Research Report

Studies of Possible Effects of *GRE*® *ScoreSelect*® on Subgroup Differences in *GRE*® General Test Scores

ETS RR–22-13

David M. Klieger
Lauren J. Kotloff
Vinetha Belur
Megan E. Schramm-Possinger
Steven L. Holtzman
Hezekiah Bunde

December 2022

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Studies of Possible Effects of *GRE®* *ScoreSelect®* on Subgroup Differences in *GRE®* General Test Scores

David M. Klieger, Lauren J. Kotloff, Vinetha Belur, Megan E. Schramm-Possinger, Steven L. Holtzman, & Hezekiah Bunde

ETS, Princeton, NJ

Intended consequences of giving applicants the option to select which test scores to report include potentially reducing measurement error and inequity in applicants' prior test familiarity. Our first study determined whether score choice options resulted in unintended consequences for lower performing subgroups by detrimentally increasing score gaps in ways and for reasons that the research literature had suggested. Our follow-up study explored possible determinants of changes in score gaps attributable to score choice options. Using *GRE®* *SCORESELECT®*, the score choice system for the GRE general test, we concluded that unintended consequences were few, small in magnitude, and usually undetectable. To the extent that unintended consequences occurred, they were limited to effects for citizenship subgroups and generally benefited lower performing subgroups.

Keywords fairness; diversity; equity; race; ethnicity; gender; citizenship; age; parental education; socioeconomic status; *GRE®*; higher education; graduate school; professional school; score choice options; *ScoreSelect®*; superscore; super score; test scores

doi:10.1002/ets2.12356

Measurement error is a major issue in high-stakes testing. All test scores will reflect some degree of measurement error, many sources of which are unrelated to the test itself.[1] For example, an examinee might suffer an illness during the test, the thermostat in a testing center might suddenly break causing the room to become unbearably hot, or a flash mob might unexpectedly interrupt an examination with a rendition of *Les Miserables* (Goodman, 2011). Even for tests high in statistical reliability, measurement error may occur if test items do not happen to cover an examinee's knowledge of a subject area with which the examinee is truly familiar. All of the foregoing sources of error can have serious consequences for test validity, defined as "the extent to which the evidence supports or refutes the proposed interpretations and uses" (Kane, 2006, p. 17). These sources of error can reduce test scores, resulting in an understatement of an examinee's true skill level. A program, institution, or organization that selects applicants based on affected test scores might reject a truly qualified individual. One way to reduce this risk is to allow examinees to eliminate aberrant scores from the decision-making process.

A second major issue in high-stakes testing is variability among examinees in access to prior knowledge about and experience with the kinds of items on a test, the best test-taking strategies, and opportunities to prepare for a test. This disparity might be due to differences in financial resources (e.g., the ability to pay for extra test study guides) and associations with others who are familiar with the test (e.g., family members and friends who have taken it). One way to simultaneously reduce the downside of measurement error and inequity in prior test familiarity is by giving examinees opportunities to take a test under realistic conditions and then to permit the examinee to observe the outcome (i.e., test scores) without forcing examinees to include those scores in any future score reports. Several high-stakes tests offer this opportunity to choose which scores to send to institutions, including the *SAT®* (*Score Choice®*), ACT, and *GRE®* (*ScoreSelect®*), used in higher educational admissions.

Under the assumption that examinees generally wish to maximize their probability of being admitted or receiving some other benefit (e.g., a scholarship or fellowship), the introduction of a score choice option might cause some examinees to become selective in their score submission choices. An examinee who receives a less than perfect score has the opportunity

Corresponding author: David M. Klieger, E-mail: dklieger@ets.org

to improve that score by retaking the test. With a score choice option, if the examinee receives a retest score that is higher than previous scores, then the examinee has the opportunity to report just the higher score to an institution. If the examinee receives a retest score that is lower than what the examinee wishes an institution to consider, then the examinee has the opportunity to report just a higher score that the examinee may have received prior to the retest. Because future scores can help but never need hurt an examinee's prospects, a score choice option might encourage at least some examinees to retake an examination.

