in the broader context of the issues explored in the preceding sections.

[89] Perhaps OJT emphasis affects completion time in a significant way. Anecdotal evidence suggests that health policy and national security dissertations appear to have fewer funding difficulties. However, the data for this analysis had limited information on OJT areas. The admissions committee did not generally have an indicator of project emphasis. Many students change OJT emphasis over time. While the survival analysis could include OJT indicators for national security and health policy, it led to vast missing data problems for those observations without a completed dissertation. Such limited information did not lead to the discovery of any statistical significance for OJT indicators in predicting time to completion. Perhaps further research could uncover something along these lines, with both more observations in general and the possibility of developing a more reliable indicator of OJT focus.

X. CONCLUSIONS

Institutional and Selection Objectives

Empirical research such as that just presented often provokes one of two reactions. Some people express disappointment in the low correlations between selection criteria (or ratings) and performance measures. Others express a range of emotions from smug satisfaction with those low correlations to hostility toward the concept of quantifying any of the attributes related to academic selection and performance. These reactions often have obvious motives. Their relevance here lies in a less obvious, but fundamental source: disagreement and confusion regarding institutional and selection objectives.

Policy analysis as a rule avoids questioning objectives. Analysts focus instead on the impact that various actions have on the achievement of those objectives. In the case of admissions policy and academic performance measurement, studying the objectives can yield insights into the empirical analysis. This dissertation does not attempt to identify any objectives as the most appropriate for any institution. Rather, the complexity and uncertainty of objectives underscores the variation for empirical analysis to explain. Translation of institutional objectives into selection objectives and narrower selection criteria introduces still more potential error variance. The process of combining individual objectives and criteria through persuasion and consensus building into a single admissions committee decision might suggest that statistical significance in any empirical analysis represents a miracle.

Of course, consideration of what institutions do belies the miracle apparent in such statistical significance. The persuasion and consensus building undoubtedly tend to attenuate correlations. Admissions committee ratings may not fully reflect decision makers' true judgments about the desirability of each applicant. Such judgments also reflect more than mere predictions of performance. But differences among committee members do not imply fundamental opposites. Institutions tend to bring people together on those points where they agree. Thus, members of admissions committees will not likely hold diametrically opposing views of the meaning and purpose of public policy analysis. They may even share many beliefs

about what a graduate program in policy studies should accomplish. Therefore, the formation of objectives and criteria, and ultimately the process of making decisions will not reflect pure randomness any more than pure determinism.

The statistical relationships presented in this report show a balance between the two. The attempt to find correlations among selection criteria, ratings and performance measures does not fully resolve the debates that led to this dissertation (see Preface). The results do suggest that ratings embody a complex objective that includes but also goes beyond predicting performance. Understanding the complexity and uncertainty of objectives at least allows interested parties to see why empirical analysis cannot provide definitive guidance for admissions policy. This understanding also reinforces the insights that empirical analysis can provide.

Selection Criteria, Ratings and Performance

GRE Quantitative score dominates other criterion measures as a predictor of various outcomes at RGS. However, this dominance does not extend over all performance measures. Other criterion measures have significance in predicting particular performance dimensions. Sometimes this significance exceeds that of GRE Quantitative score. Table X-1 summarizes the importance of criterion measures in predicting each of the outcome variables considered in this research. This table identifies the single most important criterion measure in each case by comparing partial correlations within each regression sample. Other important measures include all variables with statistical significance at the 5% level. For each outcome measure considered, GRE Verbal Score has no statistical significance after adjusting for GRE Quantitative Score.

The RGS admissions committee obviously considers a number of factors in judging applicants. Two facts highlight the complexity of admissions committee ratings.[90] First, objective criterion measures explain about half the variance in average committee ratings (see Section VI). Only GRE Quantitative score and Undergraduate GPA have statistical significance in predicting committee ratings. While number of judges also has significance in regression equations for ratings, this has no apparent bearing on admissions decisions or subsequent performance measures.

---

[90] Subsequent discussions highlight additional implications of these emprical findings: see "The Role of Judgment, below.

Table X-1
Importance of Criterion Measures in Predicting RGS Outcome Variables

Outcome Variable	Most Important Criterion Measure	Other Important Measures
Average Rating	GRE Quantitative Score	Undergraduate GPA Number of Judges
Admission	GRE Quantitative Score	Undergraduate GPA
GPA	GRE Quantitative Score	Age Rating of Undergraduate Institution
Any C Grade	GRE Quantitative Score	MS Degree Earned
Attrition	GRE Quantitative Score	
Passage of Qualifying Exams	Age	
Distinction on Qualifying Exams	Undergraduate GPA	Rating of Undergraduate Institution Age
Time to Completion	MS Degree Earned	

Although other factors have no statistical significance, clearly ratings reflect more than just an evaluation of a few quantitative measures of an applicant. Second, ratings provide only a partial explanation of the variance in admissions decisions. Ranking applicants in order of average rating can match actual admissions decisions for about 90% of applicants. (This rule still requires determination of the number of applicants to admit in each cohort. In any case, ratings clearly do not equate exactly to an applicant's ultimate desirability.)

Different performance measures have varying levels of predictability, using either criterion measures or ratings. Table X-2 summarizes the r-square values for the best-fit equations to predict each of the performance measures considered in Sections VIII and IX. Table X-2 shows that GPA and distinction on qualifying exams have much higher predictability than other measures. This reinforces the importance of considering the multiple dimensions of the performance concept. Furthermore, these results show that, in general, outcomes further removed in time have lower predictability, regardless of the predictors used. (Distinction on qualifying exams seems to represent the exception to this rule.) This may help to explain why most validation studies focus on GPA, and why many narrow the focus still further to first-year GPA.

