ETS GRE®

Assess Ability. Predict Performance.

ETS GRE® Board Research Report
ETS GRE®–14-04
ETS Research Report No. RR–14-27

A Descriptive Study of Universities' Use of *GRE*® General Test Scores in Awarding Fellowships to First-Year Doctoral Students

JoAnn Leah Rock

Rachel M. Adler

December 2014

The report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

****

****

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

GRE-ETS

PO Box 6000

Princeton, NJ 08541-6000

USA

To obtain more information about GRE programs and services, use one of the following:
Phone: 1-866-473-4373
(U.S., U.S. Territories*, and Canada)
1-609-771-7670
(all other locations)
Web site: www.gre.org

*America Samoa, Guam, Puerto Rico, and US Virgin Islands

GRE Board Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

A Descriptive Study of Universities' Use of *GRE*® General Test Scores in Awarding Fellowships to First-Year Doctoral Students

JoAnn Leah Rock & Rachel M. Adler

Educational Testing Service, Princeton, NJ

The purpose of this study was to investigate the ways in which universities use the *GRE*® General Test scores to award merit-based fellowships to first-year graduate students in doctoral programs. While GRE use in fellowship award decisions is a common practice, there is very little validity evidence to support its use in this context. This exploratory qualitative study uses an argument-based validity approach as a conceptual framework to document GRE use systematically in award decisions. The findings from this study indicate that universities use merit-based fellowships to recruit *the best and the brightest* graduate students. The GRE is used as part of the nominee selection process in departments and is also used in the final selection process by university-wide committees. The GRE scores are primarily used to infer whether students have the baseline knowledge and skills needed to succeed in graduate school. These academic competencies were one of the characteristics the decision makers identified as required for successful doctoral students. The participants stated that the factors for success in graduate school are nebulous; however, the students with high GRE scores and other noteworthy supporting materials are expected to perform strongly in graduate school and, therefore, are worthy of the investment of a merit-based fellowship.

Keywords Score use; score interpretation; GRE, validity; merit-based fellowships

doi:10.1002/ets2.12027

The purpose of this report is to describe the ways in which universities use the *GRE*® General Test scores to award merit-based fellowships to first-year doctoral students. GRE use in awarding fellowships requires validity evidence based on the uses and interpretations of the scores by decision makers in this context. Currently, there is very little systematic documentation of the use and interpretations of scores by fellowship decision makers. The goal of the exploratory qualitative study reported here is to address this gap by documenting use and interpretations of scores in a sample of universities. The exploratory findings from this study are an initial step toward building an interpretive argument for GRE use as part of an argument-based approach to validity.

In an argument-based approach to validity, the uses, interpretations, and claims made by score users are presented in an interpretive argument, and then theoretical and empirical evidence is collected to support or negate the validity of the network of arguments (Kane, 2001). A simplified example might be the use of an achievement score in reading to measure children's morality. Using an argument-based approach, this would not be a valid use of the score because the interpretation of the score is not a plausible inference from what is measured.

An argument-based approach to validation uses the interpretive argument as the catalyst for collecting and presenting validity evidence (Kane, 1992). An interpretive argument consists of explicit statements about the uses and interpretations made when using test scores (Kane, 2001). The goal of validity is to confirm that the score use and interpretations are supported through quantitative and theoretical evidence (Messick, 1989).

According to Kane (2006), when a score is used to make a decision, the interpretive argument includes both a semantic interpretation made when using the score and the decisions based on the semantic interpretations. For example, in examining the use of GRE scores in awarding fellowships, the interpretative argument consists of both the traits and characteristics inferred to students admitted to doctoral programs, when GRE scores are used in the award process, and the student outcomes hypothesized using these interpretations to make decisions.

Corresponding author: J. L. Rock, E-mail: jrock@ets.org

The qualitative study reported here uses this approach to validity as a conceptual framework for the study. In qualitative research, the conceptual framework is a written or visual tool that defines the key factors, concepts, and variables for study and guides the development of each component of the research design (Miles & Huberman, 1994). With this framework in mind, the qualitative study reported here was designed to systematically explore the uses, interpretations of, and claims made when using GRE scores in fellowship decisions.

Qualitative research methodology is a good fit for gathering the information needed to build an interpretive validity argument. Qualitative methods are used to understand the meaning, intentions, and consequences participants give to their actions (Merriam, 1998). For example, Miles and Huberman (1994) stated, "A main task [of qualitative research] is to explicate the ways people in particular settings come to understand, account for, take actions, and otherwise manage their day-to-day situations" (p. 7). The methodological tools are designed to interpret the meaning participants give to their actions. Thus, a strong match to the goals of an argument-based approach to validity is to document and examine the meaning individuals give to test scores.

This study takes a small step toward building an interpretative argument for validity use in fellowship decisions. The patterns and themes identified in the exploratory sample can be confirmed and revised through a larger sample and, subsequently, can be used to build an interpretive argument upon which to build validity evidence for GRE use in this context.

The study sample consisted of nine universities that award first-year merit-based fellowships and offer doctoral programs in the science, technology, engineering, and math fields (STEM); the humanities; and the social sciences. The schools were selected to produce variation in three categories that were identified in preliminary research as potentially important to GRE use in this context: admission selectivity, public or private governance, and geographical region. Using text-based data collected via 50 in-person interviews conducted at the universities in the study sample, nine information-rich case studies were developed. These text-based data were analyzed using a constant comparative multicase qualitative analytic strategy to identify themes and patterns in the use of the GRE in awarding fellowships.

The focus of this study was on the use of GRE scores when awarding merit-based fellowships to first-year doctoral students through graduate deans' offices. While the GRE is used in other contexts to award fellowships, such as master's programs and through foundations, centralized fellowships in universities may use more systematic procedures and provide numerous academic contexts within which patterns of usage can be examined.

Research Questions Guiding the Study

The primary research questions addressed in this study were:

- What are the university goals when awarding merit-based fellowships to first-year doctoral students?
- What is the process used to award merit-based fellowships to first-year doctoral students?
- How are GRE scores used in the decision-making process?
- What inferences are made by decision makers when using GRE scores in the process for awarding fellowships?
- What outcomes do universities intend for students awarded a fellowship when GRE scores are used as part of the decision process?
- Does GRE use in fellowship decisions differ across disciplines or university characteristics?

Review of the Literature

While there is substantial validity evidence to support the use of the GRE in admission decisions (Burton & Wang, 2005; Kuncel, Hezlett, & Ones, 2001; Kuncel, Wee, Serafin, & Hezlett, 2010; Schneider & Briel, 1990), there is little specific evidence to support the scores' use as criteria in awarding fellowships to graduate students. Standards for educational and psychological tests require specific evidence on test validity in each of the contexts in which a test is used (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Validity evidence is shaped by the context in which the score is used (Messick, 1989). One of the primary functions of the GRE is to contribute to the prediction of a student's potential for success in graduate school (Conrad, Trisman, & Miller, 1977; Willingham, 1976). Most of the current validity for the GRE score is based on use within these parameters.

To support the focus on readiness for graduate school, a number of studies have been conducted to confirm the predictive validity of the GRE for use in making admission decisions for graduate students. Scores on the GRE have consistently been found to be predictive of first-year graduate grade point averages (GGPA), cumulative GGPA, and faculty ratings of students (Burton & Wang, 2005; Kuncel & Hezlett, 2007; Kuncel et al., 2001; Kuncel et al., 2010; Lannholm, 1968; Schneider & Briel, 1990; Willingham, 1974). In a study of more than 1,300 students from 128 departments, Burton and Wang (2005) concluded that a combination of GRE scores and undergraduate GPA strongly predicted students' cumulative GGPA, ratings by faculty, mastery of the discipline, potential for professional productivity, and communication ability. Kuncel et al. (2001) conducted a meta-analysis of more than 1,000 studies and found that the GRE was successful in predicting first-year GGPA, cumulative GGPA, faculty ratings, comprehensive examination scores, and citation counts.

In contrast, a few small-scale studies have questioned the value of the GRE in predicting success in graduate school (Milner, McNeil, & King, 1984; Morrison & Morrison, 1995; Sternberg & Williams, 1997). For example, Sternberg and Williams (1997) found that although the GRE predicted Yale psychology graduate students' first-year GGPA, the students' GRE scores were not predictive of their second-year GGPA. Also, one study found that the GRE failed to predict the performance of students in a social work graduate school (Milner et al., 1984).

While there is a great deal of research supporting the use of GRE scores in graduate admission decisions, the evidence base is limited when scores are used as one of the criteria in awarding fellowships to graduate students. King and Besco (1960) evaluated the degree to which research fellowship recipients' GRE Aptitude Test scores predicted faculty ratings of their overall performance in graduate work. While the results indicated a "slight but useful relationship" (p. 856) between the GRE Verbal scores and the faculty ratings, there was no significant relationship between the ratings and GRE Quantitative scores. Based on a survey completed by 245 members of the Council of Graduate Schools, Burns (1970) concluded that "the methods used in selecting students to receive fellowship awards varies considerably, and no clear pattern emerges for either field or type of institution" (p. 31). In this same study, GRE Aptitude Test scores were found to be the fourth most important criterion in making fellowship decisions, preceded by college transcript, letters of recommendation, and completed application form. According to Chapman and McCauley (1993), it is not the case that the fellowship process identifies exceptional students, but that the self-fulfilling expectations associated with receiving the fellowship drives positive student outcomes. In their study of National Science Foundation (NSF) Graduate Fellowship applicants, Chapman and McCauley found that when the NSF awards were distributed quasi randomly, receiving an award had a small, but reliable effect on the likelihood of Ph.D. completion.

