*Article*

# To Test or Not to Test? The Graduate Record Examinations: Predictive Validity toward Graduate Study Success on Research Masters' Programs in a Large European University

**Anastasia Kurysheva [1,*], Gönül Dilaver [1], Christine Merie Fox [1], Harrison Kell [2,†], Matthias Robert Kern [3] and Harold V. M. van Rijen [1]**

[1]   Biomedical Sciences, Center of Education and Training, University Medical Center Utrecht, 3508 GA Utrecht, The Netherlands; g.dilaver@umcutrecht.nl (G.D.); c.m.fox@umcutrecht.nl (C.M.F.); h.v.m.vanrijen@umcutrecht.nl (H.V.M.v.R.)

[2]   Human Resources Research Organization (HumRRO), Alexandria, VA 22314, USA; hkell@humrro.org

[3]   Heidelberg Institute of Global Health, Heidelberg University Hospital, 69120 Heidelberg, Germany; matthias.kern@uni-heidelberg.de

[*]   Correspondence: a.kurysheva@umcutrecht.nl

[†]   Harrison Kell worked at Educational Testing Service Princeton, NJ 08541, USA when the study was conducted.

**Abstract:** Graduate admissions committees in Europe have a challenging task of selecting students from an increasingly large pool of candidates with diverse application files. Graduate standardized testing can ease the comparison of application files. The purpose of this study was to examine whether the Graduate Record Examinations (GRE) is predictive of several dimensions of graduate success on English-taught research masters' programs in science, technology, engineering, and mathematics at a large European university. The data from 167 masters' students were collected. Hierarchical regression analyses were conducted. All GRE scales predicted Graduate Grade Point Average. Individual GRE scales predicted internship grade and supervisors' assessments of students' research performance and the content of their research report. None of the individual GRE scales predicted supervisors' assessments of students' practical skills, but the three GRE scales taken together improved the explanatory power of the model. The structure and style of students' research reports was not predicted by the GRE. All relationships were held after accounting for socioeconomic status. Overall, the GRE appeared as a reasonable predictor of graduate study success. Both the benefits and drawbacks of the implementation of the GRE in European masters' programs are discussed, as well as the legal limitations.

**Keywords:** Graduate Record Examinations; GRE; STEM; study success; European university; research masters' programs; SES; rubrics; internship; graduate education

## 1. Introduction

In recent years, European higher education has seen several reforms. The most overarching, comprehensive, and long-lasting of these was the Bologna Process, which was introduced in 1999 to help strengthen the competitiveness and attractiveness of European higher education. The Bologna Process helped initiate the formation of the European Higher Education Area (EHEA), which aims to ensure the comparability and compatibility of higher education degrees across 49 member countries [1]. One of the main achievements of the EHEA has been the harmonization of frameworks for qualifications through the introduction of the three cycle (bachelor's/master's/PhD) degree structure. For higher education institutions (HEIs), this harmonization initiated three changes to the admissions process for graduate school applicants.

The first change was that students could transfer more easily between undergraduate (bachelor's) and graduate (master's and PhD) education not only within their own

countries, but also between countries. While this change increased student mobility across EHEA member countries, this mobility is not being evenly distributed. A majority of EHEA countries have seen more outgoing than incoming students [2]. A smaller number of countries have experienced a steady increase in incoming international students. According to the most recent Bologna Process Implementation Report [3], the Netherlands belongs to the top three EHEA countries, where incoming international students substantially prevail over outgoing students (with the UK and Denmark being the other two; see Figure 1).
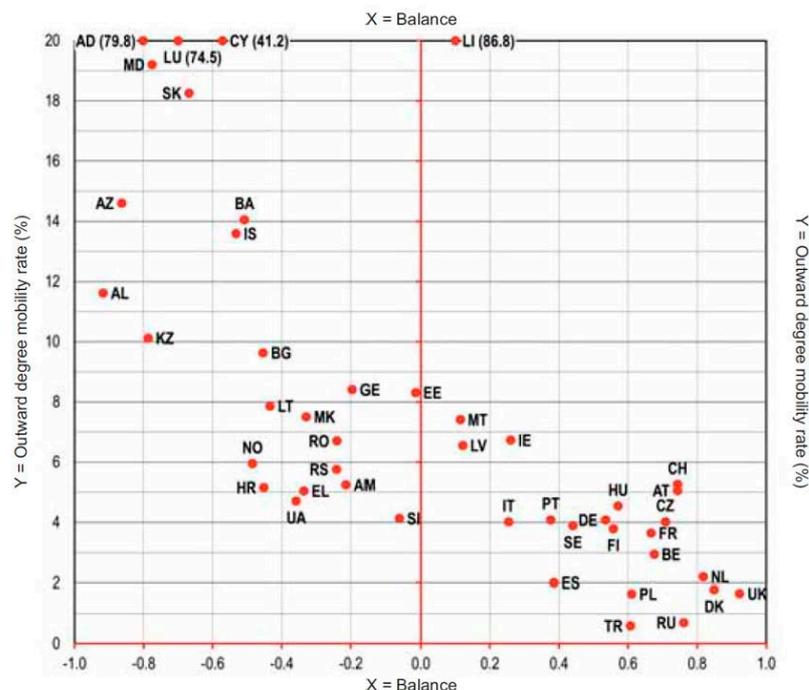


**Figure 1.** The Degree Mobility Balance in 2017, with Reference to the Outward Degree Mobility Rate of the Respective Country. © The European Higher Education Area in 2020 Bologna Process Implementation Report. *Note 1.* 1. AD−Andorra; AL−Albania; AM−Armenia; AT−Austria; AZ−Azerbaijan; BA−Bosnia and Herzegovina; BE−Belgium; BG−Bulgaria; CH−Switzerland; CY−Cyprus; CZ−The Czech Republic; DE−Germany; DK−Denmark; EE−Estonia; EL−Greece; ES−Spain; FI−Finland; FR−France; GE−Georgia; HR−Croatia; HU−Hungary; IE−Ireland; IS−Iceland; IT−Italy; KZ−Kazakhstan; LI−Liechtenstein; LT−Lithuania; LU−Luxembourg; LV−Latvia; MD−Moldova; MK−North Macedonia; MT−Malta; NL−The Netherlands; NO−Norway; PL−Poland; PT−Portugal; RO−Romania; RS−Serbia; RU−Russia; SE−Sweden; SI−Slovenia; SK−Slovakia; TR−Turkey; UA−Ukraine; UK−United Kingdom. *Note 2.* Explanation from the report on how to read the figure. "Both axes include mobility flows within and outside the EHEA: The higher the importing balance, the lesser the outward mobility rate. For graphical readability purpose, balance is computed as the absolute difference (incoming—outgoing students) divided by the total number of incoming students (when the balance is positive) or by the total number of outgoing students (in case of negative balance). This results in a smoother continuum, more readable when plotted than taking the ratio (incoming/outgoing) which is below 1 for most countries" [3] (p. 142). "The United Kingdom, Denmark, and the Netherlands are situated on the right side of the X-axis with the highest imbalance (above 82% each) and very low shares of outgoing mobile students (below 2.5%)" [3] (p. 143).

The second admissions-related change was that it became easier for students to pursue graduate degrees in related fields, but not necessarily the same field as their undergraduate studies [4]. For example, graduate programs in the life sciences started receiving applications from students with bachelors' degrees in areas such as psychology and mathematics. Finally, the third admissions-related change was that in countries with binary higher education systems, such as the Netherlands; students from universities of applied sciences

(which focus on higher professional education) were now allowed to directly apply to graduate programs at research universities (which focus on research-intense education).

While there are many benefits to these Bologna-related policy changes, there is also increased complexity due to diversity of admissions files and the absence of one common metric of comparison. The diversity of admissions files and the fact that in many HEIs the number of admissions exceeds the number of seats has resulted in a process of selection. This process of selection has two main purposes: (1) to determine which applicants are eligible in principle and (2) to rank applicants on their predicted success on a specific graduate program.

The complexity behind the student selection process, however, makes it challenging for admissions committees to understand and compare education systems and grading scales of other countries, especially of those outside of the EU. Within the EU, these comparisons are facilitated by the knowledge gained during the long-time investments in credit and grade transfer systems under the ERASMUS program—the EU's program to support, among others, mobility in higher education in Europe [5]. However, even within the EU, the comparisons of education systems and grading scales are not always straightforward and are more complicated to conduct than within one country.

The admissions committees might also not understand the quality of education at an applicant's undergraduate HEI. For example, admissions committees, using international rankings to judge the quality of universities, can misinterpret those rankings because perceived quality does not usually align with the methodologies used by the ranking committees [6]. Likewise, admissions committees might find it challenging to compare the learning outcomes of applicants' programs. Attempts for such comparisons were shown to be susceptible to assessing only the differences in knowledge, while a student's learning outcomes are greatly conditioned by the profession and discipline [7].

One possible solution to ensure the comparability of graduate school applications would be to introduce standardized admissions testing. Standardized admissions testing provides admissions committees with a metric for assessing students from diverse educational backgrounds. Therefore, potentially, they can be used to help create a fairer, more reliable, and transparent admissions process. On the other hand, the standardized testing is prone to different types of biases, and it requires financial investments (usually from the applicants), while the direct financial benefit goes to the testing companies.

There are a variety of tests for the undergraduate level. For instance, there are the Scholastic Assessment Test (SAT), American College Testing (ACT), and Advanced Placement (AP) Exams. For graduate levels, there are the Graduate Record Examinations (GRE), the Graduate Management Admissions Test (GMAT), the Law School Admissions Test (LSAT), the Pharmacy College Admissions Test (PCAT), the Medical College Admissions Test (MCAT), and the Miller Analogies Test (MAT). In this paper, we focus on graduate level, because the international mobility of students is highest at the graduate level (19% at the master's level and 31% at the doctoral level, compared to 8% at the bachelor's level). Furthermore, we concentrate on the Graduate Record Examinations (GRE) despite other standardized tests for graduate programs because it assesses a wide range of skills and is widely accepted for admissions to a broader spectrum of graduate programs compared to other graduate tests.

## 2. GRE

### 2.1. The GRE History

The GRE emerged from a collaborative project on studying outcomes of college education, which was run in the 1930s by the Carnegie Foundation for the Advancement of Teaching and four Ivy League universities (Columbia, Harvard, Princeton, and Yale; [8]). The primary aim of the project was to develop a metric for assessing the scholastic knowledge obtained by students throughout their undergraduate studies, assuming that the previous knowledge would be the best predictor of future study success [9]. As its popularity grew, it was eventually taken over in 1948 by the recently founded Education Testing

Service (ETS). Over the years, the GRE had differentiated into two types of tests: a test that measures general skills, which are not related to a specific field of study, such as critical thinking, analytical writing, and quantitative and verbal reasoning (GRE General); and achievement tests that measure knowledge in a particular field of study (GRE Subject Tests in psychology, physics, mathematics, and chemistry). This study is dedicated to the GRE General and does not address the GRE Subject Tests.

*2.2. The GRE Structure*

Since its inception, the GRE General has undergone several additions. Presently, the GRE General measures verbal reasoning (section GRE-V), quantitative reasoning (section GRE-Q), and analytical writing (section GRE-AW). Graduate school applicants must take all three sections of the GRE General (it is not possible to take one section of the test). For this study, we use the terms "GRE General" and "GRE" interchangeably.

## 3. GRE Usage in the US and Europe

In the US, testing is firmly rooted in the education system, education regulations, and expectations of stakeholders [4]. This has contributed to the widespread test taking of the GRE in the US and beyond. In 2015–2016, a milestone was reached worldwide when 584,577 GRE tests were taken (61.9% of those were in the US; [10]). These numbers have gradually decreased over the years: from 532,826 tests taken in 2018–2019 (56% of those in the US; [11]) to 341,574 tests in 2021–2022 (36% of those in the US; [12]). This decrease reflects a recent turn away from standardized testing in the US, a so-called "GRExit", as an increasing number of PhD programs across science, technology, engineering, and mathematics (STEM) fields have dropped their GRE requirement in admissions [13,14]. Notably, this trend is not specific for admissions on a graduate level. Some leading universities, such as the University of California (UC), decided to suspend undergraduate standardized testing (American College Testing (ACT)/Scholastic Aptitude Test (SAT)) until fall 2024—the time needed to develop a new undergraduate test that better aligns with the UC's expectations for incoming students [15]. Also, the COVID-19 pandemic has had a continuous impact on access to testing for applicants. Therefore, some elite higher education institutions have decided to allow applicants to submit their applications without requiring standardized test scores (e.g., Harvard College made such a decision for the classes of 2027, 2028, 2029, and 2030). However, in April 2024, *Harvard Magazine* (www.harvardmagazine.com) broke the news that Harvard, Dartmouth, Yale, and Brown had decided to reinstate the compulsory requirement for applicants to submit standardized test scores starting from fall 2025 (for the class of 2029). A year earlier, MIT had made a similar decision. In justifying their actions, the universities cited research indicating that standardized tests are predictive of students' academic success.