Some institutions have objected to the offering and use of score choice options in the belief that they will unfairly advantage certain demographic groups, such as wealthier students, who may be better able to take advantage of those score choice options (see, e.g., Matthews, 2009). An unintended consequence of score choice options might be an increase in the difference between the mean scores of groups historically underrepresented in a sector (e.g., college or graduate school) and those of groups not historically underrepresented in that sector. Research about fairness in selection often focuses on groups defined by race/ethnicity, gender, citizenship, age, socioeconomic status, or citizenship. When decision makers use some degree of rank ordering or cut scores to make selection or benefit decisions, mean score differences between groups generally will translate into differences in selection or other benefit rates across those groups (see Sackett & Ellingson, 1997; Sackett & Wilk, 1994). If the score choice option increases the mean score difference between a lower scoring group and another higher scoring group, then even at a moderately selective institution, the score choice option would be expected to attenuate the selection rate of the lower scoring group vis-à-vis the comparison group.[2] Conversely, if the score choice option decreases the mean score difference between a lower scoring group and another, higher scoring group, then the score choice option would be expected to increase the selection rate of the lower scoring group.

The impact of score reporting options on fairness and the diversity of selected individuals is unknown.[3] It is possible that, under those options, score submission and retesting behaviors will vary significantly based on race/ethnicity, gender, citizenship, age, socioeconomic status, or citizenship. For example, examinees sometimes have to pay a fee to exercise a score choice option. They also generally have to pay a fee to retake a test. Consequently, members of groups with lower average incomes and citizens of lower socioeconomic nations may typically have less opportunity to take full advantage of a score choice option. Even if the effects of a score choice option on score submission and retesting behaviors are uniform across demographic groups, the score choice option may lead to different outcomes across demographic groups. For example, a score choice option may encourage subgroup A to retest as often as subgroup B does and to submit scores in the same way as subgroup B does, but subgroup B's Graduate Record Examination (GRE) scores might on average improve much more upon retesting than subgroup A's GRE scores do on average.

To understand the effects of a score choice option, we conducted two related studies. In the first study, we investigated the practical impact of ScoreSelect, the score choice option for the GRE, on fairness and the diversity of selected individuals. Given that there are many stakeholders who highly value such diversity (see, e.g., Walpole et al., 2002), our analytical approach was conservative in that it was based on a set of assumptions that we expected would most likely lead to practically significant changes in subgroup mean score differences. In the second study, we investigated possible determinants of any impact that ScoreSelect might have on fairness and diversity.

Study 1 Method

Instruments

GRE Revised General Test

The GRE is a multiple choice and constructed response test used for graduate school admissions and funding decisions. It is offered worldwide. Most examinees take it by computer in a testing center, although in some cases it is offered in paper-based format. The GRE has a verbal reasoning section (GRE-V), a quantitative reasoning section (GRE-Q), and an analytical writing (GRE-AW) section. The most recent revision of the GRE became operational in August 2011. There are subject-specific GRE tests as well that offer score choice options, but in this article, GRE refers only to the GRE revised General Test.

GRE Accounts and Background Information Questionnaire

In order to register to take the GRE test, a registrant typically creates an online GRE account prior to the test date. As part of the process of creating this account, registrants are asked to complete a background information questionnaire (BIQ) indicating their birth date (month, day, and year) and gender (female or male). To continue the registration process, the registrant is asked to provide additional personal background information. This information includes the registrant's country of citizenship, race/ethnicity (American Indian or Alaska Native, Asian, Black, Latino, or White) if a U.S. citizen and the highest education levels attained by the registrant's mother (or female guardian) and father (or male guardian).

GRE Score Reporting and ScoreSelect®

An examinee can request that ETS (the administrator and scorer of the GRE) send a score report to a graduate institution or organization affiliated with graduate education that has been approved by ETS to receive score reports. A report provides separate scores for GRE-V, GRE-Q, and GRE-AW. Prior to July 1, 2012, GRE score reporting was cumulative, and reports included all scores earned in the previous 5-year period. Under the ScoreSelect option introduced July 1, 2012, an examinee has additional choices. At the end of a computer-delivered test, a candidate may choose to send (a) the scores from the current test administration, (b) scores from all of the candidate's GRE General Test administrations in the last 5 years, or (c) no scores at that time.[4] After test day, a candidate may send scores from one or multiple test administrations from the last 5 years or no scores at all. Candidates select one or more specific test dates, and all scores (i.e., GRE-V, GRE-Q, and GRE-AW) from each selected test date are reported. For paper-based GRE tests, the examinee must make score submission choices prior to the test date. For a fee of $28 (U.S.) per report, a test taker may send scores from any or all administrations within the 5-year score-reporting period. Fee-based score reports are called additional score reports (ASRs).