Method

This study uses a constant comparative multicase study method to answer the research questions. In qualitative research, the data are text or words collected from participants in their natural environments. Using the constant comparative multicase method, the data from multiple case studies are entered into analytic matrices shaped by the conceptual framework for the study (Merriam, 1998). Using the completed matrices, an inductive process is used to identify a theme in one university, and then to compare and refine the theme through examining the words or data in the next university.

This study uses the argument-based approach to validity as the conceptual framework or guide for the qualitative research method used in this study. In qualitative research, the conceptual framework is a written or visual tool that serves as a theoretical guide for the study. This framework shapes and guides the development of each component of the research design toward the purpose of the study as articulated in the research questions (Maxwell, 2005). In this study, the argument-based approach to validity serves as a guide for the qualitative methods used to collect preliminary information on the proposed uses and interpretations decision makers give to scores in the fellowship process.

Conceptual Framework Guiding the Study

Figure 1 outlines the major components of the conceptual model used in this study based on the argument-based approach to validity. The conceptual framework provides a theoretical guide for the research direction and method used in the study. Maxwell (2005) stated, "The function of the theory is to inform the rest of your design — to help you to assess and refine your goals, develop realistic and relevant research questions, select appropriate methods and identify potential validity threats to your conclusions" (p. 33).

Figure 1 Schematic of conceptual framework used in study based on argument-based validity framework.

The components in the conceptual framework are based on the argument-based approach to validity and were used to define the research questions and the qualitative research methodology used to answer the research questions. An argument-based approach to validation uses the interpretive argument as the framework for gathering validity evidence (Kane, 1992). The interpretations and claims made by score users are defined as explicit statements, which are then confirmed or negated through the collection of theoretical and empirical evidence (Kane, 2001).

When a score is used to make a decision, the interpretive argument includes the semantic interpretations of the test takers' characteristics made by score users, as well as the decisions made based on these interpretations (Kane, 2006). Accordingly, an interpretive argument for the use of GRE scores in awarding fellowships includes the traits and characteristics inferred to students and the student outcomes hypothesized to occur when using these interpretations to make decisions.

Because very little is known about the uses and interpretations made by test score users in the fellowship context, the goal of this study was to gather preliminary information of these uses in a sample of universities as an initial step toward building an interpretative argument.

Study Sample

Sample Selection

A multistage sampling strategy was used to select the study sample (see Appendix A). The strategy was designed to produce a diverse sample of universities that offered one or more merit-based fellowships to first-year doctoral students. The goal was to select a sample of up to 10 universities that offered doctorates in each of three disciplines (STEM, humanities, and social sciences) and differed in three characteristics: selectivity of the university, public or private governance, and geographical region. These criteria were identified through preliminary research as potential explanatory factors in the ways universities use GRE scores in making award decisions.

In the first step of the sample selection process, universities that offered at least one doctoral program in each of the three disciplines (STEM, humanities, and social sciences) were identified using the Carnegie Classification System, a comprehensive list of U.S. colleges and universities. The resulting pool of schools was merged with information from the National Research Council (NRC) Data-Based Assessment of Research-Doctorate Programs, which provides more detailed information about the characteristics of the universities (Ostriker et al., 2010).

This pool of schools, consisting of those on both the Carnegie and NRC lists, was further reduced by excluding schools without *valid* fellowships. In this study, a valid fellowship or assistantship was awarded through the graduate dean's office to incoming doctoral students based on merit. These criteria reduced the recruitment pool to 65 schools. Three additional pieces of information were collected for this pool to facilitate recruitment of a diverse sample: governance structure; geographical region; and estimates of the selectivity of the STEM, humanities, and social sciences doctoral programs.

Estimating Selectivity

Selectivity was calculated for each university using the 2011 U.S. News Report rankings. In the case that the U.S. News Report did not rank any subject within a particular discipline for the university, program selectivity was not calculated. To calculate selectivity for an institution, all of the doctoral programs of universities that the U.S. News Report ranked were listed and sorted into the categories of social sciences, humanities, and STEM. The social sciences consisted of

criminology, economics, political science, psychology, and sociology. The humanities consisted of the arts, education, English, and history. Finally, the STEM programs consisted of biology, chemistry, computer science, earth science, engineering, mathematics, physics, and statistics. For each program within each school, a *selectivity score* was calculated by dividing the U.S. News Report ranking by the total number of schools ranked. This resulted in a number ranging from 0 to 1, such that the lower the number, the more highly ranked the program. The lowest selectivity score was chosen for each division within the school (social sciences, humanities, and STEM) and that number was used as the school's final selectivity score for each division. Finally, each selectivity score was categorized accordingly: high = 0 – 0.33, middle = 0.34 – 0.66, low = 0.67 – 1.0.

The recruitment of universities began with 10 schools selected randomly from the pool of 65 to represent the diverse characteristics of the universities desired in the final sample. A recruitment e-mail was sent to the graduate dean, along with a brief description of the program. Deans were asked to simply indicate interest or no interest. If interest was expressed, a follow-up e-mail or call was initiated, asking the dean to select a contact person for follow-up. Deans were contacted at least twice if no response was received. To establish a diverse sample, the next schools contacted depended on the characteristics of the schools indicating interest in participating. The process continued until a diverse sample was constructed. Twelve schools were ultimately selected and agreed to participate. For various reasons, several schools ultimately chose not to participate, and the final sample was made up of nine universities.

Description of the Sample

The final sample consisted of nine universities that differed on the three selection criteria: selectivity, private or public governance, and geographical region (see Table 1). The schools were fairly evenly distributed across the United States, with three schools in the Midwest, one from the Southwest, two each from the South and Northeast, and one from the West. Most of the universities in the sample were located in urban areas (seven out of nine universities). The seven urban universities were split among public and private universities (four and three, respectively). One rural and one suburban university made up the remainder of the sample.

Of the nine institutions that participated in the study, four were included in the Carnegie Classification System's *very high* research activity category, and the remaining five were in the *high* research activity category. Almost every institution in the sample (eight out of nine) was classified by the Carnegie Classification System as a comprehensive doctoral program, which meant it offered multiple doctoral programs in all disciplines. The final school was deemed a doctoral program that emphasized professions other than engineering (e.g., education and social work).

The universities differed in selectivity using the classification system developed for this study. With regard to the STEM fields, there were three institutions categorized as high, one as middle, and three as low; for the humanities, five schools

Table 1 Characteristics of the Nine Universities Included in the Study

						Selectivity		
University	Region	Governance	Size	Research activity	Setting	STEM	Hum	SS
1	South	Public	Medium	High	Urban	N/A	Low	Low
2	Northeast	Private	Medium	Very high	Urban	N/A	N/A	High
3	Midwest	Public	Large	High	Rural	Low	High	Low
4	Northeast	Public	Large	Very high	Urban	Middle	High	High
5	South	Private	Medium	High	Urban	Low	Middle	Low
6	Midwest	Public	Large	High	Suburban	Low	Middle	Low
7	Southwest	Public	Medium	Very high	Urban	High	High	High
8	Midwest	Public	Large	Very high	Urban	High	High	High
9	West	Public	Large	Very high	Urban	High	High	High

Note. A medium-sized university was defined as having a full-time enrollment of 2,000 to 4,999 students; a large university had a full-time enrollment of 5,000 to 9,999 students. Research activity was determined based on a combination of two indices: the university's aggregate level of research activity and per-capita research activity. Size and activity determinations were reported as indicated in the Carnegie Classification System. STEM = science, technology, engineering, and math; Hum = humanities; SS = social sciences; N/A = not applicable. Selectivity was labeled N/A if the 2011 U.S. News Report did not rank any subject within the discipline for the university.

were categorized as high, two as middle, and one as low; finally, for the social sciences, there were five schools in the high category, none in the middle, and four in the low category. Three institutions had selectivity scores in the high category for all three areas of study.

Measures

On the basis of the conceptual framework for the study, detailed structured interview guides were designed for systematic, in-person data collection. A separate interview guide was developed for each of three types of participants: a dean or director, a faculty member, and an administrator. While the guides covered similar areas of interest, they were customized to capture information unique to the type of study participant. This allowed for consistency across participant roles, while acknowledging divergent viewpoints by position. In general, the guides were developed to cover four main areas: purpose of the fellowship program, characteristics of fellowship applicants, process used to award fellowships, and use of GRE scores in the award process. See Appendix B for a sample interview guide.

Data Collection Procedure

Participant Selection

University deans who agreed to participate in the study were asked to recommend two additional people for interviews. It was suggested that the additional participants include a faculty member who served on an award committee and a faculty member who recommended a student for an award or advised a student who had received an award. In most cases, the university scheduled more than three interviews, with five being the average number of interviews. In total, 50 in-person interviews were conducted.