Table X-2
Predictability of Various RGS Performance Measures

	R-Square	
Performance Measure	Criterion Measure(s)	Average Rating
GPA	0.26	0.19
Any C Grade	0.09	0.02
Attrition	0.04	0.00
Passage of Qualifying Exams	0.04	0.00
Distinction on Qualifying Exams	0.35	0.17
Time to Completion	0.01	0.02

Despite their complexity, admissions committee ratings have predictive value for RGS outcomes, even after adjusting for criterion measures. The statistical syntheses presented in Sections VIII and IX demonstrate that ratings have significance in the prediction of GPA, distinction on qualifying exams and time to completion. Indeed, for time to completion, ratings have higher correlations with performance than do any or all criterion measures. The relatively high correlation between ratings and distinction on qualifying examinations indicates that the RGS admissions committee may attach relative importance to the selection objective of choosing applicants to maximize probability of a *great success*.[91] Furthermore, the relatively low correlation between ratings and attrition implies that the committee attaches less importance, at least implicitly, to the selection objective of choosing applicants to minimize the probability of failure.

Of course, neither ratings nor criteria explain much of the variance in time to completion. The most predictable outcomes still have more than half their variance unexplained after the most exhaustive regression analyses. While criterion measures generally predict outcomes better, the statistical syntheses suggest that human judges do add value in predicting performance at RGS.

Including the rating of an additional judge in Average Rating increases the explained variance in admissions decisions and several performance measures, at least for the first few Judges. This reinforces the assertion that judgment has some predictive value. It also

[91] However, the preceding discussion of differential predictability may indicate that this outcome simply has greater inherent predictability, for whatever reason.

emphasizes the reliability problem with respect to committee ratings. When individual measures have imperfect reliability, their average improves overall reliability.

Transformations

Transforming admissions committee ratings increases the correlation between average rating and the admissions decision. This reflects a fundamental characteristic of the admissions process. RGS chooses to limit its matriculating cohort size to about 15 students per year. It can influence somewhat the number of applicants with advertising, personal contacts at other institutions and other recruiting tactics. It has even less control over the general quality of the applicant pool. Variance in the number and quality of applicants results in differing admissions rates by cohort (see Section VII). These differing admissions rates virtually require a relative rating of applicants. For example, for a large pool of applicants each with Average Rating above 3.5, the admissions committee would need to reject some very qualified applicants in order to keep a manageable cohort size. Of course, the committee could choose to ignore cohort size completely. This might still lead to varying admissions rates stemming from the quality of applicant pools.

The committee appears to strike a balance between cohort size and applicant pool quality (and perhaps other factors). Admissions rates vary by cohort, but so do cohort sizes. Thus, admissions decisions seem to involve a combination of relative and absolute standards of evaluation. In fact, the transformation of ratings that best explains admissions decisions also reflects a blend of relative and absolute standards. This transformation used applicant pools for each judge as sub-samples, with the mean and standard deviation of GRE Quantitative score for each judge's applicant pool. GRE score represents an absolute standard of reference. The transformed ratings still also represent relative comparisons within each judge's pool. The explanation of variance in admissions decisions increases when it accounts for the number of applicants in each cohort. However, the GRE based transformation has greater explanatory power than a simple z-score transformation or the Elliott-Strenta transformation. This reinforces the balance between relative and absolute standards.

This balance also surfaces in the explanation of variance in GPA. Transforming grades increases their predictability, whether using criterion measures or ratings. In this case, however, additionally transforming ratings does not improve the fit by much (see Section VIII). However, it seems clear that performance measurement also involves a blend of relative and absolute standards of evaluation. The greater impact of transforming grades suggests that grades have more consistency and reliability problems than ratings. A more thorough assessment would note that grades in different courses (particularly in different departments) measure different constructs. Nevertheless, overall GPA has greater predictability than departmental GPA. These narrower GPA measures, in turn, have greater predictability than individual course grades. The impact of transformations makes it clear that instructors evaluate students partly by comparing them to their peers.

The Role of Judgment

This dissertation does not contradict the consensus of previous research that statistical predictions have greater accuracy than clinical predictions. However, as statistical syntheses demonstrate, committee ratings do have predictive value for several outcomes, even after adjusting for criterion measures. In other words, judgment seems to have value. Even if ratings did not have statistical significance, replacing them with a formula would not help admissions decision makers understand their problems much better.

This research has shown several difficulties in establishing screening rules. The divergence between ratings and admissions decisions suggests a subtlety that a screening formula will fail to capture. Namely, the students with the highest predicted ratings may not represent the ideal matriculating cohort. On the margin, at least, considerations such as the happy bottom quartile may matter more to the admissions committee than predicted performance or even overall desirability of a single candidate. Actual decisions result from unspecified but important differences in the influence wielded by individual committee members. Even a simple average rating combines more than individuals' predictions about an applicant. Furthermore, divergent performance measures would require even a narrow predictive screening rule to choose which performance dimension to maximize.