Research Design

We analyzed data for up to 2,015,024 GRE score reports ordered between July 1, 2012 (when ScoreSelect went into effect) and July 2013. The numbers of reports belonging to various demographic groups are indicated in various tables in this report. Scores were limited to those from the revised GRE (i.e., from GRE tests taken as of August 1, 2011), because it was unclear if any mean group differences in GRE scores would be different for the revised version in comparison to the former version, irrespective of any effects of ScoreSelect. Also, except for comparisons specifically for citizenship groups, all analyzed demographic subcategories (i.e., subcategories for race/ethnicity, gender, age, and socioeconomic status) are U.S. citizens only. Relevant differences among subgroups within those subcategories may differ country by country; we did not have a firm enough basis to create hypotheses for those subcategories outside of a U.S. context, and the majority of GRE examinees are U.S. citizens.

We examined mean group differences based on information from examinees' GRE accounts or BIQ responses. For analyses based on race/ethnicity, we relied upon the given categorizations from the BIQ: American Indian or Alaska Native, Asian, Black, Latino, or White. Likewise, we analyzed mean group differences for gender based on the given categorizations from GRE accounts (i.e., female or male). Due to mean citizenship-based differences that have been found in GRE scores (see, e.g., ETS, 2013, discussed further below), citizenship information from the BIQ was classified into three subgroups: United States.; Chinese (citizens of mainland China/the People's Republic of China); or non-U.S., non-Chinese citizens (e.g., Canadian, French, Japanese). To determine age from GRE account information, we subtracted an examinee's birth date from the date of the respective score report and rounded to the nearest whole year. Based on a definition of a "traditional" graduate student as a student starting graduate school before age 25, and based on research regarding age-based changes in GRE test performance (especially a discontinuity in performance found at age 40; see ETS, 2013 and Trapani, 2013), we divided age into three categories: up to age 25, age 25 to 40, and age 40 and older.

Finally, we collected data about socioeconomic status via the BIQ question about the highest education level attained by an examinee's mother (or female guardian) and father (or male guardian). If the examinee reported for more than one parent or guardian, we then took the highest level reported for each set of parents or guardians under the assumption that the parent or guardian with the higher education level would have the dominant influence on the examinee's test performance and decisions regarding the graduate and professional school application process. To reduce the number of analyses that would be required for the nine educational levels provided by the BIQ, we combined the parental/guardian educational attainment levels into the following four categories: (a) some high school, grade school, or less; (b) high school diploma or equivalent, business, trade school, or some college; (c) associate or bachelor's degree; and (d) at least some graduate or professional school. With the exceptions discussed below, the higher performing group on the GRE test is White, male, a U.S. citizen, under age 26, and has a parent who attended at least some graduate school.

Study 1 Hypotheses

Given that the sample sizes involved were very large (in the thousands to hundreds of thousands), we focused our initial hypothesis, analyses, and conclusions on practical significance rather than statistical significance. Because we could not run a controlled experiment to determine the effects of ScoreSelect, we limited our hypotheses, analyses, and conclusions to associations (i.e., evidence of possible causal relationships). Specifically, we examined the following set of hypotheses: ScoreSelect is associated with practically significant changes in mean group differences. The mean score differences will increase to a practically significant extent in favor of the higher performing groups (i.e., White, male, U.S. citizen, under age 26, with a parent who attended at least some graduate school), resulting in even higher graduate school admission rates for these groups vis-à-vis comparison groups. Exceptions to this hypothesis will be for the following groups for the following GRE sections:

- Chinese citizens; non-U.S., non-Chinese citizens; and Asian Americans on GRE-Q
- Examinees age 25 to 40 and age 40 and older on GRE-V

Study 1 Analyses

Measure of Practical Significance

For each GRE section (GRE-V, GRE-Q, and GRE-AW), standardized mean group differences across subgroup comparisons were calculated using Cohen's *d* value point estimates, where Cohen's *d* was the difference in the subgroup means divided by the pooled standard deviation.[5] The point estimates and sample sizes are separately reported for ASRs and free reports in Appendix A, Tables A1–A4, to address the concern that financial resources might be impacting score choice and retesting behaviors. Statistical significance is not a focus here, given that the smallest sample size for a compared subgroup is 1,389 and most subgroup sample sizes are a great deal larger.[6]

Meaningful Interpretation Using Ratios

The focal analyses appear in Appendix A, Table A5, and report ratios based on research by Sackett and others (see the first table in Sackett & Wilk, 1994 and in Sackett & Ellingson, 1997). Tables A1–A4 are insufficient alone because practical significance in selection depends on the larger context (e.g., cut scores, selection ratios). Even Cohen himself (1988) cautioned about reliance on his often-cited guidelines to determine practical significance of a standardized mean difference. Therefore, we used the modified and extended approach of Sackett and his colleagues (Sackett & Ellingson, 1997; Sackett & Wilk, 1994) in which we estimated how many times more often the higher scoring group would be admitted in comparison to the lower scoring group under varying levels of selectivity in top-down selection (admission of top 10% vs. top 50% vs. top 90% of scores).[7] These estimations were made separately for (a) a hypothetical condition under which all reportable scores on the current version of the GRE test were reported by all examinees who ordered score reports (*Hypothetical All*)[8] and (b) the actual scores reported under ScoreSelect (*Actually Sent*). Thus, two ratios were calculated at a time, one ratio representing a hypothetical world without ScoreSelect and one ratio representing the actual world with ScoreSelect available (abbreviated as *No SS* and *SS*, respectively, in Table A5).[9] Practical significance was defined as, after rounding, a 5% or greater difference between the ratios for the hypothetical all and actually sent conditions for highly selective admissions systems (selection of top 10% of scores).

Study 1 Results

Table A5 indicates that the only comparisons for which practically significant differences (a 5% or more change in mean group differences) were found were citizenship subgroups. In general, then, our hypotheses were not supported. The two cases in which they were supported were for ASRs when a highly selective admissions system with top-down selection admits students based solely on GRE-Q. In such a case, the admissions rate advantage for Chinese citizens vis-à-vis U.S. citizens increased by 6% and for non-U.S., non-Chinese citizens vis-à-vis U.S. citizens increased by 5%. Few other practically significant changes were found. These other changes resulted in the lower scoring group (non-U.S. citizens generally, Chinese citizens in particular, or non-U.S., non-Chinese citizens in particular) improving their performance vis-à-vis the performance of the higher scoring group (U.S. citizens). Although the lower performing groups (non-U.S. citizens) still scored lower on average on GRE-V and GRE-AW than did the higher performing group (U.S. citizens), the score gap shrank to some extent (i.e., the *d* value decreased, and the selection ratio in favor of the higher performing group declined). Note that this result is unsurprising, given that one would expect the English language skills of U.S. citizens on average to be stronger than the English language skills of non-U.S. citizens. These practically significant changes were more frequent and greater in magnitude (percentage change) for ASRs than for free ones. As indicated in Table A5, for scenarios with highly selective top-down selection on GRE-V, there was a

- 27% reduction in how many times more often U.S. citizens would be admitted compared to Chinese citizens for ASRs;
- 20% reduction in how many times more often U.S. citizens would be admitted compared to Chinese citizens for free reports;
- 16% reduction in how many times more often U.S. citizens would be admitted compared to non-U.S. citizens in general for ASRs;
- 7% reduction in how many times more often U.S. citizens would be admitted compared to non-U.S. citizens in general for free reports;
- 7% reduction in how many times more often U.S. citizens would be admitted compared to non-U.S., non-Chinese citizens for ASRs; and
- 5% reduction in how many times more often U.S. citizens would be admitted compared to non-U.S., non-Chinese citizens for free reports.