External admissions testing is not a widespread tradition in Europe. Most European graduate schools base their admissions decisions primarily on prior grades and other admissions tools that are indicative of a student's merits. There are, however, several graduate schools that accept the GRE. Some European business schools have developed admissions procedures similar to the US model; see the list of European business programs that accept GRE scores (https://www.ets.org/gre/consider/business/programs; accessed on 16 March 2024). There are also a handful of masters' programs in other study fields that accept the GRE. For example, in the Netherlands, these are masters' programs in economics (e.g., from Utrecht University or Tilburg University), engineering (TU Delft), mathematics (University of Twente), and health sciences (Maastricht University), which either require or make the GRE General optional for students with international bachelors' degrees. As an alternative to the GRE General, these programs usually also accept score reports on the Graduate Management Admission Test (GMAT). Despite these few exceptions, the GRE is not normally used for STEM programs in the Netherlands (and in EU countries in general). This is reflected in the open data of the ETS, where a stable 3% of the GRE test takers are from Europe, and most of those take the GRE with the intention of studying in the US and

Canada [12]. While the GRE has been broadly accepted by graduate schools in the US (and, to a smaller extent, beyond the US), there are several questions regarding its usage. The most prominent questions are (1) whether the GRE is a valid indicator of study success for STEM disciplines and (2) whether there are biases in standardized testing.

## 4. GRE as a Predictor of Graduate Study Success: Review of Literature

The empirical evidence shows that the GRE is predictive of first-year graduate grade point average (GGPA; [16–19]), total GGPA [19–24], and faculty ratings of a student's academic and professional skills [17,19,21,22]. More specifically, faculty ratings mean that several faculty members (usually two) who are familiar with a student rate the student on "(a) professional knowledge, ability to apply that knowledge, and ability to learn independently (mastery of the discipline); (b) judgment in choosing professional issues and creativity and persistence in solving the issues (professional productivity); and (c) ability to communicate what was learned (communications skills)" [20]. This is also confirmed in the meta-analytical studies [25–27].

As for other dimensions of graduate study success, scientific research has not provided a conclusive answer. For example, the primary empirical studies find no predictive validity of the GRE toward degree completion [19,28–33], with the most recent study pointing in the opposite direction and showing that scores on GRE-V and GRE-AW, but not GRE-Q, have a large effect in predicting persistence to a PhD degree versus dropout [34]. The meta-analyses, which correct for statistical artifacts of primary studies (e.g., the restriction of range of a predictor), also indicate a small but positive significant relationship between the GRE and degree completion [25,26,35]. Another interesting outcome that gained research attention was research productivity of GRE test-takers: while the primary studies do not find a relationship [19,36–38], the meta-analytical evidence points to a small positive relationship [25,26,35]. Finally, there are multidirectional findings regarding the GRE as a predictor of graduate time to degree. While some authors have not found any predictive power [19,29,33,36,38], others have found a negative relationship [22,30] and a positive relationship [39].

*Problem Statement Regarding Predictive Validity of GRE*

An overwhelming majority of the studies mentioned above were conducted within the US, and their conclusions might not be directly transferable to the European context because of differences in the structure of higher education systems. If we consider the scientific evidence available on the research aspects of graduate study success within the US higher education context, we find that those studies examined only productivity aspects of graduate students' research work, such as the number of publications, conference presentations, awards, grants, and fellowships [19,22,27,36,38]. It can be assumed that this focus on productivity relates to the fact that, in the US, master's education is often integrated into overarching graduate programs that award students with a PhD degree at the end of their studies. PhD students obviously have higher chances to have research output published or presented. This is not necessarily the case in the European higher education context and its distinct master's cycle (i.e., master's as the final degree).

In European countries, to the best of our knowledge, only two studies on the validity of the GRE exist [24,40]. They both found that GRE is predictive for graduate GGPA. In addition to that, Schwager et al. [40] found no relationship between GRE and degree completion, while Zimmermann et al. [24] detected a weak positive relationship between GRE and study progress, but no relationship between GRE and thesis performance (defined as grade obtained for their master's thesis).

The study of Zimmermann et al. [24], which examined thesis performance, might be regarded as the first attempt to examine the quality of graduate students' research work. The authors explain the finding on the absence of relationship between the GRE and thesis performance by the differences in the grading schemes for theses, used by supervisors [24].

This leaves an open question regarding the predictive validity of GRE toward students' research skills at a graduate level.

## 5. Bias in Standardized Testing: GRE and SES

Another consideration is whether standardized tests discriminate against students from certain groups, including racial/ethnic minorities, gender groups, and lower SES. Regarding racial inequality in the US context, there is evidence that applicants from underrepresented groups (African American, Puerto Rican, Native American, and Mexican American) obtain lower GRE scores than White and Asian American applicants, as well as evidence that females obtain lower GRE scores than males [41].

Racial/ethnic minorities and certain gender groups might also be susceptible to the stereotype threat: "a psychological phenomenon in which a member of a negatively stereotyped group underperforms on an activity because of increased anxiety that they may confirm the negative stereotype" [42] (p. 387). The research indicates that stereotype threat influences the test performance of minority groups in laboratory settings. For instance, Black students demonstrated lower performance on the GRE when informed that their intellectual capabilities were being assessed (in a "threat" condition) compared to a situation where Black students were informed that the test did not diagnose their cognitive abilities in a "non-threat" condition [43]. Awareness of the practices that trigger stereotypes (e.g., instructions or subtle cues that are embedded in the instructions before taking a test or formal testing environment) is seen as a solution to mitigate the effects of stereotype threat [42,44]. Stereotype threats do not affect mean score differences between minority and majority students [43]. The between-group comparison of stereotype threat has been often mischaracterized in research articles and psychology textbooks, as shown by researchers who investigated this and continued the misinterpretation of stereotype threat in cognitive testing [45,46]. In the real-life settings, however, where the testing of abilities, in essence, is a high-stakes situation, the worry level might already be so high that the stereotype threat has a negligible impact [47].

Regarding SES, critics claim that students from families with a higher SES achieve higher scores on standardized tests such as GRE on account of their access to professional test training and having more test experience [48,49]. They also argue that standardized testing places a financial burden on economically disadvantaged students [50]. An experimental study showed that low-SES participants perform worse than high-SES participants on GRE-like and IQ-like tests if a test is described as a measure of intellectual capability, but the performance of these two groups does not differ if they are presented with a test as nondiagnostic of intellectual capability [51].

Some researchers regard the weaker performance of low-SES students on standardized tests not as a problem of standardized testing itself, but as a societal issue (the lack of financial and other resources puts them at a disadvantage). Therefore, the differential performance of socioeconomic groups on standardized tests, such as the GRE, simply reflects these existing societal inequalities [26]. The meta-analytical research on standardized testing at the undergraduate level shows that even though SES is related to performance, the statistical control for SES reduces the estimated correlations between standardized tests and study success only to a small extent [52]. This means, according to Sackett et al. [52], that the relationship between standardized admissions tests and study success is not an artifact of SES. There is no comparable meta-analysis on this topic on a graduate level. However, there is some evidence from primary studies, which points to the same conclusion [40,53].

*Problem Statement Regarding GRE and SES*

In order to sustain statistical power and the focus of our study, our research concentrates on whether SES affects the results on students' GRE scores (and not ethnicity or gender). Out of the two studies conducted on the GRE in the European context, only one included SES in the model. This study found that the detected significant relationship between GRE and GGPA is independent from SES [40]. Despite this result and the

meta-analytical evidence on undergraduate level, it is important to account for SES while the debate surrounding the relationship between standardized test scores, SES, and study success is still ongoing [54].

## 6. Current Study

To expand the existing scientific evidence concerning the GRE within the European context, particularly its significance in English-taught STEM research masters' programs, we applied the principles of differential psychology [55,56] to examine the relationships between student abilities and their academic achievements. For this examination, we conducted a quantitative study that incorporated a regression analysis and the visualization of findings pertaining to predictive validity. This visualization demonstrates whether the inclusion of the GRE in the model results in an improved prediction or not.

We first examined whether the GRE is predictive above and beyond Undergraduate Grade Point Average (UGPA) toward several dimensions of graduate study success on research masters' programs. Next, we examined whether this relationship between the GRE and graduate study success still holds after SES is accounted for. Our focus on research masters' means that, in addition to widely examined GGPA as one of the dimensions of graduate study success, this study also considers other, research-related dimensions of graduate study success, such as assessments of the quality of independent research work of graduate students by their supervisors.

## 7. Materials and Methods

### 7.1. Design and Procedure of the Study

The study was approved by the Netherlands Association for Medical Education Ethical Review Board (dossier number: 2018.5.9). One hundred and sixty-seven students from a large Dutch research university volunteered to participate in the study. These students started their studies in 2019 at one of three STEM-focused graduate schools: the Graduate School of Life Sciences (GSLS), the Graduate School of Natural Sciences (GSNS), and the Graduate School of Geosciences (GSGS). As compensation for their time, they were offered a free lunch and a professional photograph for their LinkedIn profile.

The data were collected across two time points. Time point 1 was at the beginning of their two-year research master's program. The students, who had already been admitted and were thus enrolled in the programs, filled in a consent form, completed a paper-based GRE and a questionnaire on their background information during a 4-hour session. Information on study success was gathered after the students completed their masters' programs (Time point 2). The dimensions of graduate study success, which we considered, as well as predictors and control variables, are described in the "Measures" section (Section 8) below.

### 7.2. The Institutional Context

The three graduate schools that participated in this study offer research-oriented English-taught education at a master's level (two-year programs of 120 ECTS credits in total) and a PhD level (four-year programs). For this study, we focus on the master's level programs. The demand for study placement at these graduate schools increases annually. For example, there were over 1700 applicants with complete dossier in the 2020–2021 academic year at the GSLS, yet there were only 500–550 student places. The application files are diverse in these graduate schools. For example, almost half of the GSLS applicants (44%) were international students in the 2020–2021 academic year. The selection procedure is as follows. First, the eligibility of applicants is established based on their undergraduate degree. For international students, the university international admissions office assesses the extent to which their undergraduate degree corresponds to a Dutch undergraduate degree. Second, the applicants are ranked according to several aspects of their undergraduate degree, such as quality of undergraduate HEI (based on curriculum and/or international university rankings) and study success during undergraduate degree, including UGPA, grades for undergraduate courses, relevant to the content of a master's

program, as well as their motivation, extracurricular activities, and recommendations from referees. It is important to note that GRE scores are not considered during the selection process for these programs.

Once enrolled, the students receive intense research training during courses and internships. At the GSNS and GSGS, a majority of ECTS credits are earned through courses (compulsory and elective), while internships constitute a smaller part of the curriculum (typically, around 45 ECTS credits at the GSNS and 22.5–60 ECTS credits at the GSGS). At the GSLS, in contrast, a majority of ECTS credits are earned through internships (so-called major research project and minor research project or a research profile of 51 and 33 ECTS credits, respectively), while courses constitute a smaller part of the curriculum (usually, 27 ECTS credits). Nevertheless, students are recommended to take mandatory courses of at least 15 ECTS credits before taking an internship. The GSLS is also distinct from the other two graduate schools in terms of using specific rubrics for assessment of different aspects of students' performance during internships, which will be explained in detail in the section below.

## 8. Measures

### 8.1. Independent Variables

**Undergraduate Grade Point Average (UGPA).** The UGPA was obtained through the university administrative system. In cases where the UGPA was not registered, a self-reported UGPA was used instead. Since the participants had UGPAs from different higher education systems, we brought them on one scale. We converted their UGPAs on the original scales from their countries into UGPAs on a US scale from 1 to 4.

**Graduate Record Examination (GRE) General Test, Paper Version.** The GRE General Test measures student abilities that are important at the graduate level within any field of study [57]. The General Test consists of three sections: verbal reasoning (GRE-V), quantitative reasoning (GRE-Q), and analytical writing (GRE-AW). GRE-V assesses a wide range of abilities related to understanding, analyzing, and interpreting the kinds of texts commonly included in graduate schools' curricula. GRE-Q measures basic mathematical skills and the ability to answer questions using quantitative methods, and GRE-AW assesses critical thinking and analytical writing skills [58]. GRE scores are reported for GRE-V on a scale of 130–170, in 1-point increments; for GRE-Q, on a scale of 130–170, in 1-point increments; and for GRE-AW, on a scale of 0–6, in half- point increments.