Such difficulties demonstrate that searching for optimal screening rules may not even address the most appropriate question. Obviously, the empirical results presented in this report and elsewhere could form the basis for prospective admissions decisions. But they cannot determine institutional and selection objectives. Moreover, if admissions decision makers wanted to focus on narrow, predictive selection criteria, their ratings would almost certainly show a better fit with criterion measures. (Some argue that the lack of fit reflects arrogance or ignorance of admissions committee members who insist that they can predict performance better than objective criteria can.) Rather than suggesting a screening rule, which replaces human judgment, this dissertation recommends several actions for admissions committees to facilitate that judgment. Other admissions committees and school administrators might also consider similar changes

Policy Recommendations

Clarify Objectives

If the RGS admissions committee wishes to understand which factors to emphasize in selection, it should explicitly articulate the nature of and relationships among its institutional objectives, selection objectives and selection criteria. For example, the preceding discussion of screening rules suggests that emphasizing a particular performance measure would highlight not only a specific set of statistically significant selection criteria, but also the appropriate weights to give each. Whether or not the committee chose to employ a screening rule, it would still have a greater understanding of how selection criteria relate to its specific objective.

This does not imply that the committee needs to choose a single objective. Nor does it imply that the set of chosen objectives should pertain only to the performance measures considered in this analysis. The committee might decide to make additional explicit performance measures, such as OJT performance. It might also decide to include one or more explicit measures of institutional well-being. Section II reviewed the range of issues regarding objectives, but this research avoided making value judgments about them. Such judgments properly constitute a fundamental task for decision makers. This will not resolve existing differences or disagreements about particular applicants. (In fact, it may increase the length of

debate. The Dean might wish to set some limits on the extent of discussions.) But if committee members will specify their objectives, they will probably better understand the reasons for their disagreements regarding specific candidates. Furthermore, they can discuss those disagreements with more information about the implications for specific performance measures corresponding to various objectives.

Committee members can further illuminate their differences (as well as similarities) by assigning priorities to each objective[92]. Obviously, most decision makers have sufficient knowledge to "game" a system of fixed weights, so such weights should best serve as general guidelines, not hard and fast rules. In addition to weights, the committee would identify specific selection criteria that reflect each objective, as part of this clarification process. Lastly, the committee could also clarify the outcomes that it seeks to optimize, again prioritizing them and specifying how to measure them. This would make it possible to enhance evaluation policies. Before discussing evaluation guidelines, this section includes a somewhat obvious but practical recommendation to optimize committee resource use.

Rationalize Decision Processes

The analysis presented in this report supports a number of general recommendations to rationalize decision processes at RGS. In fact, the RGS admissions committee has already changed its processes significantly from the description in Section III. The following discussion first outlines the general recommendations and then illustrates how the changed RGS process embodies them.

If RGS wishes to simplify its admissions process and elucidate the contribution of additional ratings on the margin, then it should assign committee members to evaluate applicants so as to maximize the reliability of average ratings. This would require that each applicant receive one rating before any receive two. For the same total number of ratings[93] represented by the data analyzed for this dissertation, each applicant could have received three ratings. RGS could further refine this policy by selectively assigning candidates to receive additional ratings

---

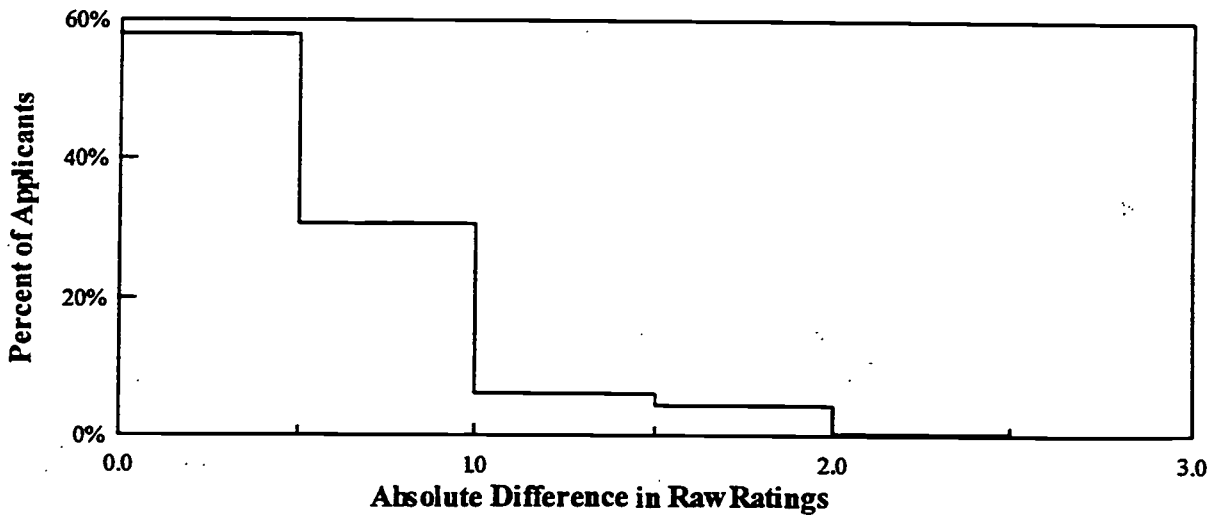
92 This assumes that the admissions committee articulates multiple objectives.

93 Since additional ratings of an individual may require diminishing work on the margin, this analysis cannot equate the reduction in number of ratings with a reduction in total work.

when the absolute difference in the first two ratings exceeds some threshold. As Figure X-1 shows, most pairs of ratings lie within less than one half point on the traditional four-point scale used in this analysis. In fact, 25% of such pairs receive *identical* ratings.

If RGS set the threshold difference at one half point, approximately 42% of applicants would receive a third rating. The total number of ratings would fall by about 20%, compared with the total number of ratings for the applicant pool analyzed for this dissertation. If RGS set the threshold at zero, the total number of ratings would fall by about 10%, and 75% of applicants would receive three ratings. The admissions committee could use the resulting time savings to increase the time allocated for discussion of applicants during committee meetings. These rough estimates assume that all ratings take an equal amount of time. Intuitively, applicants with differences between two randomly selected ratings would take marginally more time. However, the total number of ratings would still fall, and the committee would still have improved information to use in its deliberations.