Interviewees

All interviews were conducted in person at the participants' universities. Most of the interviews were conducted individually. However, due to scheduling conflicts, three interviews were conducted with more than one participant at a time. Prior to the interviews, each participant received a written brief about the study. The interviews were scheduled for 1 h and were conducted by one or more of the three ETS staff members on the research team.

Interviewers

The majority of the interviews were conducted by two ETS staff members (either individually or as a pair) at each of the universities in the sample. (Interviews at one site were conducted with an additional ETS staff member.) Each interviewer received a binder for the school that included information designed to facilitate the interview process. The binder included detailed background information on the university and participants, structured interview guides, write-up forms, consent forms, interview schedules, and descriptions of the study to distribute to participants. Interviewers were provided digital audio recorders and were instructed to obtain written permission from participants to record the interviews or, if permission was denied, to obtain permission from participants to conduct the interviews without recording. All materials taken to the universities were coded to protect confidentiality. Interviewers were trained to provide a brief overview of the study; to keep within the 1-h time commitment; and to use the guide to facilitate conversation, rather than dictate it. Finally, the interviewers completed the write-up forms with initial thoughts and reactions immediately after the interview.

All interviewees signed a consent form, which stated that all identifying information would be removed from the audio recordings and that only ETS researchers would have access to the recordings. Furthermore, participants were informed that they could end the interview at any time, and any notes or recordings would be deleted.

Preparation of Individual Case Records

The data in qualitative research are words or text from study participants collected in their natural environment (Miles & Huberman, 1994). The verbatim text is coded with qualitative analytic tools to identify patterns and themes. The data for

Table 2 Classification Categories and Themes of Analytic Matrices

University goals for fellowship	Fellowship award process	Semantic interpretations	Hypothesized student outcomes
• Increase diversity	• Student criteria	• Academic	• Recruitment
• Grow a department	• Use of GRE	• Research	• Time to completion
• Recruit students	• Role of committee	• Noncognitive	• Professional success
• Improve reputation of university	• Decision-making process		
• Contribute to field	• Other criteria		

this study were prepared for analysis by developing university-specific case records. Case records are text-based documents that serve as the primary resource for qualitative analysis (Patton, 1990).

The case record for each university was developed using the transcripts of the interviews conducted at the university. The interviews were transcribed verbatim by an external firm and included both interviewees' and interviewers' comments. Notes written by the interviewer during and after the interview were also considered part of the case record.

Analytic Strategy

A constant comparative multicase analytic approach was used in this study. The case records were coded or categorized within analytic matrices to identify and confirm patterns in responses within and across the universities. Using the research questions and the completed matrices, themes were developed and confirmed using an inductive process while comparing responses across cases (Merriam, 1998).

Category Construction

The first stage of analysis is the development of meaningful categories of classification based on the conceptual framework, which are then used to code and sort the data. These categories are used in the development of the analytic matrices that allow for within- and across-site analyses. The categories of classification were developed through an initial review of the transcribed data and insights from the in-person interviews. According to Merriam (1998), "Devising categories is largely an intuitive process, but it is also systematic and informed by the study's purpose, the investigators' orientation and knowledge and the meaning made explicit by the participants themselves" (p. 79).

Analytic Matrices

Analytic matrices are displays of text-based data in categories of classification that allow for comparative and inductive analysis. Four analytic matrices were developed, and data from the case records were coded and entered into the matrices manually.

Four theme-based analytic matrices were developed to organize and permit the coding of the data for analysis. The themes were identified using the conceptual framework and the research questions that guided the study. The four themes were: University Goals Supported by the Fellowship, Processes for Awarding Fellowships, Semantic Interpretations of Scores, and Hypothesized Student Outcomes. The categories of classification identified for coding were organized within each theme (see Table 2). All of the universities were included in one matrix to allow for analysis across universities. Three factors used in sample selection—geographical region, governance, and selectivity—were also included in the matrix to help identify factors that could explain patterns in the use of GRE scores in fellowship decisions.

The text from the nine university case records was coded and entered into the matrices manually. Each case record included the text from all of the participants within the university. For example, text or words that discussed granting fellowships in order to improve the university's reputation were entered in the University Goals for Fellowships category. The completed matrix reduced the data and allowed for analyses within and across universities.

Analytic Methodology

A constant comparative multicase analytic strategy was used with the completed matrices to answer the research questions. This method was used to propose, compare, contrast, analyze, and generate patterns from the text-based data that were coded and entered into the analytic matrices (Miles & Huberman, 1994). While qualitative research is partly an intuitive process, an analytic strategy provides a systematic process for developing theories from the words of study participants, which can be tested and elaborated through additional data collection (Patton, 1990).

In a constant comparative strategy, multiple case studies are used to generate theories from the text-based data to answer the research questions. Generalizations developed from one case or university are compared to and refined by the next university in the matrix. Each matrix is analyzed separately. The data within the matrix are analyzed within the classification categories included in the matrix and in which the text-based data have been entered. For example, the process might begin by analyzing the data within the University Goals for Fellowships cell, which contains text addressing recruitment as a university goal for the fellowship. In the first step, the data would be analyzed within one university, using the multiple participants coded within the cell. These data would be used to develop a generalization, such as: recruitment is a major goal for the fellowships awarded at University 1. Then, this generalization would be reviewed for plausibility in the next university. The generalization might be refined or confirmed based on data from the next university. This method would continue across the universities until a general theory was developed that could be further refined and tested over time. After the theories or patterns were identified, the matrices and case records were used to identify particular quotes that were representative of the differences or similarities in the theories across universities.

Results

In the results section, we describe the patterns and themes that emerged from the qualitative analysis of the nine case studies developed on GRE use in fellowship decisions, using an argument-based approach to validity as a conceptual framework. The goal of this study was to describe score use in a sample of universities as an initial step toward the development of an interpretive argument for GRE use in this context. This section is divided into four broad subsections that incorporate the research questions guiding the study. Each of these subsections is structured to provide insights into the factors that are important in building an interpretive argument for GRE use in awarding fellowships.

First, we explore the purposes for the merit-based fellowships for first-year doctoral students. This section includes a discussion of the goals for the fellowships from the viewpoint of the graduate offices and various academic departments. Second, we discuss the characteristics decision makers are looking for in fellows. Third, we discuss the processes used to award merit-based fellowships. In this section, we consider the use of the GRE in the decision-making process and the inferences drawn about student characteristics when using the GRE to make decisions. Fourth, we discuss the way in which differences in university characteristics may influence score use.

The Purpose of Merit-Based Fellowship Programs

Understanding the purpose of the merit-based fellowship program is important in evaluating the plausibility of the score's use in the decision-making process. In this section, we discuss the reasons for using GRE scores from the viewpoints of administrators and faculty involved in the decision-making process. An important part of building a validity argument for the validation process is examining whether the test score is a reasonable measure for the purpose for the contexts in which it is used (Cronbach, 1988).

Fellowships as Recruitment Tools: Sweetening the Pot

The primary purpose of the merit-based fellowship is to recruit students to the university in order to advance the goals of the graduate school. "Yes, the whole process is to attract talented students," said a graduate school administrator at a large urban university. The theme of recruitment was consistent across universities and was supported by both administrators and faculty across a variety of disciplines.

While the universities' goals differed—some aspired to become top-tier universities, others to maintain their selectivity—attracting *the best and the brightest* was part of the solution and, in the participants' views, the merit-based fellowship was an important tool for doing this. One STEM professor explained, "Exactly, it is used as a recruitment tool

and the whole fellowship process is a recruitment tool. That's why we offer these fellowships, to provide an incentive for the most promising graduate candidates to actually come to our program."

The universities viewed the merit-based fellowship as a way to *sweeten the pot* for students who had been offered admission to a doctoral program but had not yet accepted the offer. In other words, the fellowship was used as an incentive to encourage attractive students to accept the program's admission offer. One faculty member put it this way: "It used to be students chasing research opportunities; now PIs [principal investigators] chase fewer, but better qualified students."

The Ball Is in Their Court

The most common timing for use of the fellowship as a recruitment tool is between an offer by a doctoral program to a student and the student's acceptance of the offer. The fellowship is used to encourage a student on the fence to accept the university's offer. "We offer them to candidates that really are good and have shown interest and we hope will come," said a chair of graduate admissions in a STEM field. However, in some cases, the possibility of a fellowship may be mentioned early in the recruitment process. For example, possible candidates to the university might be told at conferences or recruitment weekends about the fellowships as an early incentive to apply.

Defining the Best and the Brightest

While the universities consistently described the fellowship as a way to advance their goals by attracting the best and the brightest, the goals differed across the universities and, subsequently, the meaning of *the best and the brightest* differed. In general, the universities were looking for the same characteristics when referring to the best and brightest, but the definitions changed according to the selectivity of the university and the program.

For example, the goal of the highest tier universities might be to maintain the rigor of the academic programs by enrolling students with multiple offers from similarly competitive programs. "These are the best students, they are students who are being courted by really good schools," said one graduate dean at a highly selective school.