### 8.2. Dependent Variables

**Graduate Grade Point Average (GGPA).** The GGPA is registered on the Dutch grading scale from 1 to 10, with 5.5 as a passing grade. At the moment of conducting the final analysis, 88% of students had completed their studies. For those students who had not finished their master's yet, a proxy of GGPA was calculated based on the components of the curriculum that they had already finished. It was chosen to calculate proxies only for those students who completed at least 15 ECTS credits of their master's program, because this is the minimal number of credits obtained for the mandatory courses before starting internships. In our sample, only four students earned as little as 15 ECTS credits. The mean was 95 credits, and the median was 103.5 credits. Therefore, most of the students were close to obtaining the required 120 credits for their master's program. There were three students who did not earn any credits and three students who earned less than 15 credits. These six students were excluded from the analysis.

**Internship Grade at the GSLS**. This variable represents a grade for the performance during the research internship of nine months and 51 ECTS credits. Research internships of such length are an obligatory requirement only at one graduate school (the GSLS) out of three participated graduate schools. These internships are typically performed in a laboratory with full immersion in research practice. During the internship, students implement a wide range of research tasks, including the design, implementation, and

reporting on a research project. The grade is given on the Dutch grading scale from 1 to 10, with 5.5 as a passing grade.

**Rubrics at the GSLS**. The GSLS applies three rubrics with the aim to facilitate the supervisors in giving comprehensive feedback to their masters' students on the quality of their performance during internships and to provide general criteria and standards for this feedback [59]. These rubrics are (a) "research skills" with three main categories: "performing research", "practical skills", and "professional attitude"; (b) "research report", with three main categories, namely "content", "structure and style", and "professional attitude"; (c) "presentation", with three main categories, namely "content", "presentation technique", and "composition and design". Each main category consists of several criteria.

The criteria of these three rubrics were developed following the learning objectives of the internship and refined during several pilots [59]. Notably, internship grades are not calculated from the supervisors' scores on rubrics. The rubrics are meant to provide justifications for the overall internship grade, but they constitute a separate assessment instrument. For this study, we chose to focus on those main categories of the rubrics, which are most closely related to the research skills of students: "performing research", "practical skills", "content" (of a research report), and "structure and style" (of a research report).

*"Performing Research" Main Category.* This category consisted of three criteria: design research plan/experiments, data analysis and interpretation, and discussion research outcomes. Each criterion is assessed following the three levels of quality descriptors: insufficient, satisfactory, and excellent. Following the approach of Postmes et al. [59], we converted the supervisors' ratings on the rubric into five scores: 1.0 for insufficient level, 1.5 when both insufficient and sufficient levels were indicated, 2.0 for sufficient level, 2.5 for both sufficient and excellent, and 3.0 for only excellent level. Cronbach's alpha of "performing research" was acceptable: $\alpha = 0.78$.

*"Practical Skills" Main Category.* This category had four criteria: technical skills, efficiency, organization lab journal/log/work records, and organization working space. The quality descriptors and their conversion into the metric scale were similar to what we described for the main category, "performing research". Cronbach's alpha of "practical skills" was acceptable: $\alpha = 0.75$.

*"Content of Research Report" Main Category*. The following criteria constitute this category: title, abstract, Layman's summary, introduction, methods section, results, tables and figures, and discussion and conclusion. The quality descriptors and their conversion into the metric scale followed the principle described above. Cronbach's alpha of "content of research report" was good: $\alpha = 0.83$.

*"Structure and Style" Main Category.* This category consisted of three criteria: structure and line of reasoning, referencing, and writing skills. The quality descriptors were the same as for other categories, as well as their conversion into the metric scale. Cronbach's alphas of "structure and style" category was acceptable, $\alpha = 0.72$.

We did not collect data on "graduate time to degree", because the duration of studies of these students would not be comparable to a typical duration that one would have observed before the COVID-19 pandemic.

### 8.3. Control Variables

**Socioeconomic Status (SES).** Students were asked to indicate the level of the education of each of their parents. The highest level of parental education of either parent was derived as an indicator of SES (following Gooding [60] and OECD [61]). The scale was then transformed as closely as possible to International Standard Classification of Education (ISCED) and European Qualifications Framework (EQF), resulting in the following ordinal scale: 1= less than secondary education, 2 = upper secondary education, 3 = short-cycle tertiary education, 4 = bachelor's or equivalent level, 5 = master's or equivalent level, and 6 = doctoral or equivalent level. We also gathered information on other indices of students' SES, such as yearly family income, parental occupation, and self-perceived SES.

**Number of Graduate ECTS Credits.** To control for the fact that not all students earned the same amount of ECTS credits during their graduate program, we included the number of ECTS credits as a control variable in the last step of the analysis. This allowed us to adjust for the possibility that students with a lower number of ECTS credits after 26 months are those who take a longer time to earn higher GGPAs and vice versa.

## 9. A Priori Power Analysis

We conducted an a priori power analysis with five predictors: the UGPA, three scales of the GRE, and the SES. We based the calculation on the effect sizes found in two studies in the European context [24,40]. When predicting GGPA, the first study found an effect of 0.14 on a sample of 282 students, and the second study found an effect of 0.19 on a sample of 369 students. Based on this information, we conducted the power calculations for our study (Table 1). As follows from Table 1, to detect a small-to-medium effect, our sample size must be above 154. After the announcement of this study, 234 students signed up to participate; 167 students appeared at the testing event.

**Table 1.** A priori power analysis.

| Power | Interpretation of Effect Size | Effect Size ($R^2$) | Effect Size ($f^2$) | Sample Size |
|---|---|---|---|---|
| 0.95 | Medium | 0.13 | 0.15 | 139 |
| 0.80 | Medium | 0.13 | 0.15 | 92 |
| 0.95 | Small–medium | 0.08 | 0.09 | 234 |
| 0.80 | Small–medium | 0.08 | 0.09 | 154 |
| 0.95 | Small | 0.02 | 0.02 | 975 |
| 0.80 | Small | 0.02 | 0.02 | 635 |

Note. There are various metrics to measure effect size, with $R^2$ and $f^2$ being prominent ones. $R^2$ is for the magnitude of the effect of the entire model, aiding in understanding how much variance in the dependent variable is explained by the model. $f^2$ is for the magnitude of the effect of the entire regression model or individual predictors; it allows us to focus on individual predictors.

## 10. Participants

We collected data on 167 students from three graduate schools at the start and at the end of their master's program. After the exclusion of 6 students who earned less than 15 ECTS credits within two and a half years of being enrolled in their master's program, we formed a sample of 161 students (Sample 1). On this sample, we tested the predictive validity of the GRE above UGPA toward one dimension of graduate study success, namely GGPA. For the analysis of other dimensions of graduate study success (internship grade and rubrics' main categories), we formed a subsample of the original sample (Sample 2), which consisted of 84 GSLS students who completed their nine-month internship, meaning that their grades and assessments on rubrics were available as well. The characteristics of each sample are provided in Table 2.

**Table 2.** Samples' demographical, socioeconomical, and educational characteristics.

| Characteristics | Sample 1 (N = 161) | Sample 2 (N = 84) |
|---|---|---|
| Gender: females | 53% | 67% |
| Age | *M* = 23.1 *SD* = 2.4 Median = 22.0 | *M* = 22.8 *SD* = 2.0 Median = 22.0 |
| GSLS | n = 93 | n = 84 |
| GSNS | n = 54 | 0 |
| GSGS | n = 14 | 0 |

**Table 2.** *Cont.*

| Characteristics | Sample 1 (N = 161) | Sample 2 (N = 84) |
|---|---|---|
| Country of origin Netherlands | 66% | 61% |
| Other EU | 19% | 26% |
| Outside of EU | 14% | 13% |
| Students with at least one parent with a higher education degree | 69% | 65% |
| Parental income | | |
| No knowledge of it | 36% | 42% |
| Less than EUR 25,000 | 11% | 8% |
| EUR 25,000–EUR 50,000 | 16% | 17% |
| EUR 50,000–EUR 75,000 | 14% | 8% |
| EUR 75,000–EUR 100,000 | 8% | 11% |
| EUR 100,000–EUR 150,000 | 9% | 12% |
| More than EUR 150,000 | 6% | 2% |
| Self-perceived SES on a scale from 1 (low) to 10 (high) | $M = 6.7$ $SD = 1.5$ Median = 7.0 Range = 3–10 | $M = 6.7$ $SD = 1.5$ Median = 7.0 Range = 3–10 |
| Previous HEI | | |
| The same research university they enrolled for their masters' | 45% | 45% |
| Another research university in the Netherlands | 18% | 21% |
| University of applied sciences in the Netherlands | 5% | 1% |
| HEI outside of the Netherlands | 24% | 25% |
| Other type of HEI | 8% | 7% |
| Missingness (values) | 0.2% | 4.2% |

Note: GSLS—Graduate School of Life Sciences; GSNS—Graduate School of Natural Sciences; GSGS—Graduate School of Geosciences; EU—European Union; SES—socioeconomic status; HEI—higher education institution.

## 11. Data Analysis

The regression analyses were conducted in SPSS version 26. Missing values were handled by using the expectation maximization (EM) technique, because the percentage of missingness was no larger than 5%. First, the hierarchical regression analyses were used. We had six outcome variables. The model with the first outcome variable (GGPA) was analyzed, using Sample 1. The models with the other five outcome variables (Internship grade and four categories of rubrics) were analyzed using Sample 2. In the first step of each regression, we included UGPA. In the second step, we added three sections of the GRE and in the third step, we controlled for SES. The analysis with GGPA as an outcome variable had an additional fourth step, in which we controlled for the number of graduate ECTS credits.

We set an alpha level of 0.10 and considered it as marginal significance because, according to our a priori power analysis (Table 1), the sample size of Sample 1 (n = 167) is not big enough to detect small effect sizes under α-level 0.05, and the sample size of Sample 2 (n = 84) is not big enough to detect small and small-to-medium effect sizes under α-level 0.05. In other words, we were willing to increase the Type I error (we take a slightly higher risk to detect an effect, which is not there) in order to decrease the Type II error (we take a lower risk of missing an effect, which is there). We made this decision because of the exploratory nature of this study toward the dimensions of graduate study success, related to quality of independent research work of graduate students.

*Interpretation of the Figures*

We supplemented the regression analysis by presenting the figures (Figures 2–7), two per each outcome (i.e., per dimension of graduate study success). The first type of figures (with an extension 0.1 at the end, namely Figure 2(1), Figure 3(1), Figure 4(1), Figure 5(1),

Figure 6(1) and Figure 7(1)) allowed us to contrast the predicted score of each individual on outcome (X-axis) against the actual score of this individual on the outcome (Y-axis). On the left side of each figure, the predictions (X-axis) were based only on UGPA. On the right side of each figure, the predictions were based on UGPA and three GRE scales. If the predictions had completely corresponded to the actual values, we could expect all data points to have the same score on the X-axis as on the Y-axis.
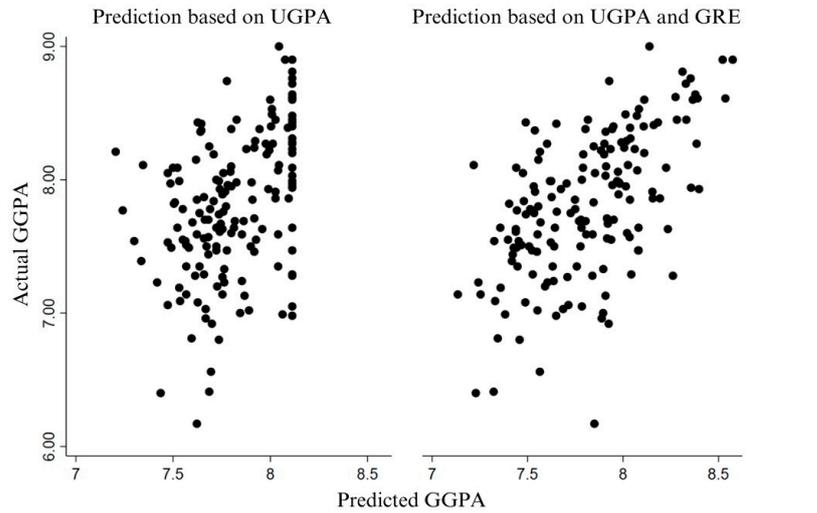