Figure X-1
Differences in Raw Ratings



Note: 500 applicants with at least two ratings

If the admissions committee wishes to maximize the information content of additional ratings on the margin, then RGS should assign members to evaluate applicants without regard to the content of prior evaluations. Committee members would evaluate applicants without any knowledge of how other committee members rate them. This might make some evaluations harder for committee members who rely on other members judgments. However, to the extent one member's evaluation relies on another's, such work represents replication and waste, as well as potential reinforcement of an unreliable measurement.

The current RGS admissions process has implemented a particular version of the preceding general recommendations. The committee still requires the same information from applicants as described in Section III (according to roughly the same annual cycle). However, two individuals now screen all applicants.[94] (For several years, the same two individuals have served in this capacity. This eliminates approximately two thirds of the applicant pool from further consideration. (As Table V-2 shows, approximately 55% of applicants previously received three or more ratings.) Thus, the changes in the process improve reliability of ratings in two ways: they reduce variability due to changing composition of the judging pool and they ensure a minimum of two evaluations for each candidate. The full committee then considers the remaining one third of the applicant pool, focusing on those candidates considered on the margin between admission and rejection. Thus, the process devotes relatively more resources to those candidates where additional evaluations might actually influence the admissions decision. Approximately 15% to 20% of the applicant pool receive admission (considerably lower than the data for this analysis reflect).

Enhance Evaluation Policies

Since the RGS admissions committee has not yet implemented the recommendation to clarify objectives, the final recommendation made by this dissertation must remain somewhat incomplete. However, the specific changes that follow should make it clear how to adapt these recommendations to any specific objectives that the committee articulates.

---

[94] The Dean no longer chairs the admissions committee.

If RGS adopts multiple selection objectives, then it might also wish to require multiple ratings (one corresponding to each objective) for each candidate. First, committee members could explicitly predict future performance with one or more ratings. One rating could represent predicted GPA (or relative GPA, depending on performance measurement scale). Another could represent predicted probability of attrition, another predicted probability of distinction, and so on. The admissions committee should determine which performance measures to emphasize and in what proportion, if it wishes to include each of these performance measures in its selection objectives. Differences among the measures considered in this dissertation suggest more rather than fewer measures, since performance has several dimensions. Committee members could predict any and all future performance that they consider in their institutional objectives (for example, OJT performance, student satisfaction or achievements after graduate school). These multiple ratings can make it easier to quantify the uncertainty inherent in the admissions process. The admissions committee could also then have an outcome criterion measurement for each of its selection objectives, making it possible to track progress against each.

Second, RGS should assign an overall rating of desirability for each applicant, to the extent that it wishes to make such a measure explicit. This might reconcile the apparent divergence between average rating and admissions. Such divergence may reflect unequal influence of individual committee members. Thus, an overall rating would not likely improve the predictability of admissions decisions. In any case, the rational assignment of applicant ratings discussed above will make it easier to detect such influence.

Since evaluation involves both relative and absolute aspects, RGS could require at least two evaluations in place of course grades. The first could represent the instructor's relative rating of each student within a class. The second could refer explicitly to some external standard.[95] Of course, no such standard can have complete objectivity, but some standards will allow for easier comparisons of relative ratings across courses, departments and cohorts. For example, instructors could compare each course explicitly with the previous course (usually taught in the previous year, mostly with students from a prior cohort). Alternatively, the external reference

[95] This will allow decision makers to analyze whether or not such an external standard appears to move anyway, as proponents of grade inflation theories assert.

point could represent the ideal student, though each instructor will undoubtedly conceive this ideal in different ways.

RGS could also establish explicit dimensions of performance evaluation in place of course grades. For example, instructors could evaluate the following attributes:

- technical competence in subject material
- contribution to class discussion and group learning
- ability to apply techniques to practical policy problems

Each dimension could have both an absolute and a relative rating. In other words, RGS could develop multiple performance dimensions independently of relative and absolute measures.) In addition, course performance measurement could include predictions of qualifying examination and dissertation performance (such as completion time). This could make it possible to identify potential problems relating to specific institutional and selection objectives.

RGS could begin to develop these attributes by studying the comments currently given by course instructors along with course grades. These attributes need not replace instructor comments. However, they could help to give such comments a discipline so that students would always receive feedback about specific aspects of their performance. In addition, specifying these attributes would help admissions decision makers clarify their objectives, particularly regarding performance.

Any such enhancements to performance measurement would probably also require a process to include input from current and prospective faculty members, in order to establish their feasibility. This process should involve some formal training or indoctrination regarding the meaning and purpose of various measures, if RGS wishes to encourage consistency across faculty members.

Similarly, enhancing the measurement scale for qualifying exam results might both increase predictability of results and add meaning to the measurement. RGS could (and clearly should) retain some absolute standard for passing the exams (for example, a minimum score of 80 on a 1-100 scale). These would provide the same, highly reliable performance measures used in this dissertation. But future analyses could treat qualifying exam results as a matter of degree,

in addition to using a binary indicator. Furthermore, more continuous measurement scales for exam results would give students a more precise indication of their performance. (Some might not value this increased precision.) Dissertation performance measurement could also benefit from similar changes.

Implications

As noted in Section IV, several considerations may limit the applicability of this research beyond the specific case study. Previous validation studies do not explicity address public policy as an academic field. Institutional and selection objectives at RGS probably differ at least slightly from those at other institutions. The selection and evaluation process at RGS also differ from those at other institutions. Thus, empirical analysis of RGS admissions and performance data do not fit the exact context of the literature review.