In other cases, the goals of the university may be to become a top-tier program, and to use the fellowships to attract students who are strong candidates, but who may not be admitted to the most selective programs. A faculty member from a program self-described as "second tier," said:

> With the type of university we are, we're not getting the truly elite students, we know that. But once you get past the top 5% of graduate students, we think we're pretty good and we certainly have had our share of graduate students who can and have gone on to be extremely productive researchers and faculty members at other universities.

While both graduate administrators and faculty discussed the fellowships as a recruitment tool, at the department level, the fellowships were discussed as a tool used in an ongoing dialogue with a student. For example, the possibility of a fellowship might be raised in an early conversation with a student who initiated a conversation with a faculty member, once the faculty senses a match. The faculty member "wants to see that there is a plausible match" before moving on to an in-depth discussion of the fellowship. When discussing what the faculty members are looking for, a STEM faculty member said, "The successful fellowship students are ones who had a good amount of research background and ideally some of that will match or correlate with research that's going on in the University. When you see that, that seals the deal."

On the other hand, some departments felt less strongly about a specific match between faculty and student early on, and the fellowship conversation would be more general. As one STEM professor said:

> When we admit a student, the first thing we look for is that the student ought to be a strong student, almost independent of what their research interests are, partly because I would say in the vast majority of cases, . . . their research interests really haven't matured at that stage. . . . And the students themselves recognize and say so in their goals statement that they are open to areas . . . they might encounter once they get to the department.

Providing a student with a fellowship has both substantive and financial implications for a department and its faculty. A fellowship allows the student's advisor to defer funding the student from her external grants, and for the department as a whole, it decreases the amount of money needed to support incoming students. "A fellowship is attractive because it does not count against the internal clock," said a financial officer at a university where students are guaranteed a limited number

of years of funding. Funding for students is an ongoing challenge, particularly in doctoral programs where the time to completion can vary, and the fellowship provides an important increase in the limited funds available to departments. "I'm sure this is a problem for all publics across the country. The privates don't seem to be having the same problem or at least to the same extent," offered a dean at a large public university when discussing funding students.

The importance of obtaining funding through a fellowship program also seemed to differ by discipline. In departments or disciplines where external funding is more easily available or other university fellowships are more lucrative, the merit-based fellowship was less important to the department funding. "In our field there are not many opportunities to get funding, such as an NSF grant, so winning a fellowship is more important."

The only school that did not use GRE scores as a primary tool for awarding fellowships was the one historically Black college and university (HBCU) in the sample. This was partly due to the timing of the award of fellowships, which generally occurred after a student's matriculation in a program. However, a few of the faculty members expressed hesitation in using the GRE to measure student ability for fellowship decisions. Because there was only one HBCU in the sample, it is not clear whether this was a pattern among HBCUs or an individual case.

Characteristics of Fellows: Looking for the Best and the Brightest

It is important to the validation process to understand the characteristics that the test score users are looking to identify in the test takers. The traits or characteristics inferred from a test score are central to the interpretive argument linking a claim to a score, which is the foundation for an argument-based validation approach (Kane, 2006).

Academic Skill, Research Experience, Perseverance: All Cylinders Blasting

The traits or characteristics the decision makers were looking for fell into three broad categories: (a) academic skills and knowledge required for graduate school; (b) research experience; and (c) noncognitive skills, such as perseverance. These characteristics are similar to those described as important for students admitted to doctoral programs, the pool from which the fellows are selected. "We don't expect them to be Einstein," joked one faculty member.

In general, rather than looking for distinctive traits in the fellows, the universities tended to describe the characteristics of fellows in terms of risk. One faculty member from a highly selective program said, "I think the notion really goes back to investments. In the simplest terms, these fellowship students are basically supposed to be no or little risk for a high performance." Another faculty member stated that "most PIs are very risk averse."

Academic Skill

The academic skill needed to succeed in graduate school was a central trait the decision makers were looking for in both their admitted doctoral students and the fellowship recipients. When describing the ideal doctoral student, one dean stated that the student should "have a good baseline academic profile to suggest success." In general, the participants talked about a threshold of academic skill and the ability to advance to a higher level intellectually. As one faculty member explained, "Our primary interest or expectation of a Ph.D. student is someone who is going to have the ability to function at a high level intellectually, to be able to function independently."

When assessing the fellowship candidates, the decision makers were looking for strong indicators that the student had what was considered the baseline skills needed to succeed in graduate school, as opposed to different types of skills than other admitted students. "We have thought about this a long time and we have decided that this is a merit award and academic skill is important," said one member of an award committee.

In the sciences, the necessary academic skills are somewhat more clear-cut than those in the humanities or social sciences. One professor in the sciences said, "We expect high grades, especially in key courses; I mean we look particularly at courses in, say, genetics [and] cell biology . . . as indicators for their ability to do graduate level work."

Higher Order Thinking

Within the framework of academic skills, the participants discussed higher-order reasoning skills, such as logical thought and the ability to synthesize information, as important characteristics in doctoral students. One department head said:

> Any indication that we can have ahead of time as to the student's ability in doing verbal reasoning, in their ability to understand the written word and infer what it is trying to say, and at the same time, try to convey their own ideas in a clear fashion, would be something that would be useful to us.

Communication Skills

The participants also emphasized the importance of writing and communication for success in graduate school. One department head said, "The ability to write, the ability to synthesize and to invest in a sponsor's domain problem area: those are the skills that make someone successful in their research." Another professor expanded on this topic, stating:

> If you want to have a career in research, yes, you have to have the technical abilities, but you do not succeed unless you have the communicative abilities as well. You have to be able to speak, you have to be able to write, and you have to be able to get your point across. That's the only way you get funded, in terms of writing grants, it's the only way you get published, and it's the only way that people will remember what you say at a conference, is to communicate that well. And so you really also have to be a pretty good story teller.

Research Experience

The decision makers saw research experience as important both because of the technical skill involved but also because it suggests that the students know what they are getting into. "I get very frustrated with students who have kind of a low tolerance for failure because, in this field, you tend to have more failures than successes," was the view of one professor who conducts social science research. In some fields with wet labs, where research involves handling biological material as opposed to theoretical work, experience in a lab is essential. A science professor at a highly selective university said:

> We don't take anyone who hasn't had some level of experience in a laboratory as an undergraduate. That's our way of evaluating how good their hands might be when they get in the lab and start doing research when they get here, because that's really the big part of what they do when they arrive.

Noncognitive Skills

When discussing important characteristics of successful graduate students, the conversations frequently turned to noncognitive skills. "Our view is that tenacity, industriousness, the ability to work hard, the ability to stay focused, are additional attributes that we would like to see," said a professor at a competitive private university. In general, although many participants felt ambivalent about identifying these qualities, the traits they deemed important were described as maturity, perseverance, the ability to just "sit and do it," and an ability to manage and recover from failure. At the end of the conversations, the decision makers appeared uncertain about their ability to define and assess these qualities, which they felt were essential to successful doctoral students. One department head maintained that it was "harder to measure certain problem-solving abilities and the motivation . . . the dedication . . . to keep with it." With a shrug of his shoulders, another professor said, "In the end what can you get from a pile of papers? You never really know."

Process for Awarding Merit-Based Fellowships

The universities in this study use a two-phase process to award merit-based fellowships to first-year doctoral students: department nominations and then selections made by a cross-discipline university committee. In the first step of the process, each department nominates students to the graduate school for consideration. "We nominate, we don't award," said a department head at a rural university. Depending on the structure of the university, the department recommendations might go through an additional level of filtering, with the associate dean of each school reviewing the nominees from the various departments and deciding which nominees would be forwarded to the university-wide committee.

In the second step of the process, a university-wide committee reviews the nominees and selects the fellowship recipients. These two phases of the selection process operate according to the criteria for the fellowship set by the graduate school. Typically, the final list of students recommended by the university selection committee is approved by the graduate dean's office.

Within this two-phase structure, the process differed in two major ways across the universities. First, the academic eligibility criteria for the nominees, as determined by the university, differed in the level of specificity. Second, there was some variation in the selection processes used by departments to nominate students and by the university committees to select students to receive the fellowships.

The universities differed in the specificity of the academic success indicators required for fellowship nominees; this was particularly the case with GRE scores and GPA. In about half of the universities, the fellowship had a minimum for GPA and GRE scores. The minimums might include specific cut scores for each of the GRE sections along with a minimum undergraduate GPA, and in at least one case, a minimum combined GRE score. A public university dean said:

> We had a large discussion about that [weighting criteria], and the bottom line, most of us agreed, was that when it comes to merit fellowships, these are supposed to be awards given to the best and the brightest students, and academic record should carry the day; and one element . . . that should be strongly considered are test scores.

While there was still an emphasis on academic skill at the other universities, the requirements were more general and did not include specific cut scores. However, at all of the universities—even at schools where there were no required minimum GRE or other standardized test scores—participants felt that high GRE scores were essential for a student to be considered competitive in the university-wide fellowship selection process. The sentiment, "If you don't have a relatively high GRE score, you are dead in the water," was common across the schools. This may be because of the correlation between GRE scores and other materials reviewed in the fellowship process. "GRE scores do count, but if you look at the top applicants, it's unusual if you have a 3.8, 3.9 GPA, did lab work, and have low GRE scores. I mean these all kind of correlate together," said a professor involved in the decision-making process.