(**1**)
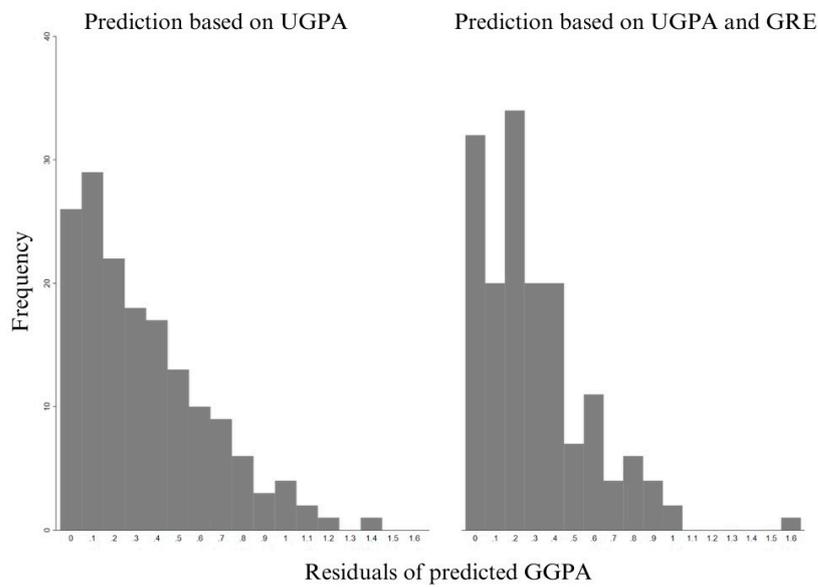


(**2**)

**Figure 2.** (**1**) Predicted individual scores versus actual scores of GGPA. (**2**) The absolute value of difference between predicted and observed scores of GGPA.
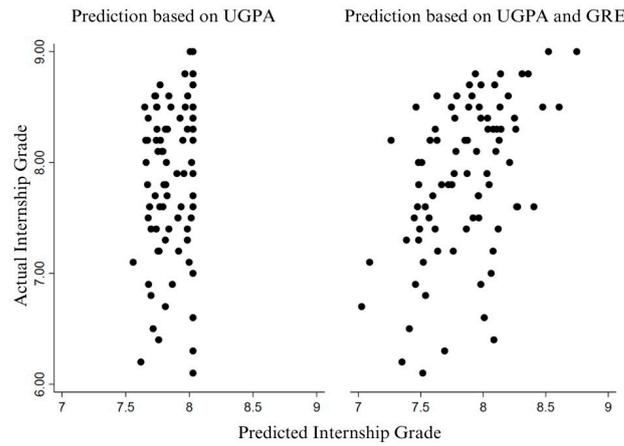
(1)
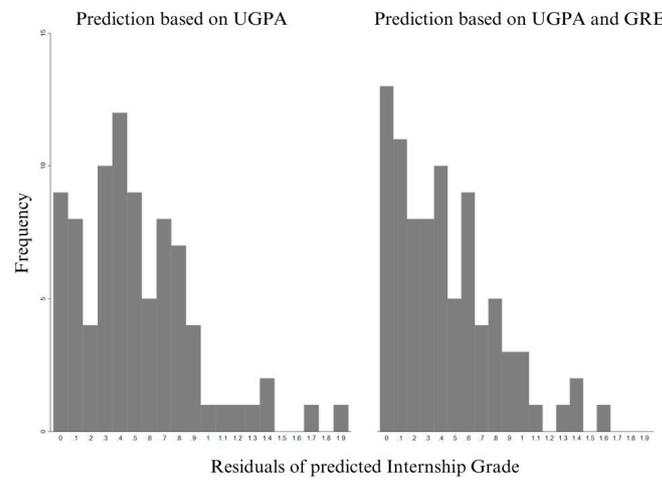


(2)

**Figure 3.** (**1**) Predicted individual scores versus actual scores of internship grade. (**2**) The absolute value of difference between predicted and observed scores of internship grade.
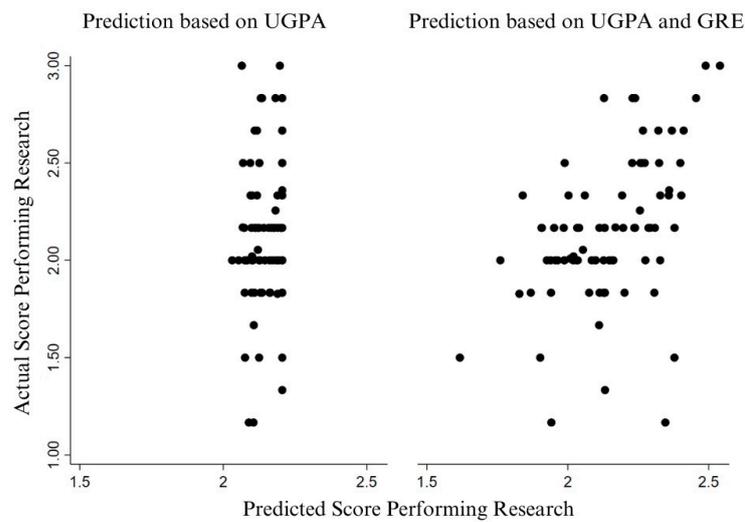


(1)

**Figure 4.** *Cont.*
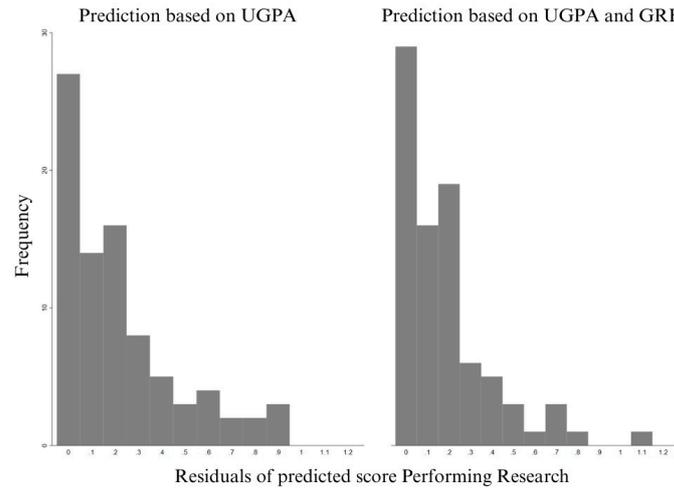
(**2**)

**Figure 4.** (**1**) Predicted individual scores versus actual scores of "Performing Research" category of "Research Skills" rubric. (**2**) The absolute value of difference between predicted and observed Scores of "Performing Research" category of "Research Skills" rubric.
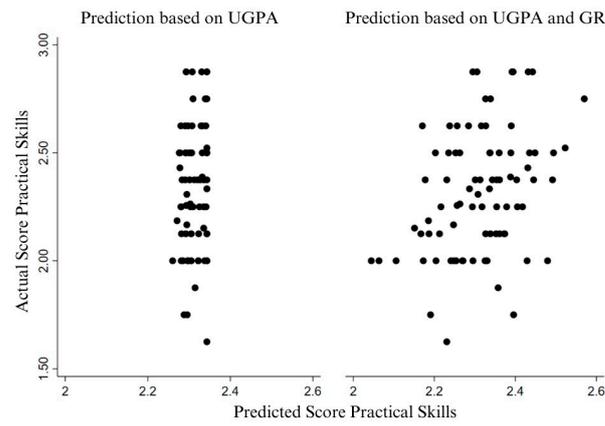


(**1**)
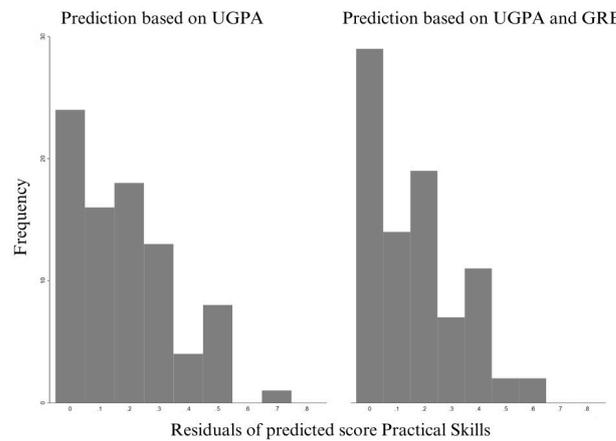


(**2**)

**Figure 5.** (**1**) Predicted individual scores versus actual scores of "Practical Skills" category of "Research Skills" rubric. (**2**) The absolute value of difference between predicted and observed scores of "Practical Skills" category of "Research Skills" rubric.
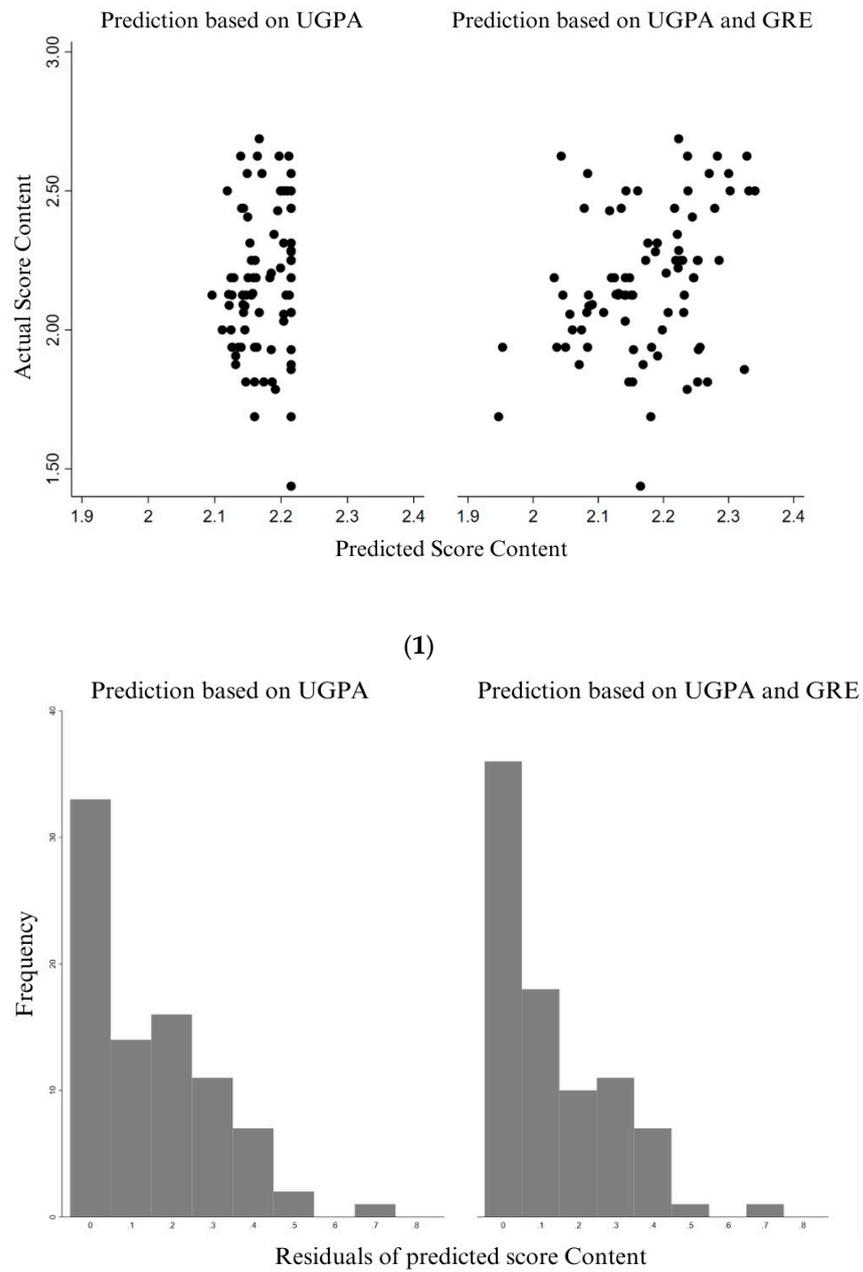
(**1**)



(**2**)