Despite these limitations, this analysis does seem to have broader implications. Regarding previous validation studies (and related educational psychology issues), this analysis did not attempt to disprove or contradict the established literature. In fact, this dissertation more or less confirms the superiority of statistical models over clinical predictions. More to the point, this research provided an illustration of how to consider the multiple dimensions of the "criterion" in criterion-referenced validation. This illustration also illustrated the impact of measurement scales in a setting with clear policy implications for a particular academic institution. (This departs from much of the validation research, which emphasizes the implications for a particular test used by multiple institutions.)

Although RGS represents a unique set of objectives and processes, the analytical framework established here can accomodate those of different institutions. Indeed, only within a particular institutional setting could decision makers make any practical, empirical link between institutional and selection objectives, on the one hand, and selection criteria, ratings and performance evaluation, on the other hand. RGS, or any other school, will always have a unique particular way that these factors interact. Every school, though, can discover their own particular realization of these interactions.

Some of the preceding policy recommendations would set a more clear direction for empirical analyses.[96] More importantly, they also provide an appropriate context for understanding and interpreting such analyses. Decision makers can focus on predicting and maximizing performance if they wish. They may even, quite logically, choose to implement a screening formula to accept or reject some or all applicants. But such choices should not obscure their fundamental task of determining and clarifying objectives.

The role of judgment in admissions extends beyond articulating objectives. Regardless of selection criteria (formulaic or not), decision makers need to assess periodically how well their institutions have achieved their objectives. The empirical analyses and broader conceptual framework presented in this dissertation will hopefully provide some useful tools for that assessment, as well as an appreciation for the magnitude of the task.

[96] Enhancing evalution to include additional measures, including more explicit predictions and indications of overall desirability would make it possible to reconsider the conclusion that overall ratings fail to capture certain dimensions of both "desirability" (as embodied in actual admissions decisions) and future performance. Furthermore, development of absolute and relative measures, as well as multidimensional performance and selection evaluation, would change the nature of the data analyzed for this dissertation.

BIBLIOGRAPHY

Adler, Mortimer, *Ten Philosophical Mistakes*, MacMillan Publishing Co., New York, 1985.

Akemann, Charles A., Bruckner, Andrew M., Robertson, James B., Simons, Stephen, and Weiss, Max L., "Conditional Correlation Phenomena With Applications To University Admisisons Strategies," *Journal of Educational Statistics*, Volume 8, Number 1, pp. 5-44, Spring 1983.

Anderson, C. Arnold, and Bowman, Mary Jean (editors), *Education and Economic Development*, Aldine Publishing Co., Chicago, IL, 1965.

Angoff, William H., "Scales, Norms and Equivalent Scores," in Thorndike, Robert L. (editor), *Educational Measurement*, American Council on Education, Washington, DC, 1971.

Ashenfelter, Orley, and Mooney, Joseph D., "Graduate Education, Ability and Earnings," *Review of Economics and Statistics*, Volume 50, Number 1, pp. 78-86, February 1968.

Assessment in a Pluralistic Society, Proceedings of the 1972 Invitational Conference on Testing Problems, Educational Testing Service, Princeton, NJ, 1972.

Baird, Leonard L., "Do Grades and Tests Predict Adult Accomplishment?" *Research in Higher Education*, Volume 23, Number 1, pp. 3-85, 1985.

Bell, Robert M., *Medical School and Physician Performance: Predicting Scores on the American Board of Internal Medicine Written Examination*, R-1723-HEW, The RAND Corporation, Santa Monica, CA, 1977.

Berg, Ivar, *Education and Jobs: The Great Training Robbery*, Praeger Publishers, New York, NY, 1970.

Blandy, R., "Marshall on Human Capital: A Note," Journal of Political Economy, Volume 75, Number 1, pp. 874-875, December, 1975.

Blaug, Mark (editor), *Economics of Education I*, Penguin Books, Baltimore, MD, 1968.

Blaug, Mark, *Economics of Education: A Selected Annotated Bibliography*, Pergamon Press, New York, NY, 1970.

Bloom, Allan, *The Closing of the American Mind: How Higher Education Has Failed Democracy and Impoverished the Souls of Today's Students*, Simon and Schuster, New York, NY, 1987.

Bornheimer, Deane G., "Predicting Success in Graduate School Using GRE and PAEG Aptitude Test Scores," *College and University*, Volume 60, Number 1, pp. 54-62, Fall 1984.

Bowman, Mary Jean, Debeauvais, Michel, Komarov, V. E., and Vaizey, John (editors), *Readings in the Economics of Education*, United Nations Educational, Scientific and Cultural Organization, Paris, 1968.

Braun, Henry I., and Szatrowski, Ted H., "Validity Studies Based on a Universal Criterion Scale," *Journal of Educational Statistics*, Winter 1984, vol. 9, no. 4, pp. 331-344.

Braun, Henry I., and Szatrowski, Ted H., "The Scale-Linkage Algorithm: Construction of a Universal Criterion Scale for Families of Institutions," *Journal of Educational Statistics*, Winter 1984, vol. 9, no. 4, pp. 311-330.

Cahn, Steven (editor), *Classics of Western Philosophy*, Hackett Publishing Co., Indianapolis, IN, 1977.

Chan, Steve, "Expert Judgments Under Uncertainty: Some Evidence and Suggestions," *Social Science Quarterly*, Volume 63, Number 3, pp. 428-444, 1982.