Several of the universities had strict restrictions on using cut scores or using the GRE or other standardized tests as the sole criteria for admission or granting fellowships. In some instances, these decisions were largely the result of court cases that ruled that using a cut score (in particular, a composite score) did not provide equal opportunity for all students.

The departments were required to get a waiver from the graduate dean's office when a nominee did not meet the specific eligibility requirements for the fellowship. The topic of waivers was mentioned most frequently when discussing nominees that met or exceeded the GRE Quantitative Reasoning score minimum set by the graduate school, but did not meet the GRE Verbal Reasoning score minimum.

The participants were divided about this use of the waivers for both admission and fellowship decisions, with some feeling the dichotomy was acceptable for a successful applicant and others expressing concern. One faculty member in a STEM field said:

> Occasionally, one does hear rumblings now and then with people saying, "Well, if somebody's in a technical field, perhaps the GRE Verbal requirement ought not be as strong as it might be in the humanities field," for example. But I think they all are beginning to realize that communication is just as important, written and oral, on the technical side as it is on the humanities side.

The Department Nomination Process: Put Your Candidates Forward

The fellowship selection process begins with the departments nominating candidates for review by the university committee. The departments' goal is to select fellowship nominees from the pool of students with admission offers who would be most successful in a cross-discipline competition. A professor from a competitive social sciences program explained:

> The fellowships are competitive, of course, more competitive than the admission process itself, which is already pretty competitive. So we know that we only have a chance by nominating files that are just kind of strong and firing on all cylinders, really.

The department nomination process is prompted by a request from the graduate dean's office for fellowship nominees. As one committee member joked, the call goes out to each department to "put your candidates forward." One concern among participants from several universities was that the fellowship-granting process needs to start earlier to allow offers to be used more effectively as recruitment tools. The nominee request from the dean's office includes the number of

students that the department or school can nominate, which is governed by university finances, the department size, and how important it is to spread the fellowships across the various disciplines.

In general, a small committee conducts the nomination process within each department. In some cases, this was an admission committee and, in other cases, a separate committee nominated fellowship recipients. The committee either reviewed the credentials of all admitted students or only the students nominated by particular faculty members. For each student, the committee received materials that were part of the admission packet. This included: (a) GRE or other standardized test scores required for admission, (b) academic transcripts and GPA, (c) the student's statement of purpose and description of research experience, and (d) recommendation letters. In general, the committee was looking for students who were stellar in all categories.

This two-phase process — department nomination followed by university committee selection — supports a balance between the department needs and the university needs in the fellowship-granting process. For example, the department nomination process allows the criteria important to each department to be considered through the nominees chosen for the university-wide selection process, where the student characteristics important to each particular discipline might be less evident. The graduate dean at a public university said:

> I think you need some input from the programs, in terms of the types of students that they see that they need, because we don't fully understand every one of the programs, what their current goals are within their program. . . . But I also think that you have to make sure that the overall academic quality of the graduate school or all graduate programs is moving forward. I think it has to be done sort of centrally at the graduate school to make sure that we're bringing in the best students, regardless of the disciplines that they apply into.

Departments' Processes for Ranking Students

All departments had a system for ranking their internal nominees from the admitted pool of students. In most departments, the ranking process was complex and involved weighing the factors important to their discipline with an eye toward the cross-disciplinary factors. "The departments really look at the students . . . that would make a contribution to the program, and . . . they know the students that will compete well with the fellowships," said an administrator for one university's fellowship programs in reference to the department nomination process.

In general, the product of the internal department process is a ranked list of candidates for the cross-discipline university selection process. The characteristics of graduate students identified as important to success — academic, research, communication, and noncognitive skills — are all considered in ranking the students. The department fellowship committee assesses these factors through the students' GRE scores, GPA, transcripts, statements of purpose, and recommendation letters included in the fellowship application package.

The process varied across universities, particularly in the extent to which a formulaic average of the different components was used to rank the students. For example, in one department, a student was given a 1 for every 50 points over a threshold GRE value, which reflected the importance of the score in the discipline; if the threshold GRE Verbal Reasoning score was 550, and the student scored 600, that student would receive a 1. This same process would be used for the GPA and recommendation letters (a point given over a set threshold), resulting in a rank score for the student. In other cases, the departments were given instructions on how the ranking should be determined, such as: one third academic skill, one third research experience, and one third faculty references.

The GRE scores and GPA were always included in the decision-making process, but they were not used as the sole determinant. "We rely heavily on GRE scores to select and rank our nominees for fellowships and it's our perception that it's the most efficient way [to do this]," said one department head where the GRE was weighted heavily in the ranking process. Others felt strongly that there was too much emphasis on the GRE in the decision-making process. For example, a science faculty member expressed this view: "Each of these pieces of the puzzle — the [GRE] scores, the written materials, and the face-to-face [interview] — each of them has its own merits and limitations, and so we judge the whole package."

However, in no case were the GRE scores used as the single determinant. "The GRE score is one piece of information used to assess students' quality," said a public school dean. This was partly due to the university's interest in assessing qualities not easily measured by the scores, but also because at that point in the process, there was limited meaningful variation in the GRE scores. "I don't see a large differentiation between the GRE scores that we see. We get very few students that are at the top. . . . Many of our students score in the 1,000 to 1,200 range."

While keeping in mind the eventual evaluation of nominees by reviewers outside of the discipline, most of the departments weighed the various factors, particularly the GRE scores, in terms of their significance to the discipline. That was one way by which the departments put their personal stamp on the selection process. For example, a reasonable GRE Quantitative Reasoning score matched with a high GRE Verbal Reasoning score might not be considered problematic in a humanities department. The discipline focus was more readily apparent in departments' reliance on GPAs, which indicate whether a student had taken and succeeded in the types of courses necessary to thrive in a particular program.

At the same time, the departments looked for balance and consistency across the measures in the fellowship application packet, partly because of what it said about the student, and partly because the departments did not want nominees to be eliminated by the university-wide committee for these possible *red flags*. For example, when there was a noticeable discrepancy between a student's GRE scores and GPA, the departments acknowledged the need to provide a justification to overcome the perception that the candidate is less attractive to the committee selecting students for awards.

Some departments used a rubric to analyze the fellowship application materials to make the decision evidence based and to eliminate subjectivity. "Create a rubric and let the math be your guide," was the advice from the dean of a large public university involved in the university-wide committee. Depending on the size of the department and the number of admitted students, the entire committee might read all the applications and give each student a number — for example, from 1 to 5 — which would be averaged with the ratings provided by the other committee members and would result in a ranking.

Most participants indicated that strong letters of recommendation were very influential in the review process, particularly if the letters were from faculty members they knew. "The letters of recommendation can outweigh the GRE because those are people who know these students," said a faculty member in a STEM field.

Moving toward a more holistic review of the fellowship application materials seemed to be a common directive from the graduate dean's office. A dean in a medium-size public university said:

> [The GRE] can't be the only thing we look at. We have to look at the total range of the elements that we ask for and have a defensible reason for asking for those things and a defensible reason for the way that we view those things. Now that's been my crusade here. . . . It's still a long fight. We still have departments that will tell students, "You have to have x." . . . and when I know the department tells someone that, . . . I ask them to stop.

While the goal was to move toward a holistic review of the fellowship application materials, there was a fair amount of cynicism about how holistic the process was. As one participant stated, "To tell the truth, it is basically a GRE competition." This may partly be because the students' materials are often correlated strongly with one another. "The students who have the high GRE scores, have strong references." On the other hand, one professor argued that the emphasis on the GRE score in the department was not just shaped by the university committee, but by outside agencies that rank universities and faculty grant applications by the GRE scores of their students.

The issue of whether some students just do not test well was also mentioned when discussing GRE scores. One department graduate studies director said:

> What the GRE has going for it is that it's absolutely standardized and everybody is taking essentially the same test. . . . So it allows you to do a comparison. . . . [But] it bothers me to use the GRE as strongly as we do because some students don't thrive in that kind of sit-down testing environment. Some students who will be brilliant at English and at extended analysis of text and writing 20-page papers aren't so great when they are doing the fill in the circle or . . . click on the circle. It's a very different task from what we're going to ultimately be asking them to do in our department. So it's a very imperfect measure in a lot of ways and yet it's the only one that is standard.

Ultimately, a variety of additional subjective and relative qualities often determined the final list of ranked students. Discussion among the committee members weighing the pros and cons of each applicant, the needs of the department, and the interests of faculty members were important parts of the nomination process. For example, the qualitative discussion about potential fellowship recipients might take into account whether a faculty member has a strong interest in recruiting a particular student or whether a junior faculty member has adequate external funding.

The nomination process at the department level results in a ranked list of candidates, which is sent to the university committee for consideration. A letter of nomination is included with the students' materials and can be an important

factor in the university-wide committee process. One professor explained, "That is the department's chance to make a more personal sort of argument for this student as being a particularly good candidate."

Use of the GRE Scores in Department Nomination Process

In this section, we discuss in detail the use of GRE scores in the department nomination process. First, we describe the type and form of scores used; second, we discuss the semantic interpretations of the scores made by users; and, finally, we discuss the student outcomes predicted by decision makers when using the scores to nominate fellows for merit awards. Each of these three components is important to developing an interpretative argument that will guide the validation process for the use of the GRE score in fellowship-granting decisions.