**Figure 6.** (**1**) Predicted individual scores versus actual scores of "Content" category of "Research Report" rubric. (**2**) The absolute value of difference between predicted and observed scores of "Content" category of "Research Report" rubric.
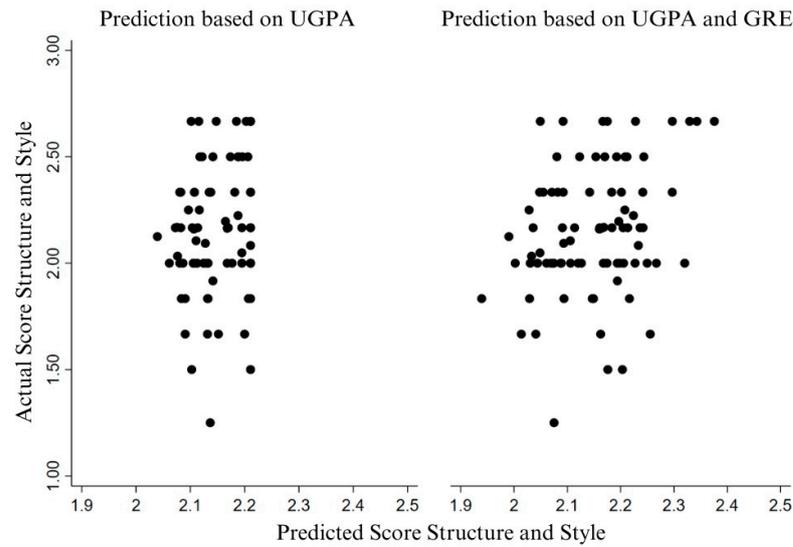
**(1)**
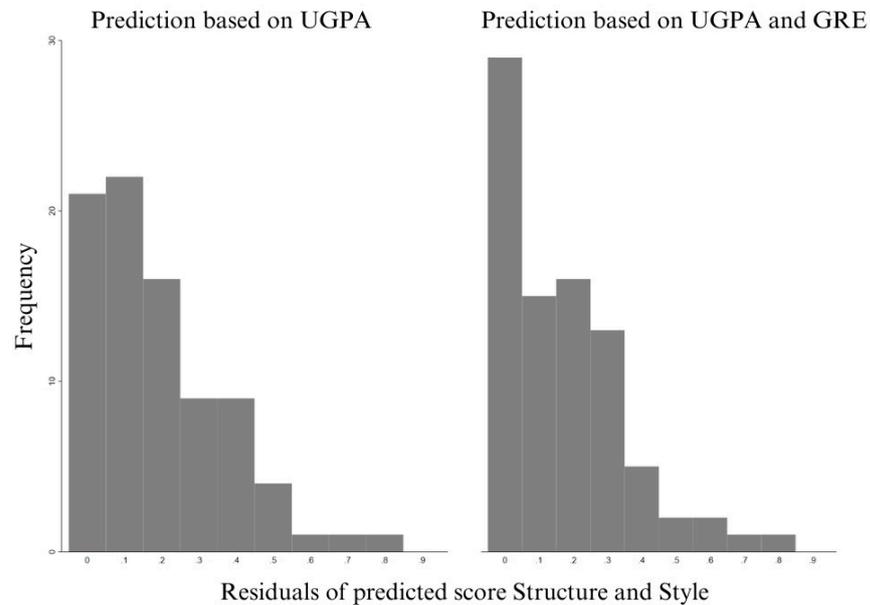


**(2)**

**Figure 7.** (**1**) Predicted individual scores versus actual scores of "Structure and Style" category of "Research Report" rubric. (**2**) The absolute value of difference between predicted and observed scores of "Structure and Style" category of "Research Report" rubric.

The second type of figures (with an extension 0.2 at the end, namely Figure 2(2), Figure 3(2), Figure 4(2), Figure 5(2), Figure 6(2) and Figure 7(2)) was designed to show histograms of the absolute values of residuals (differences between predicted and actual scores). Similar to the first type of figures, the left side of a figure (a model with only UGPA as a predictor) was designed for comparison with its right side (a model with both UGPA and GRE). We suggest to the reader to interpret the figures as follows: the better prediction model has more cases with residuals being close to zero; the worse prediction model has less cases with residuals being close to zero. Overall, both types of figures visualized what predictive validity of GRE means for each of the examined outcomes. The graphs were designed in Stata.

## 12. Results

### 12.1. Predicting GGPA

The analyses presented below are conducted on Sample 1: the full sample of students from three STEM graduate schools. Table 3 shows an overview of bivariate correlations between the study variables of Sample 1. The scores on all GRE sections are positively intercorrelated. Also, GRE-V and GRE-AW (but not GRE-Q) are correlated with UGPA. Three sections of the GRE, UGPA, and ECTS credits achieved during graduate studies are all positively related to GGPA. We also note that SES is positively related to GRE-V and, on a marginal level, to GGPA.

**Table 3.** Intercorrelations between study variables of Sample 1.

| Variable | n | M | SD | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. GRE-V | 161 | 151.83 | 6.27 | 1 | | | | | | |
| 2. GRE-Q | 161 | 153.91 | 6.86 | 0.42 *** | 1 | | | | | |
| 3. GRE-AW | 161 | 3.57 | 0.57 | 0.43 *** | 0.28 *** | 1 | | | | |
| 4. UGPA | 159 | 3.33 | 0.50 | 0.31 *** | 0.13 | 0.33 *** | 1 | | | |
| 5. GGPA | 161 | 7.81 | 0.54 | 0.46 *** | 0.33 *** | 0.40 *** | 0.42 *** | 1 | | |
| 6. SES | 161 | 4.15 | 1.32 | 0.16 * | 0.10 | −0.01 | 0.08 | 0.11 | 1 | |
| 7. Graduate ECTS credits | 161 | 115.98 | 24.20 | −0.07 | 0.05 | 0.04 | 0.05 | 0.16 * | 0.06 | 1 |

Note. * $p < 0.05$. *** $p < 0.001$.

### 12.2. Regression Analysis of Sample 1

The results for incremental validity of the GRE scores are presented in Table 4. The GRE sections predict beyond and above what UGPA predicts in GGPA: Each of three GRE sections is positively related to GGPA (see Step 2). Namely, with each additional score on GRE-V, GRE-Q, and GRE-AW, GGPA increases with 0.02 (B = 0.02, SE = 0.01, t = 2.95, $p = 0.004$), 0.01 (B = 0.01, SE = 0.01, t = 1.95, $p = 0.053$), and 0.16 (B = 0.16, SE = 0.07, t = 2.21, $p = 0.029$) respectively. Altogether, three GRE sections add 16% of explained variance in GGPA to what has already been explained by UGPA.

**Table 4.** Hierarchical multiple regression analyses predicting GGPA.

| Predictor | GGPA | |
|---|---|---|
| | $\Delta R^2$ | β |
| Step 1 | 0.18 *** | |
| UGPA | | 0.42 *** |
| Step 2 | 0.16 *** | |
| UGPA | 0.16 *** | 0.27 *** |
| GRE-V | | 0.24 ** |
| GRE-Q | | 0.14 † |
| GRE-AW | | 0.17 * |
| Step 3 | <0.01 | |
| UGPA | | 0.27 *** |
| GRE-V | | 0.23 * |
| GRE-Q | | 0.14 † |
| GRE-AW | | 0.17 * |
| SES | | 0.04 |
| Step 4 | 0.02 | |
| UGPA | | 0.26 *** |
| GRE-V | | 0.25 ** |
| GRE-Q | | 0.13 † |
| GRE-AW | | 0.16 * |
| SES | | 0.03 |

**Table 4.** *Cont.*

| Predictor | GGPA | |
|---|---|---|
| | $\Delta R^2$ | β |
| Graduate ECTS credits | | 0.15 * |
| Total $R^2$ | 0.36 | |

Note. <sup>†</sup> $p < 0.10$, * $p < 0.05$. ** $p < 0.01$, *** $p < 0.001$. n = 161. UGPA—Undergraduate Grade Point Average; GRE-V—Graduate Record Examinations Verbal Reasoning; GRE-Q—Graduate Record Examinations Quantitative Reasoning; GRE-AW—Graduate Record Examinations Analytical Writing; SES—socioeconomic status, operationalized as the highest level of parental educational achievement. $\Delta R^2$ is a statistical measure that indicates the change in $R^2$ between two models in hierarchical regression. The total $R^2$ represents the proportion of variance in the dependent variable explained by all predictor variables included in the final model. β is a standardized regression coefficient associated with each predictor variable in a model.

To illustrate the findings, let us examine the figures. Figure 2(1) visualizes the findings that data points are predicted better when both UGPA and GRE are included in the model. The data points have more similar scores on the right side of the Figure 2(1), which speaks for the additional value of the GRE scores in the prediction. Figure 2(2) illustrates that once GRE is included together with UGPA into the model (the right side of the figure), more individual cases have smaller residuals.

Finally, we find that even after the addition of SES (Table 4, Step 3) and ECTS credits (Table 4, Step 4), the detected positive relationships between GRE sections and GGPA hold, and these additions explain the negligible amount of variance.

## 13. Predicting Five Internship-Related Dimensions of Graduate Study Success

The analyses that we present below are based on 84 GSLS students (Sample 2), all of whom had completed an internship of 51 ECTS credits during nine months. In addition to the grade, they received supervisors' ratings on their internship performance within rubrics. Table 5 shows the bivariate correlations between the study variables of Sample 2. The three GRE sections are significantly intercorrelated, and all correlations are positive. GRE-V and GRE-AW (but not GRE-Q) have significant positive correlations with UGPA. GRE-V is the only GRE scale that has significant positive correlations with all five examined dimensions of graduate study success. GRE-Q is correlated with four examined dimensions of graduate study success (except the "content" rubric), and GRE-AW is correlated with three dimensions of graduate study success. SES (defined as the highest education level of either parent) has only one marginally significant correlation: with the rubric "content".

**Table 5.** Intercorrelations between study variables of Sample 2.

| Variable | n | M | SD | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. GRE-V | 84 | 151.02 | 6.02 | 1 | | | | | | | | | |
| 2. GRE-Q | 84 | 151.95 | 6.21 | 0.48 *** | 1 | | | | | | | | |
| 3. GRE-AW | 84 | 3.60 | 0.60 | 0.45 *** | 0.37 *** | 1 | | | | | | | |
| 4. UGPA | 84 | 3.39 | 0.49 | 0.27 * | 0.15 | 0.22 * | 1 | | | | | | |
| 5. Internship grade | 84 | 7.86 | 0.70 | 0.43 *** | 0.38 *** | 0.31 ** | 0.20 <sup>†</sup> | 1 | | | | | |
| 6. Rubric "Performing research" | 77 | 2.15 | 0.38 | 0.47 *** | 0.19 <sup>†</sup> | 0.17 | 0.14 | 0.73 *** | 1 | | | | |
| 7. Rubric "Practical skills" | 76 | 2.31 | 0.30 | 0.28 * | 0.26 * | 0.23 * | 0.08 | 0.70 *** | 0.60 *** | 1 | | | |
| 8. Rubric "Content" | 76 | 2.18 | 0.27 | 0.30 ** | 0.08 | 0.06 | 0.13 | 0.61 *** | 0.62 *** | 0.62 *** | 1 | | |
| 9. Rubric "Structure and Style" | 75 | 2.15 | 0.32 | 0.26 * | 0.20 <sup>†</sup> | 0.27 * | 0.17 | 0.54 *** | 0.35 ** | 0.44 *** | 0.73 *** | 1 | |
| 10. SES | 84 | 4.19 | 1.36 | 0.09 | −0.03 | 0.02 | 0.12 | 0.17 | 0.18 | 0.15 | 0.19 <sup>†</sup> | 0.15 | 1 |

Note. <sup>†</sup> $p < 0.10$, * $p < 0.05$. ** $p < 0.01$, *** $p < 0.001$.

## 14. Regression Analysis of Sample 2

Table 6 shows the gain in explained variance after the addition of three GRE sections into the models with UGPA. Each model is related to one dimension of graduate study success. Table 6 also provides regression coefficients of each predictor included in the

model. As the last step in this table, the results for the models after the addition of SES are presented.