Cohn, Elchanan, *The Economics of Education*, Ballinger Publishing Co., Cambridge, MA, 1979.

Conrad, Linda, Trismen, Donald, and Miller, Ruth (editors), Graduate Record Examinations Technical Manual, Educational Testing Service, Princeton, NJ, 1977.

Cronbach, Lee J., and Gleser, Goldine C., *Psychological Tests and Personnel Decisions*, University of Illinois Press, Urbana, IL, 1965.

Cronbach, Lee J., *Essentials of Psychological Testing*, Harper and Row, New York, NY, 1970.

Cronbach, Lee J., "Five Decades of Public Controversy Over Mental Testing," *American Psychologist*, Volume 30, Number 1, pp. 1-14, January 1975.

Cronbach, Lee J., "Test Validation," in Thorndike, Robert L. (editor), *Educational Measurement*, American Council on Education, Washington, DC, 1971.

Dailey, Dennis M., "The Validity of Admissions Predictions: A Replication Study and Implications for the Future," *Journal of Education for Social Work*, Volume 15, Number 2, pp. 14-22, Spring 1979.

Daniels, Mathe J. M., and Schouten, J., *The Screening of Students*, George G. Harrap and Co., Ltd., London, 1970.

Dawes, Robyn M., "A Case Study of Graduate Admissions: Applications of Three Principles of Human Decision Making," *American Psychologist*, February, 1971, vol. 26, no. 2, pp. 180-188.

Dawes, Robyn M., and Corrigan, Bernard, "Linear Models in Decision Making," *Psychological Bulletin*, Volume 81, Number 2, pp. 95-106, February 1974.

Dole, Arthur A., and Baggaley, Andrew R., "Prediction of Performance in a Doctoral Education Program by the Graduate Record Examinations and Other Measures," *Educational and Psychological Measurement*, Volume 39, Number 2, pp. 421-427, Summer 1979.

Educational Change: Implications for Measurement, Proceedings of the 1971 Invitational Conference on Testing Problems, Educational Testing Service, Princeton, NJ, 1971.

Educational Indicators: Monitoring the State of Education, Proceedings of the 1975 ETS Invitational Conference, Educational Testing Service, Princeton, NJ, 1975.

Goldman, Roy D., and Slaughter, Robert E., "Why College Grade Point Average is Difficult to Predict," *Journal of Educational Psychology*, Volume 68, Number 1, pp. 9-14, 1976.

Griliches, Zvi, "Estimating The Returns To Schooling: Some Econometric Problems," *Econometrica*, Volume 45, Number 1, pp. 1-22, January 1977.

Hackman, J. Richard, Wiggins, Nancy, and Bass, Alan R., "Prediction of Long-Term Success in Doctoral Work in Psychology," *Educational and Psychological Measurement*, Volume 30, Number 2, pp. 365-374, Summer, 1970.

Hartnett, Rodney T., and Willingham, Warren W., *The Criterion Problem: What Measures of Success in Graduate Education?* Graduate Record Examination Board, Princeton, NJ, 1979.

Hartnett, Rodney T., and Feldmesser, Robert A., "College Admissions Testing and the Myth of Selectivity: Unresolved Questions and Needed Research," *AAHE Bulletin*, Volume 34, pp. 3-6, March, 1980.

Harvancik, Mark J., and Golsan, Gordon, "Graduate Record Examination Scores and Grade Point Averages: Is There A Relationship?" *Proceedings of the Annual Convention of the American Association for Counseling and Development*, Los Angeles, CA, 1986.

Hecht, Lawrence W., and Powers, Donald E., "The Predictive Validity of Preadmission Measures in Graduate Management Education," Proceedings of the Annual Meeting of the *American Educational Research Assocation*, American Educational Research Assocation, New York, NY, 1982.

Hecht, Lawrence W., and Schrader, William B., *Graduate Management Admissions Test: Technical Report on Test Development and Score Interpretation for GMAT Users*, Graduate Management Admissions Council, Los Angeles, CA, 1986.

Hillier, Frederick S., and Lieberman, Gerald J., *Introduction to Operations Research*, Holden-Day, Inc., Oakland, CA, 1980.

Hills, John R., "Use of Measurement in Selection and Placement," in Thorndike, Robert L. (editor), *Educational Measurement*, American Council on Education, Washington, DC, 1971.

Hoyt, Donald P., *The Relationship Between College Grades and Adult Achievement. A Review of the Literature*, ACT Research and Development Division, Iowa City, IA, 1965.

Hughes, P. W., *Academic Achievement at the University*, University of Tasmania, Hobart, Australia, 1960.

Ingram, Rick E., "The GRE in the Graduate Admissions Process: Is How It Is Used Justified By The Evidence Of Its Validity?" *Professional Psychology: Research and Practice*, Volume 14, Number 6, pp. 711-714, 1983.

Jencks, Christopher, et. al., *Who Gets Ahead? The Determinants of Economic Success in America*, Basic Books, Inc., New York, NY, 1970.

Kalbfleisch, John D., and Prentice, Ross L., *The Statistical Analysis of Failure Data*, John Wiley and Sons, New York, NY, 1980.

Keeley, Stuart M., and Doherty, Michael E., "Bayesian and Regression Modeling of Graduate Admission Policy," *Organizational Behavior and Human Performance*, Volume 8, Number 2, pp. 297-323, 1972.

Klitgaard, Robert, *Choosing Elites*, Basic Books, New York, NY, 1985.

Klitgaard, Robert, *Achievement Scores and Educational Objectives*, R-1217-NIE, The RAND Corporation, Santa Monica, CA, 1974.