As indicated in Table A5, for scenarios with highly selective top-down selection on GRE-AW, there was a

- 12% reduction in how many times more often U.S. citizens would be admitted compared to Chinese citizens for ASRs;
- 5% reduction in how many times more often U.S. citizens would be admitted compared to Chinese citizens for free reports; and
- 6% reduction in how many times more often U.S. citizens would be admitted compared to non-U.S. citizens in general for ASRs.

Differential Patterns of ScoreSelect Use by Citizenship Subgroups

Given the several findings of practically significant changes in mean subgroup differences for citizenship, we sought confirmatory empirical evidence of a connection between the changes in mean subgroup differences for citizenship groups and ScoreSelect usage. Figures 1 (for ASRs) and 2 (for free reports) illustrate that non-U.S., non-Chinese citizens and especially Chinese citizens utilized ScoreSelect more often than did U.S. citizens, regardless of whether the examinee had to pay for the report. Generally, usage patterns looked similar for ASR and free reports (see Figures 1 and 2). Tables B1 and B2 (see Appendix B) provide numerical data (including sample sizes) underlying Figures 1 and 2.[10]

Replication of Results for 2+ Score Reports

The results above include examinees who made reporting choices based on having *at least one* set of reportable GRE General Test scores. We operationalized a *score reporting choice* in this way, because the ultimate purpose of this study is to address questions about impact on subgroups' relative admission rates in a world in which, for some applicants, only one set of scores is potentially reportable at a given moment, and for other applicants, more than one set of scores is potentially reportable at that same moment. However, we also examined a subset of the data that included just those reports where *two or more* sets of scores could be reported. One could characterize a score reporting choice as a situation in which there is a choice to be made among two or more sets of scores; we refer to this as a *direct ScoreSelect decision*.[11] Because the majority

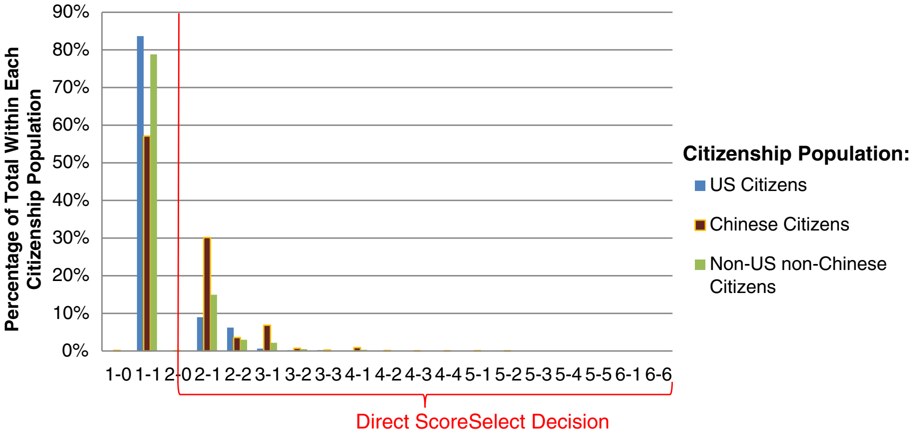


Figure 1 Differential patterns of ScoreSelect use by citizenship groups: additional score reports (ASRs). The values to the left of the hyphens (x-axis) indicate the number of sets of scores that could have been reported on a score report. The values to the right of the hyphens indicate the number of sets of scores that were reported on a score report given the number of sets of scores that could have been reported on it. The y-axis percentages for the bars for each citizenship population sum to 100%. Note that a ScoreSelect decision is directly made when the value to the left of the hyphen equals 2 or greater (bracketed in red). The numerical data underlying this figure appear in Table B1. For some scenarios (e.g., 1 – 0, 2 – 0), the percentages represented are actually near 0% rather than exactly 0%.