**Table 6.** Hierarchical multiple regression analyses predicting internship grade and scales of internship rubrics.

| Predictor | Internship Grade | | Scale "Performing Research" | | Scale "Practical Skills" | | Scale "Content" | | Scale "Structure and Style" | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\Delta R^2$ | β | $\Delta R^2$ | β | $\Delta R^2$ | β | $\Delta R^2$ | β | $\Delta R^2$ | β |
| **Step 1** | 0.04 † | | 0.02 | | 0.01 | | 0.02 | | 0.03 | |
| UGPA | | 0.20 † | | 0.14 | | 0.08 | | 0.14 | | 0.17 |
| **Step 2** | 0.20 *** | | 0.22 *** | | 0.12 * | | 0.09 * | | 0.06 | |
| UGPA | | 0.07 | | 0.02 | | −0.01 | | 0.07 | | 0.10 |
| GRE-V | | 0.26 * | | 0.51 *** | | 0.17 | | 0.37 ** | | 0.20 |
| GRE-Q | | 0.21 † | | −0.03 | | 0.16 | | −0.08 | | 0.10 |
| GRE-AW | | 0.10 | | −0.04 | | 0.11 | | −0.09 | | −0.01 |
| **Step 3** | 0.02 | 0.06 | 0.02 | <0.01 | 0.02 | | 0.02 | | 0.01 | |
| UGPA | | 0.25 * | | 0.50 *** | | −0.02 | | 0.06 | | 0.09 |
| GRE-V | | 0.22 † | | −0.02 | | 0.16 | | 0.35 ** | | 0.18 |
| GRE-Q | | 0.11 | | −0.04 | | 0.17 | | −0.07 | | 0.11 |
| GRE-AW | | 0.15 | | 0.12 | | 0.11 | | −0.09 | | −00.01 |
| SES | | | | | | 0.12 | | 0.14 | | 0.12 |
| Total $R^2$ | 0.26 | | 0.25 | | 0.14 | | 0.13 | | 0.10 | |

Note: † $p < 0.10$, * $p < 0.05$. ** $p < 0.01$, and *** $p < 0.001$; n = 84. UGPA—Undergraduate Grade Point Average; GRE-V—Graduate Record Examinations Verbal Reasoning; GRE-Q—Graduate Record Examinations Quantitative Reasoning; GRE-AW—Graduate Record Examinations Analytical Writing; SES—socioeconomic status, operationalized as the highest level of parental educational achievement. For the explanation of statistical symbols, see the note in Table 4.

## 14.1. Predicting Internship Grade

The GRE sections predict above and beyond what UGPA predicts in internship grade: They add 20% of explained variance in internship grade to what has already been explained by UGPA. GRE-V was found to be a significant positive predictor: As GRE-V increases with one score, our model predicts that internship grade will increase with 0.03 score (B = 0.03, SE = 0.01, t = 2.17, $p = 0.033$). GRE-Q is marginally significant: As it increases with one score, the internship grade increases with 0.02 score (B = 0.02, SE = 0.01, t = 1.80, $p = 0.075$). GRE-AW was not a significant predictor. Figure 3(1,2) illustrate the improvements to which the addition of GRE scales lead: More cases are better predicted once GRE scales are in the model. The addition of SES does not chance this relationship.

## 14.2. Predicting the Supervisors' Scores on "Performing Research" Category

The addition of three GRE sections to UGPA in the model significantly improves the predictive power of the model: it adds 22% to the explained variance in the supervisors' scores on "performing research". GRE-V was found to be the only predictor: With each additional score on GRE-V, the supervisors' ratings on "performing research" category increased with 0.03 (B = 0.03, SE = 0.01, t = 4.21, $p < 0.001$). Two other GRE scales were not significant. Figure 4(1,2) visualize the better predictability once UGPA and GRE are both in the model. The addition of SES to the model did not change the detected relationships.

## 14.3. Predicting the Supervisors' Scores on "Practical Skills" Category

The addition of the three GRE sections significantly improves the prediction of supervisors' ratings on the "practical skills" category in terms of explained variance, as it adds 12% of it. Figure 5(1,2) illustrate these modest improvements. However, none of the three scales reaches the chosen alpha level. These findings do not change after accounting for SES.

*14.4. Predicting the Supervisors' Scores on "Content" Category*

The GRE adds 9% in explained variance to what has already been explained by UGPA. GRE-V appears to be the scale that contributes to this improvement: With each additional score on GRE-V, the supervisors' ratings on "content" category increased with 0.02 (B = 0.02, SE = 0.01, t = 2.82, *p* = 0.006). Visual inspection of Figure 6(1,2) is in line with these findings.

*14.5. Predicting the Supervisors' Scores on "Structure and Style" Category*

The three sections of the GRE marginally improved the model with only UGPA (the improvement of 6% in explained variance). See Figure 7(1,2) for an illustration. However, none of the three GRE scales reached the chosen alpha level.

## 15. Discussion

We conducted this study in a large European university, involving participants from more than 20 research masters' programs in STEM-related fields across three graduate schools, to ensure that our findings are relevant and generalizable to programs at universities of significant size and scope throughout Europe. Overall, this study showed that the GRE is a meaningful predictor of graduate study success for STEM research masters' programs, and the detected relationships held after accounting for SES. Importantly, this study was conducted on a sample of students, who were not selected based on GRE scores, and thus the range of the GRE scores in our sample was wide and comparable to the range of future potential applicants in HEIs without a pre-set threshold.

*15.1. GRE as a Predictor of GGPA: The Role of SES in This Relationship*

Our result that the GRE is predictive of GGPA beyond and above UGPA replicated a widespread finding on a positive relation of the GRE toward GGPA in STEM disciplines [19,23,24], confirmed also by meta-analytical findings [27]. We also showed that the relationship between three GRE sections and GGPA did not significantly drop once SES was accounted for.

To place this finding in the context of the existing literature on the topic, we followed the approach of Schwager et al. [40] and computed partial correlations. Namely, we compared partial correlations between the GRE sections and GGPA after controlling for SES, to the zero-order (Pearson) correlations, as reported in Table 4. What we found is that, after controlling for SES, the correlations between GRE and GGPA stayed effectively the same: partial $r_{GRE-V-GGPA} = 0.45$, $p < 0.001$ instead of Pearson correlation $r = r_{GRE-V-GGPA} = 0.46$, $p < .001$; partial $r_{GRE-Q-GGPA} = 0.32$, $p < 0.001$ instead of a Pearson correlation $r = r_{GRE-Q-GGPA} = 0.33$, $p < 0.001$; and partial $r_{GRE-AW-GGPA} = 0.40$, $p < 0.001$ identical to a Pearson correlation $r_{GRE-AW-GGPA} = 0.40$, $p < 0.001$. These findings are in line with those of the primary study of Schwager et al. [40] on students from taught masters' programs at a Dutch university, as well as with the meta-analysis of Sackett et al. [52] on the relationship between standardized undergraduate test (Scholastic Aptitude Test; SAT) and UGPA. In both studies, the reductions in correlations between scores on standardized tests and GPA after controlling for SES ranged from negligible to small. Hence, our findings replicate the findings on both the undergraduate and graduate level.

*15.2. GRE as a Predictor of Dimensions of Graduate Study Success, Related to Independent Research: The Role of SES in These Relationships*

In addition to examining the grades for research work (i.e., internship grade), we also examined the rubric categories as dimensions of graduate study success. This allowed us to shed light on whether the GRE has any relation to various quality aspects of students' research work. We found evidence that the GRE had incremental predictive value above and beyond UGPA in predicting not only internship grades but also supervisors' ratings on "performing research" rubric, and a bit less toward "content" rubric. The individual scales of the GRE did not have incremental predictive value toward supervisors' ratings on rubric category "practical skills". However, all three scales taken together added a small amount

of explained variance to the model. The GRE did not add to explanation of variance in supervisors' ratings on rubric category "structure and style".

To the best of our knowledge, this is the first study to examine these questions, and, therefore, we do not have prior literature to relate to our findings. However, Zimmermann et al. [24] looked at the relationship between the GRE scales and thesis grades on masters' programs at a Swiss university and did not find one. The authors of that study see the possibility that "the grade earned for the thesis is strongly influenced by variations in the grading-schemes used among academic supervisors. That is, the grades assigned by different supervisors might stand for quite different achievements and, thus, are not comparable" [24] (p. 19). This is an example of an attenuating factor, which refers to when the unreliability of a study success measure reduces the estimate for a relationship between standardized tests and study success [25]. In our study, the rubric categories, consisting of items on specific aspects of students' research work, were used by all supervisors to provide justification for internship grades. We assume that this contributed to acceptable reliability in the use of these rubrics' categories and to the fact that the relationships between individual GRE scales and research-related dimensions of graduate study success were detected.

Importantly, not all GRE scales were linked to the research-related dimensions of graduate study success. GRE-V was the scale which significantly related to supervisors' ratings on "internship grade", "performing research", and "content". GRE-Q had a marginal significance once predicting "internship grade". GRE-AW did not predict any of the research-related dimensions of graduate study success. It could be that our sample size of 84 students was not big enough, and we missed the possible significant relationships of small or small-to-medium size. It could also be that only one or two and not all three GRE scales are related to the examined dimensions. Future research could explore what aspects on the GRE scales are most indicative of research-related dimensions of graduate study success.

We also found that these described relationships between GRE scales and research-related dimensions of graduate study success hold after accounting for SES. Like with GGPA, we computed partial correlations between three GRE sections and five examined research-related dimensions of graduate study success, controlled for SES. We compared these partial correlations to Pearson uncorrected zero-order correlations and revealed that they were practically the same. This, together with the negligible amount of added explained variance after inclusion of SES in our models, indicated that the detected relationships between the GRE sections and internship grade and supervisors' ratings on "performing research" and "content of research report" were not artifacts of SES.

## 16. Strengths of the Study and Future Directions of Research

The first strength of this study is the variability in the GRE scores: On each GRE scale, we had a wide variety of scores, especially on GRE-V (from 133 to 167) and GRE-Q (from 134 to 170), corresponding almost to the whole range of possible scores on these scales. This is in contrast to many studies in this field, which were often conducted on students who were already selected for admissions based on the GRE, and this restricted range in the scores on the predictor might have affected their findings, e.g., [28,37]. The wide range of scores in GRE allows us to transfer our findings more easily to the population of our interest: the diverse pool of applicants to research masters' programs, who will naturally have a potentially wide range of GRE scores, provided the institution does not set a minimum threshold score for application.

The second strength is that this is the first study to explore the predictive power of the GRE on quality-related aspects of independent research work of graduate students. From our study, it appears that individual scales of the GRE are predictive of several aspects of graduate student research work. However, they explain, at the maximum, up to one-fourth of the variance in the outcome. Future research could consider exploring why the GRE

predicts specific research skills (e.g., which cognitive constructs measured by the GRE affect the level of research skills?).

Finally, this study is only the third to be conducted on the GRE in the European context and extends our understanding of how the GRE can be applied in STEM graduate schools in Europe, considering the distinct position of master's (as a final degree) within the European three cycle system [24,40]. The possibilities and conditions of this application will be discussed in more detail in Section 18, "Practical Implications for Graduate Admissions in Europe".

## 17. Limitations

This study does not come without limitations. The first and, perhaps, the main limitation is that it was conducted in a low-stakes situation: The students had been already admitted to their master's program, so their performance on the GRE did not affect their admissions on a master's program, and they did not prepare for the test as they would usually do in a high-stakes situation. According to students' self-reports, the median preparation time was 20 min, which mostly included reading information on the GRE test content and structure before taking it. On the one hand, the fact that the participants did not prepare for the test placed them in equal positions and provided a realistic assessment of their verbal, quantitative, and analytical writing skills. On the other hand, if students' admissions to their desired masters' programs would indeed depend on the GRE scores, we expect that they would report much higher preparation rates. SES could then have played a more significant role (with more wealthy students likely making more financial investments in buying courses and books for GRE preparation, hiring tutors, and being able to make more exam attempts to obtain the desired minimum level of scores). The debate on this issue would benefit from future studies on the GRE within a high-stakes situation, conducted within the European context.

The second limitation is that we used only one aspect (namely parental education) as a proxy for SES. We acknowledge that SES is a complicated construct [54], which goes beyond parental education only. However, parental education has been shown to represent SES adequately due to the greatest load it exerts on the SES index [62] and is used in research similar to ours, e.g., [40]. We initially also gathered information on other aspects of SES (parental occupational status, family financial situation, and subjective assessments of SES). Having tested those operationalizations as proxies for SES did not deliver substantially different results.

The third limitation is also related to SES. It might be that we found the relationship between the GRE scores and graduate study success to be independent from SES because we had a relatively homogeneous sample in terms of SES. Indeed, 69% of students in our sample had at least one parent with a higher education degree, which is higher than the average of tertiary attainment in the Netherlands (38%) and the OECD average (39%) [63]. This perhaps indicates that SES does not play a major role in such a preselected sample, while it could have played a role in a more diverse sample.