Lafferty, Gladys E., *A Study of the Influence of Age on Predictability of Graduate Record Examinations Aptitude Tests for Successful Graduate Students*, South Carolina University School of Education, Columbia, SC, 1969.

Lannholm, Gerald V., *Review of Studies Employing GRE Scores in Predicting Success in Graduate Study: 1952-1967*, Graduate Record Examination Board, Princeton, NJ, 1968.

Leonardson, Gary R., "The Contribution of Academic Factors in Predicting Graduate School Success," *College Student Journal*, Volume 13, Number 1, pp. 21-24, Spring 1979.

Lord, Frederic M., "An Analysis of the Scholastic Aptitude Test Using Birnbaum's Three Parameter Logistic Model," *Educational and Psychological Measurement*, Volume 28, Number 4, pp. 989-1020, Winter 1968.

Lord Frederic M., and Novick, M., *Statistical Theories of Mental Test Scores*, Addison-Wesley, Inc., Reading, MA, 1968.

Luse, F. Dean, and Meyer, S. Ruth, "Decision Tables in Admission Procedures," *The Indian Journal of Social Work*, Volume XL, Number 4, pp. 399-406, January 1980.

Lyman, Howard B., *Test Scores and What They Mean*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

Leonardson, Gary R., "The Contribution of Academic Factors in Predicting Graduate School Success," *College Student Journal*, Volume 13, Number 1, pp. 21-24, Spring 1979.

McMahon, Walter W., *Investment in Higher Education*, Lexington Books, Lexington, MA, 1974.

Meehl, P. E.,*Clinical vs. Statistical Prediction: A Theoretical Analysis and Review of the Literature*, University of Minnesota Press, Minneapolis, MN, 1954.

Mehrabian, Albert, "Undergraduate Ability Factors in Relationship to Graduate Performance," *Educational and Psychological Measurment*, Volume 29, Number 2, pp. 409-419, Summer 1969.

Merenda, Peter F., and Reilly, Raymond, "Validity of Selection Criteria in Determining Success of Graduate Students in Psychology," *Psychological Reports*, Volume 28, Number 1, pp. 259-266, February 1971.

Messmer, Donald J., and Solomon, Robert J., "Differential Predictability in a Selection Model for Graduate Students: Implications for Validity Testing," *Educational and Psychological Measurement*, Volume 39, Number 4, pp. 859-866, 1979.

Miller, Rupert G., *Survival Analysis*, John Wiley and Sons, New York, NY, 1981.

Mitchelson, Ronald L., and Hoy, Don R., "Problems in Predicting Graduate Student Success," *Journal of Geography*, Volume 83, Number 2, pp. 54-57, March-April 1984.

Nader, Ralph, *The Reign of ETS: The Corporation That Makes Up Minds*, Allen Nairn and Associates, Washington, DC, 1980.

Noonan, Richard D., *School Resources, Social Class, and Student Achievement*, John Wiley and Sons, New York, NY, 1976.

Permut, Steven E., "Modeling the Graduate Admissions Committee," *Psychology*, Volume 10, Number 1, pp. 3-4, 1973.

Pierson, Gordon Keith, *The Investigation of the Contribution of Education to Economic Growth*, Ph.D. Dissertation, University of Washington, Seattle, WA, 1963.

Popham, W. James (editor), *Criterion-Referenced Measurement*, Educational Technology Publications, Englewood Cliffs, NJ, 1980.

Powers, Donald E., and Moss, Pamela A., *A Summary of the Results of the Graduate Management Admission Council (GMAC) Validity Study Service for 1979-1980*, Educational Testing Service, Princeton, NJ, 1980.

Psacharopoulos, George, *Returns to Education: An International Comparison*, Jossey-Bass, Inc., San Francisco, CA, 1973.

Purcell, Edward A., Jr., *The Crisis of Democratic Theory*, The University Press of Kentucky, Lexington, KY, 1973.

The RAND Graduate School of Policy Studies: Bulletin, 1988-1989. The RAND Corporation, Santa Monica, CA, 1988.

Reilly, Richard R., *Contributions of Selected Transcript Information to Prediction of Law School Performance*, Educational Testing Service, Princeton, NJ, 1971.

Rem, Richard J., Oren, Ehud M., and Childrey, Gary, "Selection of Graduate Students in Clinical Psychology: Use of Cut-Off Scores and Interviews," *Professional Psychology: Research and Practice*, Volume 18, Number 5, pp. 485-488, 1987.

Remus, William, and Wong, Clara, "An Evaluation of Five Models for the Admission Decision," *College Student Journal*, Volume 16, Number 1, pp. 53-59, Spring 1982.

Riley, John G., "Testing the Educational Screening Hypothesis," *Journal of Political Economy*, Volume 87, Number 5, Part 2, pp. 227-252, 1979.

Rolph, John E., Williams, Albert P., and Laniear, A. Lee, *Predicting Minority and Majority Medical Student Performance on the National Board Exams*, R-2029-HEW, The RAND Corporation, Santa Monica, CA, 1978.

Rolph, John E., Williams, Albert P., and Lee, Carolyn, *The Effect of State of Residence on Medical School Admissions*, R-2014-HEW, The RAND Corporation, Santa Monica, CA, 1978.

Roscoe, John T., and Houston, Samuel R., "The Predictive Validity of GRE Scores for a Doctoral Program in Education," *Educational and Psychological Measurement*, Volume 29, Number 2, pp. 507-509, 1969.