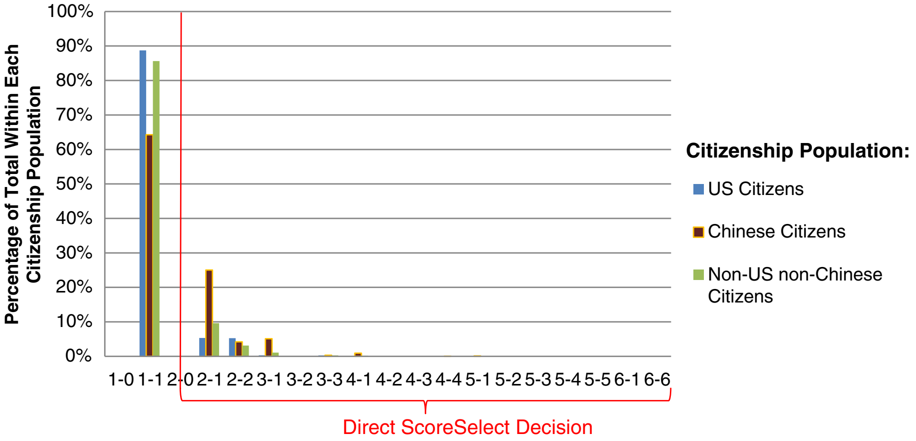


Figure 2 Differential patterns of ScoreSelect use by citizenship groups: free reports. The values to the left of the hyphens (x-axis) indicate the number of sets of scores that could have been reported on a score report. The values to the right of the hyphens indicate the number of sets of scores that were reported on a score report given the number of sets of scores that could have been reported on it. The y-axis percentages for the bars for each citizenship population sum to 100%. Note that a ScoreSelect decision is directly made when the value to the left of the hyphen equals 2 or greater (bracketed in red). The numerical data underlying this figure appear in Table B2. For some scenarios (e.g., 5 – 1), the percentages represented are actually near 0% rather than exactly 0%.

of score reports are sent where there is only one set of scores available to be reported (see Tables B1 and B2 and Figures 1 and 2), any effects of direct ScoreSelect decisions might be substantially diluted by these single-set reporting situations. It could be argued that failure to detect any effects of a score choice option is therefore due to extraneous circumstances. For the subset of data, our review of Cohen's *d* values and ratios of how many times more of ten the higher scoring group would be admitted compared to the lower scoring group revealed the exact same practically significant differences as did the analyses based on the full data set. In general, results were very similar (when not identical, after rounding).

Background for Study 2

In a second study, we ran a preliminary investigation into possible determinants of the impact that ScoreSelect had on mean differences observed for citizenship subgroups in Study 1. After conducting a literature review, we postulated that the origins of practically significant changes in mean subgroup differences related to ScoreSelect include the following eight antecedents: (a) a higher initial mean score, (b) greater disposable financial resources, (c) higher need for achievement (a.k.a. achievement motivation), (d) higher self-efficacy for obtaining higher scores upon retesting, (e) greater knowledge of ScoreSelect's existence and basic rules, (f) greater knowledge of graduate programs' GRE score requirements, (g) stronger strategic (i.e., critical) reasoning in terms of how, with ScoreSelect, there is nothing to lose in retesting if time and money are not prohibitive, and (h) greater retesting (and stronger intentions to retest). We briefly explain each of these antecedents and describe the supporting literature that links each antecedent to differences in mean subgroup scores.

Differences in Initial Mean Scores

On the whole, prior research suggests that in encouraging the retaking of an educational assessment, a score choice system will actually increase the score advantage to groups with initially higher mean scores (contrary to the assumption that a regression to the mean upon retesting would reduce mean group differences in GRE scores). Across different educational contexts, researchers and educators have observed a phenomenon of cumulating advantage to initially higher performing groups (see, e.g., Coleman et al., 1966; Cook & Campbell, 1979; Stanovich, 1986). In educational research, Stanovich popularized the term *Matthew effect* to describe this type of increase in educational performance gaps.[12] For example, Proctor and YoungKoung (2010) of the College Board reported that male examinees improved more on the PSAT/NMSQT than did female examinees on all three subtests and that male examinees saw bigger score changes than did female examinees from the PSAT/NMSQT to the SAT. For the same contexts, they also reported that Asian/Asian American and White test takers generally improved more in retesting than did other racial or ethnic groups. Nathan and Camara (1998) and Lyu and Lawrence (1998) had obtained similar findings. However, Rudner (2005) did not find large differences in gains between gender, first language, or age groups for Graduate Management Admission Test (GMAT) retesting.[13]