Type and form of GRE scores used. All of the participating universities used the GRE scores when ranking nominees for the merit-based fellowships. The particular scores included in the fellowship application packet mirrored those required for admission to the university or department. In all cases, these included the Quantitative Reasoning score and Verbal Reasoning score, and, in some cases, the Analytical Writing score.

Analytical writing score. The universities were mixed about whether the Analytical Writing score was considered in the ranking process. More than half of the participants felt that the Analytical Writing score was not useful, and they did not consider the score in their nomination process. There were three primary areas of concern with regard to the Analytical Writing score. First, the decision makers were unclear about how to reconcile the scale of this section with the scale used for the other GRE sections; second, the participants did not understand the meaning of the levels of the scale; and, finally, the participants were not sure whether the section assessed the writing skills necessary to succeed in graduate school.

While some of the comments about the Analytical Writing section were positive, most were lukewarm or even negative. "I like the [Analytical Writing] scores, they typically line up with the verbal," said one dean. More often, the opinions were not as positive. A humanities professor and graduate studies director said:

> Well, just personally, I don't give that [Analytical Writing section] quite as much weight as the verbal, and that's only because what little I know of the grading process, it sounds like it's pretty—subjective isn't quite the word. It seems like each essay is given a pretty short amount of time with someone who's looking for a specific kind of thing. And again, I think something that gets a high score is probably a pretty good piece of writing. I'm not always certain that something that gets a low score is a bad piece of writing, just because I know how little time a grader really can spend on it, and I suspect they either explicitly or implicitly have a list of what they are looking for and they are just kind of scanning and checking off, "Okay yeah, five paragraph format, whatever, fine."

One of the concerns with the score was the broadness of the scale. One participant stated, "I couldn't really tell you what the difference is, say, between a 4 and a 4.5. Yeah, the difference between a 2 and a 4.5, yeah." An additional concern was the fit between the writing skills assessed by the measure and the types of writing skills required in graduate school, which generally involve time-consuming review and revision, rather than the ability to produce an essay in 30 min. "I would say we probably end up judging the writing skills of the students . . . more from the personal statements than from the Analytical score," said a faculty member in the humanities field.

Form of the scores used. The universities varied greatly in terms of the form of the GRE scores used. In general, many decision makers focused on the percentiles for a student's score. However, in some cases, the percentiles were averaged across the three different sections. In other instances, the percentiles were averaged with the student's GPA (converted to a percentage for comparability purposes) to get a measure of academic success.

Although composite GRE scores were occasionally used, many faculty members preferred to see the individual scores and, on a number of the rubrics, these scores were weighted differently. One department director of graduate studies stated:

> We don't have a threshold or cut-offs and we don't add the scores or sum the scores or do anything like that. In fact, if applicants say they've got a total GRE score . . . I tell them, "Give me the individual scores, because the sum score doesn't mean anything to me."

Semantic interpretations of the GRE score in making fellowship decisions: it is not enough to be sure, but it's useful

In this section, we discuss the inferences made by decision makers when using the GRE in nominating admitted students for the university-wide fellowship selection process. Understanding the characteristics that decision makers attribute to a test taker based on a score is the foundation of the validation process (Kane, 2006). The participants provided a nuanced view of the inferences made when using the GRE score in awarding fellowships. These inferences are presented in terms of the characteristics decision makers discussed as important to successful doctoral students and fellows.

Academic skills and knowledge. A prevalent theme across the universities was that the GRE is a standardized measure used to infer whether a student has the baseline level of academic skill and knowledge required to succeed in graduate school. One faculty member at a top-tier university said:

> I guess I would say that the value of the GRE in this context is that everyone recognizes that it means something. . . . at least in a kind of excellent, good, poor . . . range of categories. It's hard to fake your way into a really great score, I think. I think it's possible to not get a good score despite how smart you are for any number of reasons. But I think anybody who gets 700 s, 800 s, there is something real that they know how to do right.

In general, the score is used to infer whether the student has adequate academic preparation to begin graduate school. "Perhaps what the GRE captures is a student's experience up to a particular stage in terms of acquiring knowledge," said a STEM faculty member.

In addition, some participants stated that there was not enough variation between the admitted students' scores to make the GRE very useful in differentiating the students. The STEM faculty member added, "The bottom line is that we get very good scores, on the whole, so therefore the GREs are not, for our applicants, the primary determinant in any sense."

In the fellowship competition, high GRE scores were often interpreted as indicative of a student's preparation, thereby making the student less of a risk. As one dean explained, "The fellowship students are supposed to be no or little risk for high performance." Another dean supported this sentiment, saying:

> I think very high GRE scores that are matched with other supporting material in the application as indicators of a very strong applicant, I think that is a number that's invaluable. I mean there is a number that (indicates) a student took the same test that all these other students took, and they ended up in a better percentile. That's an indication that that student knows their stuff.

While most graduate faculty and administrators inferred a level of academic preparation from a student's GRE scores, others inferred from the score a degree of *native ability* or overall intellectual ability. However, this view of the GRE scores was less prevalent than the belief that the GRE scores demonstrate a student's acquired academic skills or knowledge.

What the participants seemed to mean with regard to a measure of *baseline knowledge* for graduate school was that, while the GRE captures the academic skills required to begin graduate school, there may be other higher order analytical skills needed to succeed that may not be evaluated by the GRE. One department head said, "The skills that really make the difference in terms of success aren't as correlated with academic background as they used to be." The participants were clear about the importance of the score in measuring student readiness, but were less certain about its ability to assess the higher-order reasoning abilities required to succeed. A graduate studies director stated, "Like I said, it [the GRE] is imperfect and there's no way to make it more perfect. I think it does a good job of capturing some of the important skills, and it is a good rough predictor."

At the same time, the participants were clear about the value of the scores in determining whether a student was ready for graduate school, particularly when the student's score was at either end of the scale. One dean said:

> I think most of our faculty would say that . . . although a high GRE score might not be indicative of success in the program, they would say a low GRE score is a high indicator of potential failure. . . . I think the people who have very high GREs . . . do end up being very good students. I think for people who have very low GREs, they sometimes are actually much better than the GRE indicates. When a person ends up crashing and burning, not really succeeding, it's very rare that person was a high GRE. It's quite common that you go back and say, "They had a very low GRE." But you will also get low GRE people who succeed here fully.

Rather than inferring from the scores a particular academic skill, some participants inferred the ability to succeed in an academic environment, which one faculty member referred to as having the "knowledge needed to play the game." A high score showed that the student had succeeded in one type of environment and might therefore succeed in a similar environment. In terms of content, particularly in the sciences, some faculty members felt there were limits to what they could infer from the GRE Quantitative Reasoning score, and a few even indicated that the score does not measure anything beyond what was known at the high school level.

Noncognitive skills. In general, the participants felt that GRE scores were limited in their ability to provide information about the noncognitive skills that they thought were critical. In many ways, it was the noncognitive skills that the participants felt were essential in determining who would succeed and who would not. This seemed to have led to a sort of pessimism about the ability to predict students' likelihood of success. One faculty member lamented, "I don't think there's anything we can do. We look at everything that is there, and there is not much else we can look at to tell us whether or not the student will succeed."

Predicting Outcomes for the Fellows

The fellowships are primarily used as recruitment tools, and enrolling admitted students is the primary outcome or consequence inferred from the decision to award a fellowship. The participants did not seem to think about the outcomes for the fellows much differently from the outcomes for the other admitted doctoral students, and they provided little anecdotal or empirical evidence that the fellows performed better in graduate school or were more successful in their professional careers. A consistent theme was, "Once the fellows are admitted, they are not much different from other students." In fact, in many cases, faculty members said they were unaware whether a student was a fellow. "The students who get the fellowships at the end—will those students be the best students in the end? No idea," said a department chair who was involved in the fellowship process.

In some ways, the major difference between the fellows and other admitted students was that they had no teaching assignments early on, which theoretically let them concentrate on their coursework and research plans. In a few instances, the participants indicated that fellows complete their degrees more quickly than other students, but this seemed to be anecdotal and might have had more to do with the students having funding, rather than any personal characteristics. Interestingly, in one instance, the faculty member believed that the fellowship students felt isolated from the rest of the university members because they had no teaching assignments, and that this hindered the students' acclimation process. To address this, the university was working to build a community among the fellows to provide support for the Ph.D. process.

If pushed, the participants talked about potential future professional outcomes for fellows, but they seemed ambivalent about this, as if unsure whether the decision-making process had been successful in identifying long-term outcomes for students. A more common theme was, "We want to give them something that makes them feel good, and gives them a little bonus. . . . It's always nice for them to be able to put on their CV that they've had this fellowship," said one professor as she talked about the complexity and long trajectory from admission to finishing the degree. "Finishing: that's the outcome we are looking for."