Finally, it could be beneficial to have a larger sample size, especially for the exploratory second part of our research, where we investigate the relationships between scores on GRE scales and students' research performance, as assessed by their supervisors, using grades and rubrics. Since the GRE is not required, even optionally, for applicants to life, natural, and geosciences programs at any Dutch universities, we were unable to utilize institutional admissions data on hundreds of applicants (as was available in the study by Zimmermann [24], for example). Instead, we had to organize a special data collection event and gather data from those students who voluntarily registered and attended. This could potentially limit the generalizability of our findings to all applicants to similar European masters' programs. However, we took steps to ensure that our findings are applicable to the general applicant pool by (1) conducting an a prior power analysis, (2) conducting data collection at a single moment (preventing communication between students), and

(3) gathering data from over 20 diverse STEM research master's programs. We expect that these measures taken together provide acceptable generalizability for our findings.

## 18. Practical Implications for Graduate Admissions in Europe

Even though the results of this study demonstrate that the GRE appears to predict several dimensions of study success in STEM research masters' programs, several aspects should be considered when deciding whether to use GRE as a selection method. Some of these considerations speak for the usage of the GRE, and the others speak against it. We suggest that each STEM graduate school weighs these considerations carefully in the context of its own admissions goals and educational mission.

The considerations that speak in favor of the implementation of the GRE are as follows. First, the GRE appears to predict not only traditional dimensions of study success (such as GGPA), but it also shows a potential to predict the quality of graduate students' research performance. Second, the GRE places students from different educational backgrounds on one scale and therefore allows for a comparison of their skills, required at the entrance to a graduate school. This fact not only facilitates the work of admissions committees in terms of their invested time and efforts, but it also makes the admissions decisions transparent to different stakeholders (including applicants and their parents). Third, being able to submit the GRE scores might benefit students from not very well known and established HEIs across the world and within their own country (e.g., universities of applied sciences in the Netherlands). These students can get a serious consideration for research masters' programs in Europe, as their levels of verbal, quantitative, and analytical writing skills are then better understood by the admissions committees, who could have previously disregarded their application due to unfamiliarity with the applicant's HEI, the quality of its undergraduate programs and examinations, or by using low-validity and more subjective methods such as traditional interviews [36,64]. Fourth, using the GRE provides a possibility for benchmarking: the HEIs can compare the level of skills between own applicants who graduated from their own bachelor programs and applicants from other countries and universities.

The considerations that speak against the usage of the GRE are as follows. First, as soon as the GRE is assigned a significant weight in admissions decisions, the scores might become prone to coaching effects, because students start preparing for the test. There is a study in the US context, however, which finds that mostly GRE-AW is prone to coaching, but not the other two sections [65]. More studies (also in the European context) are required to ensure that the effect of coaching is indeed limited. Second, the standardized tests are prone to stereotype threat—a self-fulfilling prophecy of test takers from minority groups that their underperformance on standardized tests might confirm the stereotypes of their minority group's intellectual capacity, which leads to actual underperformance [66,67]. Further research in the European context on this topic would be insightful for a better understanding of the extent to which stereotype threat affects the performance of European minority groups. The third consideration is that, while the GRE is not established as a selection method in the European research-intense graduate schools, it might take additional time and effort to ensure its acceptability by stakeholders (e.g., applicants and admissions committees).

If the GRE is used, admissions committees should familiarize themselves with the best practices of implementing it in admissions processes. The ETS guidelines state that the GRE scores cannot be used as a sole or primary admissions method [68]. Moreover, the ETS does not recommend applying cut-off scores and encourages admissions committees to account for measurement error using special tables (ETS, 2018). In a qualitative study, Posselt [67] showed how easily these recommendations are disregarded in admissions practices of the top graduate schools in the US.

### 19. Legal Limitations

Another important consideration for the GRE usage is legal requirements. The Dutch law, for example, does not allow HEIs to pass costs for (admissions) examinations on students with diplomas from EEA, equivalent to the Dutch diploma [69]. Additionally, taking the GRE requires financial investments (standard GRE General test administration costs were USD 205.00 in 2021 in all areas of the world except China and India, where the prices are higher). According to Dutch law, passing the costs of admissions examinations on to an applicant is allowed only if they do not meet the requirements of a higher educational program or do not hold a diploma that gives the right to admission [69]. In practice, adhering to these requirements would mean that the GRE could have been either introduced optionally or required only from international students outside of the European Economic Area (EEA). The optional usage of GRE might have a downside that only students with higher scores submit theirs [67]. Moreover, it was found that males and applicants from small HEIs that do not grant doctoral degrees are more likely to submit their GRE scores than females and males from larger HEIs that are granting doctoral degrees [70]. Some differences across the level of education and field of study were also detected [70]. Requiring GRE scores only from students with non-EEA diplomas will assist in comparing their skills between themselves but will not help in comparing their skills with those of students with bachelors' diplomas from within the EEA.

### 20. A View on HEI Mission—A Few Questions to Consider

At its core, much of the debate surrounding standardized testing, especially the GRE, revolves around a fundamental question: Which model should a given Higher Education Institution (HEI) or an entire higher education system aspire to follow? Should it adhere to the Humboldtian model of higher education, which emphasizes accessible education transcending social divides [71], or should it lean towards a neoliberalist model with its assumptions about privatization and the necessity to reduce expenditures in socially relevant areas [72]? What conditions must be met for standardized testing to function as a tool that guarantees fairness and access, thereby bringing us closer to the ideals of the Humboldtian model? Finally, when policy makers are deciding on the introduction of standardized testing in an admissions process, they should contemplate whether any of the involved stakeholders—testing companies, students, Higher Education Institutions (HEIs), or governmental-level educational policy makers—disproportionately benefit from its implementation.

### 21. Conclusions

The results of our study indicate that GRE scores are predictive for graduate study success in English-taught research masters' programs at a major European university. All GRE scales appear to be predictive for GGPA. The best predictor out of the GRE scales turned out to be GRE-V: it is predictive of internship grade and supervisors' assessments of students' skills on performing research and quality of content of their research report. The implementation of the GRE in admissions to selective European masters' programs requires serious consideration of potential pros and cons, as well as accounting for the limitations imposed by the law.

## References

1. Bologna Working Group on Qualifications Frameworks. *A Framework for Qualifications of the European Higher Education Area*; Ministry of Science, Technology, and Innovation of Denmark: Copenhagen, Denmark, 2005. Available online: https://edudoc.ch/record/46590?ln=en (accessed on 16 March 2024).
2. Skinner, M. Student Mobility in the European Higher Education Area (EHEA). World Education News + Reviews. 2018. Available online: https://wenr.wes.org/2018/12/student-mobility-in-the-european-higher-education-area-ehea (accessed on 16 March 2024).
3. European Commission; Education, Audiovisual and Culture Executive Agency; Eurydice. *The European Higher Education Area in 2020: Bologna Process Implementation Report*; Publications Office of the European Union: Luxembourg, 2020. Available online: https://eurydice.eacea.ec.europa.eu/publications/european-higher-education-area-2020-bologna-process-implementation-report (accessed on 16 March 2024).
4. Payne, D. A common yardstick for graduate education in Europe. *J. Eur. High. Educ. Area* **2015**, *3*, 21–48. Available online: https://www.handbook-internationalisation.com/en/handbuch/gliederung/?articleID=566#/Beitragsdetailansicht/675/566 (accessed on 16 March 2024).
5. European Commission. What is Erasmus+? 2020. Available online: https://erasmus-plus.ec.europa.eu/about-erasmus/what-is-erasmus (accessed on 16 March 2024).
6. Easley, J.; Strawderman, L.; Babski-Reeves, K.; Bullington, S.; Smith, B. Perceived quality factors in higher education. *Qual. High. Educ.* **2021**, *3*, 306–323. [CrossRef]
7. Caspersen, J.; Frolish, N.; Karlsen, H.; Aamodt, P.O. Learning outcomes across disciplines and professions: Measurement and interpretation. *Qual. High. Educ.* **2014**, *3*, 195–215. [CrossRef]
8. Altman, R.; Conrad, L.; Flaugher, R.; Thompson, R.; Wallmark, M.; Willingham, W. *Graduate Record Examinations: Technical Manual*; Conrad, L., Trismen, D., Miller, R., Eds.; Educational Testing Service: Princeton, NJ, USA, 1977; pp. 1–104.
9. Carnegie Foundation for the Advancement of Teaching. *The Graduate Record Examination*; Carnegie Foundation for the Advancement of Teaching: Stanford, CA, USA, 1941.
10. Educational Testing Service. A Snapshot of the Individuals Who Took the GRE General Test July 2014–June 2019. i–73. Educational Testing Service. 2019. Available online: https://www.ets.org/pdfs/gre/snapshot-test-taker-data-2019.pdf (accessed on 16 March 2024).
11. Educational Testing Service. A Snapshot of the Individuals Who Took the GRE General Test July 2016–June 2021. i–73. Educational Testing Service. 2021. Available online: https://www-vantage-qa-publish.ets.org/pdfs/gre/snapshot.pdf (accessed on 16 March 2024).
12. Educational Testing Service. A Snapshot of the Individuals Who Took the GRE General Test July 2017–June 2022. i–73. Educational Testing Service. 2022. Available online: https://www.ets.org/content/dam/ets-org/pdfs/gre/snapshot-test-taker-data-2022.pdf (accessed on 16 March 2024).

13. King, M.R.; Jennings, K.; Chalkley, R.G.; Sealy, L.J. Questioning the value of the Graduate Record Examinations (GRE) in PhD admissions in biomedical engineering. *Ann. Biomed. Eng.* **2020**, *3*, 2155–2157. [CrossRef] [PubMed]

14. Langin, K. 'GRExit' Gains Momentum as Ph.D. Programs Drop Exam Requirement. *Science* **2022**, *378*, 936–937. Available online: https://www.science.org/content/article/gre-exit-gains-momentum-ph-d-programs-drop-exam-requirement-amid-pandemic (accessed on 16 March 2024). [CrossRef] [PubMed]

15. UC Office of the President. University of California Board of Regents Unanimously Approved Changes to Standardized Testing Requirement for Undergraduates. 2020. Available online: https://www.universityofcalifornia.edu/press-room/university-california-board-regents-approves-changes-standardized-testing-requirement (accessed on 16 March 2024).

16. Bridgeman, B.; Burton, N.; Cline, F. A note on presenting what predictive validity numbers mean. *Appl. Meas. Educ.* **2009**, *3*, 109–119. [CrossRef]

17. Burmeister, J.; McSpadden, E.; Rakowski, J.; Nalichowski, A.; Yudelev, M.; Snyder, M. Correlation of admissions statistics to graduate student success in medical physics. *J. Appl. Clin. Med. Phys.* **2014**, *3*, 375–385. [CrossRef] [PubMed]

18. Fu, Y. The Effectiveness of Traditional Admissions Criteria in Predicting College and Graduate Success for American and International Students. Doctoral Dissertation, University of Arizona, Tucson, AZ, USA, 2012. Available online: https://www.proquest.com/docview/917742486 (accessed on 16 March 2024).

19. Moneta-Koehler, L.; Brown, A.M.; Petrie, K.A.; Evans, B.J.; Chalkley, R. The Limitations of the GRE in predicting success in biomedical graduate school. *PLoS ONE* **2017**, *3*, e0166742. [CrossRef]

20. Burton, N.; Wang, M. Predicting long-term success in graduate school: A collaborative validity study. In *The Research Foundation for the GRE Revised General Test: A Compendium of Studies*; Wendler, C., Bridgeman, B., Eds.; Educational Testing Service: Princeton, NJ, USA, 2014; pp. 5.6.1–5.6.5. Available online: https://www.ets.org/s/research/%20pdf/gre_compendium.pdf (accessed on 16 March 2024).

21. Burton, N.W.; Wang, M. *Predicting Long-Term Success in Graduate School: A Collaborative Validity Study*; Report No. 99-14R. ETS RR-05-03; Educational Testing Service: Princeton, NJ, USA, 2005. Available online: https://files.eric.ed.gov/fulltext/EJ1111377.pdf (accessed on 16 March 2024).

22. Howell, L.L.; Sorenson, C.D.; Jones, M.R. Are undergraduate GPA and general GRE percentiles valid predictors of student performance in an engineering graduate program? *Int. J. Eng. Educ.* **2014**, *3*, 1145–1165. Available online: https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=2342&context=facpub (accessed on 16 March 2024).