In a meta-analysis of educational and intelligence tests taken by primary and secondary school students, Kulik et al. (1984) found that examinees with greater initial skills improved more from retesting than did examinees with lower initial skills. The difference in improvement was most pronounced for the oldest segment of students (13 – 17 years of age). The researchers opined that the students with greater initial skills learned from mere practice more quickly than did students with lower initial skills. If skill level is defined by initial test score, then based on the rationale of Kulik et al., one would anticipate that some lower scoring groups would fare increasingly worse against the highest scoring group in a retesting situation. Empirical evidence shows that mean initial GRE scores are often higher for Whites, males, U.S. citizens, those under age 26, and those who have one or both parents or guardians who attended at least some graduate school. However, Chinese citizens, non-U.S., non-Chinese citizens, and Asian Americans score higher on average on GRE-Q than do comparison groups (ETS, 2013; Gallagher et al., 2000; Klieger et al., 2022), and older examinees score higher on average on GRE-V than do younger examinees (ETS, 2013; Trapani, 2013). The latter finding is consistent with research on the effects of aging on cognition, which shows that at least some verbal abilities increase into middle and later adulthood (see, e.g., Salthouse, 2004).

Although ScoreSelect is specific to testing for graduate and professional school admissions and benefits, we note that for employee selection and benefits, there exists a recent and possibly analogous literature regarding the impact of retesting on diversity. However, in the vocational context, findings regarding Matthew effects have been mixed. In a large sample employment selection context, Schleicher et al. (2010) found that across selection ratios for a verbal skills test, voluntary retesting almost always resulted in worse adverse impact ratios for Black, Hispanic, female, and over-40 individuals. The results for a job knowledge test were considerably worse for these minority groups (especially for Black and over-40 individuals). In a promotion context involving a job knowledge test, Van Iddekinge et al. (2011) found that the mean score difference in favor of White individuals over Asian individuals shrank by more than half (from $d = 0.36$ to $d = 0.17$) after retesting. After retesting, the mean score difference where male examinees had the higher average score changed to where female individuals had it (from $d = 0.29$ to $d = -0.18$). For the age comparison, the score predominance of under-40 individuals over 40-and-older individuals expanded considerably after retesting (from $d = 0.08$ to $d = 0.45$).

Differences in Financial Resources

Another possible source of a Matthew effect are financial resources, most notably disposable income. Both the ScoreSelect option that permits the reporting of any subset of test administrations and GRE retesting have a financial cost, and mean income varies across groups based on race/ethnicity, age, gender, educational attainment, and country of citizenship (DeNavas-Walt et al., 2013; World Bank, 2014). Correlations between demographic subgroups and disposable income may indicate differences among subgroups in terms of the ability and motivation to use a score choice option and to retest.[14]

Differences in Need for Achievement and Retesting Self-Efficacy

It is possible that, because of mean-level subgroup differences in need for achievement (a.k.a. achievement motivation) or retesting self-efficacy, subgroups vary in their motivation to use a score choice option or to retest, respectively. There is a body of literature that indicates that differences exist across age groups (see Furchtgott, 1999), racial and ethnic groups, genders, socioeconomic groups, and cultures in education-related achievement motivation and self-efficacy (see review in Wigfield et al., 2008). The extent to which these differences exist across subgroups within a pool of prospective applicants that have graduated from, or who are about to graduate from, college is less clear. We are unaware of prior research that speaks specifically to whether achievement motivation drives use of score choice options or whether subgroups' decisions about retesting would be influenced by differences across subgroups in motivation or self-efficacy.[15]

Furthermore, empirical studies have failed to demonstrate that stereotype threat (i.e., effects on test performance because of stereotypical beliefs that affect motivation and self-efficacy) occurs in high-stakes assessment such as GRE testing, and in many cases these studies demonstrate results contrary to what would be expected if stereotype threat effects were actually present (see Cullen et al., 2004; Sackett & Ryan, 2012; Stricker, 1998; Stricker & Ward, 1998, 2004; Walker & Bridgeman, 2008; Walters et al., 2004). Whether there is a similar lack of effect on decisions about whether or not to retest is unknown, but any such findings may be confounded with the effects of mean-level differences on test scores across subgroups. For instance, a subgroup with a lower average test score might, because of the prevalence of lower scores, decide to retest more often than a higher scoring subgroup does regardless of whether it holds any stereotypical beliefs about test performance.