Schrader, William B., *The Graduate Management Admission Test: Technical Report on Test Development and Score Interpretation for GMAT Users*, Graduate Management Admission Council, Princeton, NJ, 1984

Schultz, Theodore W., *Investment in Human Capital: The Role of Education and Research*, The Free Press, New York, NY, 1971.

Slack, Warner V., and Porter, Douglas, "The Scholastic Aptitude Test: A Critical Appraisal," *Harvard Educational Review,* Volume 50, Number 2, pp. 154-175, May, 1980.

Sobol, Marion G., "GPA, GMAT, and Scale: A Method for Quantification of Admissions Criteria," *Research in Higher Education,* Volume 20, Number 1, pp. 77-88, 1984.

Solomon, Lewis C., and Taubman, Paul J. (editors), *Does College Matter? Some Evidence on the Impacts of Higher Education,* Academic Press, New York, NY, 1973.

Spence, A. Michael, "Job Market Signaling," *Quarterly Journal of Economics,* Volume 87, pp. 355-374, August, 1974.

Spence, A. Michael, *Market Signaling: Informational Transfer in Hiring and Related Screening Processes,* Harvard University Press, Cambridge, MA, 1974.

Srinivasan, V., and Weinstein, Alan G., "Effects of Curtailment on an Admissions Model for a Graduate Management Program," *Journal of Applied Psychology,* Volume 58, Number 3, pp. 339-346, 1973.

Statistical Abstract of the United States, Bureau of the Census, Washington, DC, 1994.

Stiglitz, Joseph E., "The Theory of 'Screening,' Education and the Distribution of Income," *The American Economic Review,* Volume 65, Number 3, p. 283, June, 1975.

Stolzenberg, Ross M., and Relles, Daniel A., *Calculation and Practical Application of GMAT Predictive Validity Measures,* Graduate Management Admission Council, Los Angeles, CA, 1985.

Stolzenberg, Ross M., *Design for a Survey of New Matriculants in Graduate Schools of Business and Management,* Graduate Management Admission Council, Los Angeles, CA, 1988.

Stolzenberg, Ross M., and Relles, Daniel A., "The Utility of Empirical Bayes Methods for Comparing Regression Structures in Small Samples," in C. Clogg (editor), *Sociological Methodology 1989,* American Sociological Association, Washington, DC, 1989.

Stordahl, Kalmer E., *Predictive Validity of the Graduate Record Aptitude Test,* Northern Michigan University, Marquette, MI, 1970.

The Promise and Perils of Educational Information Systems, Proceedings of the 1970 Invitational Conference on Testing Problems, Educational Testing Service, Princeton, NJ, 1970.

Thornell, John G., and McCoy, Anthony, "The Predictive Validity of the Graduate Record Examinations for Subgroups of Students in Different Academic Disciplines," *Educational and Psychological Measurement,* Volume 45, Number 2, pp. 415-419, Summer, 1985.

Tyler, Ralph (editor), *Educational Evaluation: New Roles, New Means,* University of Chicago Press, Chicago, IL, 1969.

Tyler, Ralph, and Wolf, Richard M. (editors), *Crucial Issues in Testing,* McCutchen Publishing Corp., Berkeley, CA, 1974.

Wainer, Howard, and Braun, Henry (editors), *Test Validity,* Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

Wallace, Marc J., Jr., and Schwab, Donald B., "A Cross-Validated Comparison of Five Models Used to Predict Graduate Admissions Committee Decisions," *Journal of Applied Psychology,* Volume 61, Number 5, pp. 559-563, 1976.

Williams, Albert P., Cooper, Wendy D., and Lee, Carolyn L., *Factors Affecting Medical School Admission Decisions for Minority and Majority Applicants: A Comparative Study of Ten Schools,* R-2030-HEW, The RAND Corporation, Santa Monica, CA, 1979.

Willingham, Warren, "Predicting Success in Graduate Education," *Science,* Volume 183, pp. 273-278, January, 1974.

Wilson, Kenneth M., *The Validation of GRE Scores as Predictors of First-Year Performance in Graduate Study: Report of the GRE Cooperative Validity Studies Project,* Graduate Record Examinations Board, Princeton, NJ, 1979.

Wilson, Kenneth M., *A Review of the Research on Prediction of Academic Performance After the Freshman Year,* College Board Report Number 83-2, College Entrance Examination Board, New York, NY, 1983.

Windham, Douglas M., *Education, Equality and Income Redistribution,* Heath Lexington Books, Lexington, MA, 1970.

Windham, Douglas M., *The Benefits and Financing of American Higher Education: Theory, Research and Policy,* Institute for Research on Educational Finance and Governance, Stanford, CA, 1980.

Wise, David A., "Academic Achievement and Job Performance," *The American Economic Review,* Volume 65, Number 3, p. 350, June, 1975.

Wise, David A., "Personal Attributes, Job Performance, and Probability of Promotion," *Econometrica,* Volume 43, Number 5-6, p. 913, September-November, 1975.

Wolff, Robert Paul, *The Ideal of the University,* Beacon Press, Boston, MA, 1969.

Wolpin, Kenneth, I., "Education and Screening," *American Economic Review,* Volume 67, Number 8, pp. 949-958, December, 1977.

Youngblood, Stuart A., and Martin, Billy J., "Ability Testing and Graduate Admissions: Decision Process Modeling and Validation," *Educational and Psychological Measurement,* Volume 42, Number 4, pp. 1153-1162, Winter, 1982.

Zedeck, Sheldon, and Kafry, Ditsa, "Capturing Rater Policies for Processing Evaluation Data," *Organizational Behavior and Performance,* Volume 18, Number 2, pp. 269-294, 1977.