University-Wide Selection Process

The universities in this study used university-wide committees organized through the graduate dean's office, which selected the fellows from the pool of students nominated by the departments. The study sample is limited to universities that offer a fellowship through the graduate dean's office. The challenge for the committees is to compare students across

disciplines in order to grant a limited number of awards. A particular concern is how to find an equitable way to weight criteria that are more important to one discipline than another. In most cases, the selection committees consist of faculty from across the university. However, in a few cases, the fellowship decisions are made directly from the graduate office administration without faculty input.

Make-Up of the Committee

The committee is generally made up of faculty members across disciplines and departments and ranges in size from two or three people to up to 15 who review hundreds of nominations. In most cases, a representative of the dean's office serves on the committee to represent the university's interests and, in some cases, acts as the chair.

Pool of Students

The university-wide committee selects awardees from the pool of students nominated by the departments or schools. The students are reviewed by committee members from a range of disciplines, which are often mismatched with those of the students. This is in contrast to the department selection process, where the students are reviewed and nominated by faculty members who are familiar with the expertise and the skills needed for their particular field.

Selection Process

In general, the committee receives an application packet for each nominated student, along with rankings assigned by the departments. In most cases, the packet includes a nominating letter from the department or faculty member. This nominating letter is very important, as it gives a general impression of the student outside of the department and provides the committee with a discipline-specific perspective on the student. The committee also usually receives GRE scores (or other standardized scores relevant to the discipline), the student's statement of interest from the admission application, and copies of recommendation letters.

While the selection process varies across universities, generally, each applicant is reviewed and ranked by several members of the committee, and the rankings are then averaged. This ranking process is somewhat more holistic than the ranking at the department level, as there is less variation among the quantitative measures, such as GRE scores and GPAs, as well as other supporting documents at this point in the process. In other words, most of the nominees have strong credentials. With regard to the university-wide fellowship committees, one graduate school dean stated:

> I will tell you, the GREs have never in four years been mentioned in the process of those boards [fellowship committees]. I'm not suggesting that people don't look at them. . . . There is some pretty fine distinctions made in the award of those things. . . . But I can never recall anybody saying, "Look at that GRE score." Part of it might be because they all tend to be pretty close to each other.

A STEM professor at a highly selective university explained:

> So it ultimately becomes, there's probably five people that are really strong, and then how do you distinguish between them? If one person's GRE isn't good, he falls out of the pile and then the next four compete; and it's a very hard situation, I think.

On the other hand, the participants expressed concerns about comparing the GRE scores across disciplines, as the scores might have less weight in some departmental decisions to nominate students. The committee tries to balance this through consideration of the rankings the departments provide. When discussing the challenge of balancing credentials across the disciplines, one graduate school dean said:

> It's difficult when you're trying to do a centralized fellowship or something like that, when you're serving a lot of different kinds of customers. How do you make it fair for all groups? I think we've done a good job of being very open with our colleagues and saying, "This is important to us, that's important to you, let's find a compromise."

Still, the participants talked about the importance of having a standardized score that could be compared across disciplines. In discussing the committee's review of fellowship nominees, one professor explained:

> [The] GRE helps. If it helps tell that story, it's a nice extra. And especially in this kind of cross-campus fellowship sort of competition, where there's less shared culture about what's good or so forth, then I think it helps more. . . . It's not really because the GRE itself is taken as so valuable that it should really be making these decisions, but it works more as a common yardstick in these kinds of broader fellowship competitions than it does . . . for us internally.

For most of the universities, the selection process consists of multiple committee members ranking each applicant, followed by a committee-wide discussion. For example, in one of the largest committees, each applicant is ranked by three to four faculty members, with each committee member assigning 40 students a number between 1 and 7. In at least one university, the process involves a purely quantitative ranking at the administrative level, without input from the faculty. This seems to be in response to complaints in the past that the system was biased. In this case, the graduate administration applies a formula, which weights the students' GRE scores and undergraduate GPA, to a ranked list of all applicants nominated across the departments. "We've had cases where the funding decisions came down to a tenth of a point. They all deserved it, but we had to make a decision. We go purely by formula score," said a graduate administrator.

Using the ranking system to select students for an award without in-depth discussion is not the norm. A professor at a highly selective university described the university-wide committee process as follows:

> We all read the files and we have a kind of a scoring system, and then we add them all up and average them all among all of us to get a kind of overall ranking. Then we spend a day-long meeting basically hashing through pretty much every single case to figure out: Is this the right ranking? How many people are we trying to let in this year? What's our general funding situation? We sort of go into that meeting with some target numbers about . . . our space limit here. And so our first cut is this quantitative ranking that we've come up with due to some internal process, but that's just the first cut. . . . There's a sort of upper range where a lot of the discussion has to take place and, in that context, GREs are certainly referred to, but usually it's referred to as, "How does it fit into the overall case?"

Differences in Score Use by University Characteristics

A preliminary discussion was held with selected academic deans to suggest university characteristics that may influence how GRE scores are used in fellowship decisions. Four factors were identified: governance structure, size of the university, region, and the selectivity of the university and program. While the small sample size and limited variation in characteristics does not support drawing conclusions, the findings can provides useful insights into how university characteristics may influence score use in this context.

Governance

On the surface, the public or private governance structure was not a major factor in how the universities used GRE scores in fellowship decisions. This may be partly because there were only two private universities in the sample. However, one difference worth noting among the public and private universities in this sample was the way in which decisions about including race as a factor in award decisions had influenced their fellowship processes. This issue was raised in at least three out of the seven public universities in the sample, but was not raised explicitly in the private universities.

The public universities differed in their descriptions of how decisions regarding race as a factor in decision making influenced their fellowship process. For example, in one university, fellowships designed to help increase minority enrollment were eliminated, and all fellowships were now open to all applicants to address issues of fairness. In another instance, there was a shift to a more holistic credential review so the GRE score would not be used as a sole criterion in the award decision due to concerns that this might be unfair to minorities. On the opposite end of the spectrum, one school created fellowships specifically to address adverse racial incidents at the university.

While all of the universities expressed interest in increasing diversity, neither the public nor private universities discussed the merit-based fellowships directly as a vehicle to increase diversity. Several of the universities felt minority students who fit the criteria had many opportunities, and the fellowship offer would not be enough to persuade these talented students to their university.

Size of the University

The size of the university is a factor in the use of GRE scores in the fellowship process when size affects the internal governance or bureaucracy of the school. In the larger universities, there are many layers of schools, departments, and programs. In these instances, the deciding committee can review a hundred applications and the faculty committee can be very large. This may influence the use of the GRE as a common yardstick for committee members who do not have insights into the skills needed for a particular discipline.

Region

The region of the university, in some instances, influenced the importance of the fellowship in attracting students. For example, some universities did not feel their region or location was particularly attractive to students. The fellowship process, including a visit to the areas, was used in these cases to bring candidates to the region to show students the quality-of-life opportunities available at the university.

Selectivity of the School

As discussed earlier, the selectivity of the school and the program was the most influential way in which GRE scores were used in the fellowship process. In the highly selective programs, the variation in scores was limited, with all accepted students scoring very high. While a high score becomes a necessary condition for a fellowship applicant, the lack of variation in the scores does not support making distinctions among applicants in the award process. In addition, in some selective schools, particularly in STEM fields, the universities felt the scores did not provide evidence of the level of math and science required for success in their program, and the departments looked for other evidence of ability to rate students. On the other hand, in less selective schools, a high score may provide more useful information in differentiating between students and may hold more weight in the selection process.

Discussion

This exploratory qualitative study examined the ways in which universities use GRE scores in the decision-making process for awarding merit-based fellowships to newly admitted doctoral students. The findings reported here are based on a qualitative analysis of nine in-depth case studies on GRE use in award decisions in a sample of diverse universities across the United States. The goal of this study was to identify patterns and themes in the universities' use of scores as an initial step toward building an interpretive argument for GRE as part of an argument-based approach to validity.

In an argument-based approach to validity, the interpretive argument documents the network of arguments made by users when using scores and examines the plausibility of these arguments through theoretical and empirical evidence (Kane, 2006). As there was little systematic documentation on how universities used scores in the fellowship process, this study was designed to provide initial exploratory findings on score use. These early findings were designed to be confirmed and revised with a larger sample, and then subsequently used to build the interpretive argument as part of an argument-based approach to validity.

In general, the participants of this study viewed the merit-based fellowships as recruitment tools to incentivize *the best and the brightest* students admitted to the university to enroll. Accordingly, the primary expected outcome for the fellows was that they accepted the admission offer and enrolled. The fellowship was viewed as a tool to support the department goals while contributing to larger university goals by bringing in high-performing students. This use of the GRE score as a measure of academic skills required for entering graduate school is a plausible interpretation of the score and suggests the direct relevance of current existing literature of GRE scores as potential validity evidence for score use in this context.

Across the universities in the study sample, the selection process for the fellows consisted of two parts: a nomination process by the departments and a cross-discipline selection process by a university-wide committee. GRE scores were used in both of these two components of the fellowship-granting process.

Interestingly, the participants did not view the fellows as having different types of characteristics than the other admitted students, but rather a greater degree of the characteristics valued in all of their admitted students, making them more of a

sure bet. While the decision makers expect the fellowship nominees to have *all cylinders blasting* in the selection process, it is not quite clear what the nominees are sure bets for. In other words, the predicted long-term consequences of the decision to award a fellowship are less clear than might be expected from the attention given to the recipients' selection.