23. Klieger, D.M.; Cline, F.A.; Holtzman, S.L.; Minsky, J.L.; Lorenz, F. *New Perspectives on the Validity of the GRE ®General Test for Predicting Graduate School Grades: New Perspectives for Predicting Graduate School Grades*; Report No. RR-14-26; Educational Testing Service: Princeton, NJ, USA, 2014. [CrossRef]

24. Zimmermann, J.; von Davier, A.A.; Buhmann, J.M.; Heinimann, H.R. Validity of GRE General test scores and TOEFL scores for graduate admissions to a technical university in Western Europe. *Eur. J. Eng. Educ.* **2017**, *3*, 144–165. [CrossRef]

25. Kuncel, N.R.; Hezlett, S.A. Standardized tests predict graduate students' success. *Science* **2007**, *315*, 1080–1081. [CrossRef] [PubMed]

26. Kuncel, N.R.; Hezlett, S.A. Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Curr. Dir. Psychol. Sci.* **2010**, *3*, 339–345. [CrossRef]

27. Kuncel, N.R.; Wee, S.; Serafin, L.; Hezlett, S.A. The validity of the Graduate Record Examination for master's and doctoral programs: A meta-analytic investigation. *Educ. Psychol. Meas.* **2010**, *3*, 340–352. [CrossRef]

28. Cox, G.W.; Hughes, W.E., Jr.; Etzkorn, L.H.; Weisskopf, M.E. Predicting computer science Ph.D. completion: A case study. *IEEE Trans. Educ.* **2009**, *3*, 137–143. [CrossRef]

29. Dore, M. Factors in the Admissions Process Influencing Persistence in a Master's of Science Program in Marine Science. Doctoral Dissertation, Nova Southeastern University, Fort Lauderdale, FL, USA, 2017. Available online: https://nsuworks.nova.edu/cgi/viewcontent.cgi?article=1117&context=fse_etd (accessed on 16 March 2024).

30. Lorden, J.F.; Kuh, C.V.; Voytuk, J.A. (Eds.) *Research-Doctorate Programs in the Biomedical Sciences: Selected Findings from the NRC Assessment*; The National Academies Press: Washington, DC, USA, 2011. [CrossRef]

31. Lott, J.L.I.; Gardner, S.; Powers, D.A. Doctoral student attrition in the STEM fields: An exploratory event history analysis. *J. Coll. Stud. Retent. Res. Theory Pract.* **2009**, *3*, 247–266. [CrossRef]

32. Miller, E.M. Promoting student success in statistics courses by tapping diverse cognitive abilities. *Teach. Psychol.* **2019**, *3*, 140–145. [CrossRef]

33. Petersen, S.L.; Erenrich, E.S.; Levine, D.L.; Vigoreaux, J.; Gile, K. Multi-institutional study of GRE scores as predictors of STEM PhD degree completion: GRE gets a low mark. *PLoS ONE* **2018**, *3*, e0206570. [CrossRef]

34. Bridgeman, B.; Olivera-Aguilar, M.; Holtzman, S. The GRE as a Predictor of Persistence to a PhD. *Front. Educ.* **2023**, *3*, 8:1182508. [CrossRef]

35. Kuncel, N.R.; Hezlett, S.A. The utility of standardized tests: Response. *Science* **2007**, *3*, 1696–1697. [CrossRef]

36. Hall, J.D.; O'Connell, A.B.; Cook, J.G. Predictors of student productivity in biomedical graduate school applications. *PLoS ONE* **2017**, *3*, e0169121. [CrossRef] [PubMed]

37.  Sanford, R. Predictors of International Graduate Student Success in U.S. Universities: Analysis of Noncognitive Student Variables and Institutional Characteristics. Doctoral dissertation, University of La Verne, La Verne, CA, USA, 2009. Available online: https://www.proquest.com/openview/0ca5804f25bd0ac8eeb599e839f8d656/1?pq-origsite=gscholar&cbl=18750&diss=y (accessed on 16 March 2024).

38.  Sealy, L.; Saunders, C.; Blume, J.; Chalkley, R. The GRE over the entire range of scores lacks predictive ability for PhD outcomes in the biomedical sciences. *PLoS ONE* **2019**, *3*, e0201634. [CrossRef] [PubMed]

39.  Frasier, H.S. An Analysis of Institutional Characteristics That Contribute to Extended Time to Doctoral Degree. Doctoral dissertation, University of Maryland, College Park, MD, USA, 2013. Available online: https://drum.lib.umd.edu/bitstream/handle/1903/14911/Frasier_umd_0117E_14734.pdf?sequence=1&isAllowed=y (accessed on 16 March 2024).

40.  Schwager, I.T.L.; Hülsheger, U.R.; Bridgeman, B.; Lang, J.W.B. Graduate Student Selection: Graduate record examination, socioeconomic status, and undergraduate grade point average as predictors of study success in a western European University. *Int. J. Sel. Assess.* **2015**, *3*, 71–79. [CrossRef]

41.  Miller, C.; Stassun, K. A test that fails. *Nature* **2014**, *3*, 303–304. Available online: https://www.nature.com/articles/nj7504-303a (accessed on 16 March 2024). [CrossRef]

42.  Gordon, A. Don't Remind Me: Stereotype Threat in High-Stakes testing. *Univ. Baltim. Law Rev.* **2019**, *3*, 387–412. Available online: https://scholarworks.law.ubalt.edu/cgi/viewcontent.cgi?article=2053&context=ublr (accessed on 16 March 2024).

43.  Steele, C.M.; Aronson, J. Stereotype Threat and the Intellectual Test Performance of African Americans. *J. Personal. Soc. Psychol.* **1995**, *3*, 797–811. [CrossRef]

44.  Walters, A.M.; Lee, S.; Trapani, C. *Stereotype Threat, the Test-Center Environment, and Performance on the GRE General Test*; Report No. 01-03R. ETS RR-04-37; Educational Testing Service: Princeton, NJ, USA, 2004. Available online: https://www.ets.org/research/policy_research_reports/publications/report/2004/hyzm (accessed on 16 March 2024).

45.  Sackett, P.R.; Hardison, C.M.; Cullen, M.J. On interpreting stereotype threat as accounting for African American-White differences on cognitive tests. *Am. Psychol.* **2004**, *3*, 7–13. [CrossRef]

46.  Tomeh, D.H.; Sackett, P.R. On the Continued Misinterpretation of Stereotype Threat as Accounting for Black-White Differences on Cognitive Tests. *Pers. Assess. Decis.* **2022**, *3*, 1. [CrossRef]

47.  Walker, M.E.; Bridgeman, B. Stereotype threat spillover and SAT®scores. *ETS Res. Rep. Ser.* **2008**, *3*, i-10. [CrossRef]

48.  Clayton, V. The Problem with the GRE. 2016. Available online: https://www.theatlantic.com/education/archive/2016/03/the-problem-with-the-gre/471633/ (accessed on 16 March 2024).

49.  Zwick, R. Is the SAT a "wealth test"? The link between educational achievement socioeconomic status. In *Rethinking the SAT: The Future of Standardized Testing in University Admissions*; Zwick, R., Ed.; RoutledgeFalmer: London, UK, 2004; pp. 203–216.

50.  Ledford, S.; Monteagudo, M.M.; Flores, A.N.; Glass, J.B.; Cobb, K.M. Geogrexit: Strategies for removing the GRE as one barrier of entry to geoscience graduate programs. *Geol. Soc. Am. Abstr. Programs* **2020**, *52*, 6. [CrossRef]

51.  Croizet, J.-C.; Dutrevis, M. Socioeconomic status and intelligence: Why test scores do not equal merit. *J. Poverty* **2004**, *3*, 91–107. [CrossRef]

52.  Sackett, P.R.; Kuncel, N.R.; Arneson, J.J.; Cooper, S.R.; Waters, S.D. Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychol. Bull.* **2009**, *3*, 1–22. [CrossRef] [PubMed]

53.  Stricker, L.J.; Rock, D.A. *Examinee Background Characteristics and GRE General Test Performance*; GRE Board Research Report No. 89-07R, ETS Research Report 92-80; Educational Testing Service: Princeton, NJ, USA, 1993. Available online: https://www.ets.org/research/policy_research_reports/publications/report/1993/ihle (accessed on 16 March 2024).

54.  Rodríguez-Hernandez, C.F.; Cascallar, E.; Kyndt, E. Socio-economic status and academic performance in higher education: A systematic review. *Educ. Res. Rev.* **2020**, *3*, 100305. [CrossRef]

55.  Stern, W. *Über Psychologie der Individuellen Differenzen: Ideen zu einer "differentiellen Psychologie" [On the Psychology of Individual Differences: Toward a "Differential Psychology"]*; Barth: Leipzig, Germany, 1900.

56.  Tyler, L. *The Psychology of Human Differences*; Appleton-Century-Crofts, Division of Meredith Publishing Company: New York, USA, 1965.

57.  Educational Testing Service. Test Content. Available online: https://www.ets.org/gre/revised_general/about/content (accessed on 16 March 2024).

58.  Educational Testing Service. Prepare for the GRE®General Test: Verbal Reasoning, Quantitative Reasoning, Analytical Writing. Available online: https://www.ets.org/gre/test-takers/general-test/prepare.html (accessed on 16 March 2024).

59.  Postmes, L.; Bouwmeester, R.A.M.; de Kleijn, R.A.M.; van der Schaaf, M. Supervisors' untrained postgraduate rubric use for formative and summative purposes. *Assess. Eval. High. Educ.* **2022**, *48*, 1–14. [CrossRef]

60.  Gooding, Y. The Relationship between Parental Educational Level and Academic Success of College Freshmen. Doctoral dissertation, Iowa State University, Ames, IA, USA, 2001. [CrossRef]

61.  Organization for Economic Co-Operation and Development. Scaling Procedures and Construct Validation of Context Questionnaire Data. PISA 2015 Technical Report. 2017. Available online: https://www.oecd.org/pisa/sitedocument/PISA-2015-Technical-Report-Chapter-16-Procedures-and-Construct-Validation-of-Context-Questionnaire-Data.pdf (accessed on 16 March 2024).

62. Caro, D.H.; Cortés, D. Measuring family socioeconomic status: An illustration using data from PIRLS 2006. In *Issues and Methodologies in Large-Scale Assessments*; Educational Testing Service and International Association for the Evaluation of Educational Achievement: Princeton, NJ, USA; Hamburg, Germany, 2012; Volume 5, pp. 9–34. Available online: https://ierinstitute.org/fileadmin/Documents/IERI_Monograph/Volume_5/IERI_Monograph_Volume_05.pdf (accessed on 16 March 2024).

63. Organization for Economic Co-Operation and Development. *Benchmarking Higher Education System Performance: The Netherlands*; Higher Education, OECD: Paris, France, 2019. [CrossRef]

64. Goho, J.; Blackman, A. The effectiveness of academic admission interviews: An exploratory meta-analysis. *Med. Teach.* **2006**, *3*, 335–340. [CrossRef] [PubMed]

65. Powers, D.E. Understanding the impact of special preparation for admissions tests. In *Advancing Human Assessment. The Methodological, Psychological and Policy Contributions of ETS*; Bennett, R.E., von Davier, M., Eds.; Springer Open: Berlin/Heidelberg, Germany, 2017; pp. 553–564. [CrossRef]

66. Garces, L.M. Aligning diversity, quality, and equity: The implications of legal and public policy developments for promoting racial diversity in graduate studies. *Am. J. Educ.* **2014**, *3*, 457–480. [CrossRef]

67. Posselt, J.R. *Inside Graduate Admissions: Merit, Diversity, and Faculty Gatekeeping*; Harvard University Press: Cambridge, MA, USA, 2016.

68. Educational Testing Service. GRE®Guidelines for the Use of Scores. Available online: https://www.ets.org/gre/score-users/scores/use-scores.html (accessed on 16 March 2024).

69. Ministry of Education, Culture and Science of the Netherlands. Regeling van de Minister van Onderwijs, Cultuur en Wetenschap 18 juni 2021, nr. 28415868, Houdende Andere Bijdragen van Studenten in het Hoger Onderwijs (Regeling Andere Bijdragen van Studenten in het Hoger Onderwijs). 2021. Available online: https://zoek.officielebekendmakingen.nl/stcrt-2021-33378.html (accessed on 16 March 2024).

70. Cho-Baker, S.; Kell, H.J. Who sends scores to GRE-optional graduate programs? A case study investigating the association between latent profiles of applicants' undergraduate institutional characteristics and propensity to submit GRE scores. *Educ. Sci.* **2022**, *3*, 529. [CrossRef]

71. Kirby, W.C. *Empires of Ideas: Creating the Modern University from Germany to America to China*; Harvard University Press: Cambridge, MA, USA, 2022.

72. Mintz, B. Neoliberalism and the crisis in higher education: The cost of ideology. *Am. J. Econ. Sociol.* **2021**, *3*, 79–112. [CrossRef]