In many ways, the participants seemed to feel that once the fellows were enrolled, they were indistinguishable from the other graduate students and followed the patterns of success of all their admitted students; some did well and some did not. The participants seemed reluctant to predict outcomes beyond an initial degree of success in graduate school for the fellows selected.

This may be due to a common belief that a number of the factors that determine success in graduate school are not easily measured by GRE scores or other supporting materials. These factors certainly include noncognitive skills, such as perseverance; tenacity; and *sitzfleisch*, or "tough skin on the butt, the ability to sit and do it." However, the participants also referred to a different type of academic skill needed for the completion of a doctoral program that may not be measured or projected by the academic skills and knowledge needed at the beginning of the process.

Understanding more about the outcomes inferred when using the GRE in the fellowship-granting process will be important to building an interpretive agreement for studying the validity of the score in this context. When a score is used to make a decision, establishing validity involves examining the plausibility of the characteristics inferred to the test taker and also the outcomes inferred to occur when the score is used to make a decision.

In this case, the inferred outcomes are not limited to individuals, but include programmatic goals, such as improving the ranking of a school. It will be important to include these global, rather than individual, goals in the interpretive argument used to examine the validity of the use of the GRE score in fellowship decisions.

The next step for this project is to refine and confirm the themes and patterns identified from the qualitative analysis using an argument-based validity framework as a guide. This could be achieved through a survey administered to a larger sample of universities, using the current interview questions and findings as a guide. The findings here suggest appropriate questions that could be included in a separate survey effort or incorporated into a larger survey on GRE use.

The findings from this study take an important initial step toward building the interpretive argument for GRE score use in fellowship decisions. In this study, an argument-based approach to validity was used as a conceptual framework to provide a theoretical direction for the study and guide the research methods used. While the findings from this study are exploratory, the findings provide an initial preliminary step in the validation process for the use of GRE score awarding merit-based fellowship to first-year doctoral students.

Acknowledgments

As we flew into cities and drove down country roads, discussions of the GRE were a unifying thread across universities. We thank all the engaging academic faculty and staff for their time.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Burns, R. L. (1970). *Graduate admissions and fellowship selection policies and procedures, Part I*. Princeton, NJ: Educational Testing Service.

Burton, N. W., & Wang, M. (2005). *Predicting long-term success in graduate school: A collaborative validity study* (GRE Board Research Report No. 99-14R). Princeton, NJ: Educational Testing Service.

Chapman, G. B., & McCauley, C. (1993). Early career achievements of National Science Foundation (NSF) graduate applicants: Looking for Pygmalion and Galatea effects on NSF winners. *Journal of Applied Psychology*, *78*(5), 815–820.

Conrad, L., Trisman, D., & Miller, R. (1977). *GRE Graduate Record Examinations technical manual*. Princeton, NJ: Educational Testing Service.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. L. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

King, D. C., & Besco, R. O. (1960). The Graduate Record Examination as a selection device for graduate research fellows. *Educational and Psychological Measurement*, *20*(4), 853 – 858.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, *315*, 1080 – 1081.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*(1), 162 – 181.

Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the Graduate Record Examination for Master's and Doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement*, *70*(2), 340 – 352.

Lannholm, G. V. (1968). *Review of studies employing GRE scores in predicting success in graduate study, 1952 – 1967* (GRE Special Report No. 68-1). Princeton, NJ: Educational Testing Service.

Maxwell, J. (2005). *Qualitative research design: An interactive approach* (2nd ed.). Newbury Park, CA: Sage.

Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco, CA: Jossey-Bass.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 3 – 103). New York, NY: American Council on Education and Macmillan.

Miles, M. B., & Huberman, M. A. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Milner, M., McNeil, J. S., & King, S. W. (1984). The GRE: A question of validity in predicting performance in professional schools of social work. *Educational and Psychological Measurement*, *44*, 945 – 950.

Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the Quantitative and Verbal components of the Graduate Record Examination with grade point average representing the criterion of graduate success. *Educational and Psychological Measurement*, *55*, 309 – 316.

Ostriker, J. P., Hinshaw, V. S., Aberle, E. D., Bradburn, N., Brauman, J. I., Holland, P. W., & Wheeler, R. P. (2010). *A data-based assessment of research-doctorate programs in the United States*. Washington, DC: National Academies of Science.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Newbury Park, CA: Sage.

Schneider, L. M., & Briel, J. B. (1990). *Validity of the GRE: 1988 – 90 summary report*. Princeton, NJ: Educational Testing Service.

Sternberg, R., & Williams, W. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychologists? *American Psychologist*, *52*, 630 – 641.

Willingham, W. W. (1974). Predicting success in graduate education. *Science*, *183*(4122), 273 – 278.

Willingham, W. W. (1976). *Validity and the Graduate Record Examinations program* (GRE Board Special Report No. 76-01SR). Princeton, NJ: Educational Testing Service.

Appendix A: Recruitment Plan

Appendix B: Example of Interview Guide

Description of University's Graduate School Programs

I thought it would be useful to begin with an overview of the graduate programs offered at SCHOOL NAME.

- **Can you provide a brief overview of the graduate programs offered at your school?**

Probes
Masters vs. doctoral
STEM
Humanities
Arts and sciences
Selectivity of programs

- **How long have you been the POSITION?**

Probes
Previous appointments
Academic positions

We are interested in hearing about the students who attend your doctoral programs.

- **How would you describe the students who attend your doctoral graduate programs?**

Probes
Full-time/part-time students
Directly from undergraduate programs
Older

- Student preparation for graduate school
- Completion rates
- Differences by disciplines?
- **What would you describe as some of the benefits of attending graduate school here?**
- **What are some of the challenges?**

Graduate Students

We are interested in your idea of a good doctoral candidate for your school, based on your experience.

- **How would you describe a good doctoral candidate for your school based on your experience?**

Probes
What academic skills would the student have?
What research skills or experience?
What non-academic skills do you think are important?
Differences by discipline?

- **When you think of this student, what would they have achieved …**
  - One year in the program?
  - Five years after admission?
  - Employment after graduation?
  - At mid-career?

Scholarships/Fellowship/Assistantship

Note to Interviewee: Define which name is most appropriate and try to use throughout.

- **While our primary research focus is merit-based fellowships offered to first-year doctoral students, can you tell us the overall process for providing funding for admitted students? For example, are some students offered full scholarships, others merit-based fellowships, others assistantships, etc.?**
  - If so, how are these decisions made?
  - What is the process?
- **What scholarships or fellowships are offered to first-year doctoral students through the Graduate Dean's office?**
  - Are these merit based?
  - How many are offered a year?
  - Do the fellowships have any work or teaching requirements?
  - Are the fellowships used as a way to attract talented students?
  - What other funding opportunities are available to first-year doctoral students admitted to your programs?

- **What are the university's goals for the fellowship?**

Probe
For example: attract talented students, increase diversity, support completion, identify future scholars

- **What are you looking for in a student to whom you award a fellowship?**

Probe
Academic skill
Non-academic skills
Research experience

We talked previously about your perception of a good doctoral candidate for your doctoral programs.

- **What else are you looking for in a strong candidate for a fellowship?**

Probe
Differences by discipline

- **What would a recipient of a fellowship have achieved, in your view …**
  - One year in the program?
  - Five years after admission?
  - Employment after graduation?
  - At mid-career?

Fellowship/Assistantship-Granting Process

Can you tell us about the application and selection process for the fellowship?

- **How does this process begin?**

Probe
When do students apply?
How do students apply?
Are students recruited?

- **What is the decision-making process for the award?**
- **Is there a committee or panel?**
- **Who sits on this committee?**
- **What instructions are the decision makers given about criteria used in the decision process?**
- **Do you consider diversity?**
- **How are decisions reached?**
- **Who makes the final decisions?**
- **What do you see as strengths of your approach?**

Probe
Weaknesses

Use of GRE Scores

We are particularly interested in how GRE scores are used in the decision-making process for your fellowship.

- **Are GRE scores requested as part of the application process for the fellowship?**
- What other sources of information or criteria are requested?

Probe
Letters of recommendation
Undergraduate grades
Research experience

- Does this differ by discipline?
- **How important are GRE scores to the decision process?**
- Are there formal rules for weighing criteria?
- Is this a quantitative process?
- Does this differ by discipline?
- **What do you learn from the GRE score?**
- Are there formal rules for weighing criteria?
- Is this a quantitative process?
- Does this differ by discipline?
- **Are decision makers given specific instructions on the use of the GRE as a criterion in awarding fellowships?**
- **What do you see as the challenges and benefits of using the GRE score in this context?**
- **Are there additional thoughts you would like to share?**

Suggested citation:

Rock, J. L., & Adler, R.M. (2014). *A descriptive study of universities' use of GRE® General Test scores in awarding fellowships to first-year doctoral students* (GRE Board Research Report Report No. 14-04, ETS Research Report No. RR-14-27). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12027

Action Editor: Brent Bridgeman

This report was reviewed by the GRE Technical Advisory Committee and the Research Committee and Diversity, Equity and Inclusion Committee of the GRE Board.