Differences in Knowledge of ScoreSelect's Existence and Graduate Schools' Score Requirements

Differences across subgroups in (a) lack of knowledge about ScoreSelect's existence and basic rules and (b) lack of knowledge about graduate schools' GRE score requirements would presumably affect the extent to which different subgroups utilize ScoreSelect. We are unaware of prior research about differences across gender, age, and citizenship groups in their knowledge about the rules regarding score reporting or graduate schools' GRE score requirements. Research does show that there are socioeconomic and racial or ethnic group differences in preparedness for the process of applying to higher education (see, e.g., Bowen et al., 2005). It is reasonable to ask whether these differences include score choice and retesting decisions, which can be a complex part of the application process.

Differences in Strategic Reasoning

Differences in critical reasoning about how to engage in score reporting and retesting under a score choice system is another possible cause of mean subgroup differences in scores. To the extent that subgroups differ in experience making these types of strategic decisions, subgroup dif ferences in mean scores may result. It may not be equally understood across subgroups that, if time and costs are not an obstacle, one has nothing to lose by retaking a test under a score choice system. As mean available time and financial resources may vary across subgroups, decision-making about use of a score choice system and retesting might be more of a hypothetical exercise for subgroups with fewer resources. For subgroups with greater resources, use of a score choice system and retesting to maximize the probability of a favorable decision might be taken for granted.

Differences in Retesting Behaviors

As mentioned previously, use of a score choice option encourages retesting behavior, which potentially affects reported scores. In fact, we believed that usages of a score choice option and retesting behavior have a reciprocal relationship with each other. At the same time a score choice system encourages retesting, the existence of multiple sets of scores from retesting can encourage greater use of a score choice system. For instance, if an examinee reports one score from a test administration (e.g., GRE-V), the examinee must also report all other scores from that same administration (i.e., GRE-Q and GRE-AW). Moreover, an applicant can send score reports to institutions with different score expectations (e.g., a program that cares only about high GRE-Q scores or another program that equally weights GRE-Q and GRE-V). Consequently, when one has multiple sets of scores, score reporting decisions can become extremely complicated. With multiple sets of scores, one can use ScoreSelect in different ways to execute different score reporting strategies.

Study 2 Method

Instrument

An online survey was developed to explore possible ScoreSelect-related determinants of the changes in mean score differences between those subgroups. The survey text appears in the Appendix C. The operational survey contained logic branching so that any follow-up questions asked of respondents were relevant. Questions were designed to measure the possible determinants or whether the respondent's responses in general were even pertinent. For instance, in Question 8, a response of *strongly agree* to "A graduate degree provides me with better career opportunities" would indicate a high need to achieve sufficiently high GRE scores. As another example, a response to Question 3 of "I am enrolled in a graduate degree program; I began the program in the year 2010" would indicate that the respondent never actually had to consider a score choice option, because ScoreSelect first became available in 2012. Therefore, the respondent's answers in general would be excluded from analyses.

Research Design

Data were collected from 1,500 GRE examinees across the racial/ethnic, gender, age, citizenship, and parental education categories previously described for Study 1. The final comparison included 1,298 U.S. citizens, 93 Chinese citizens, and 93 non-U.S., non-Chinese citizens.

Study 2 Hypotheses

We hypothesized that where there is a practically significant change in differences for citizenship subgroups, subgroups that benefit from the change will have

- a higher initial mean score;
- greater disposable financial resources;
- higher need for achievement (a.k.a. achievement motivation);
- higher self-efficacy for obtaining higher scores upon retesting;
- greater knowledge of ScoreSelect's existence and basic rules;
- greater knowledge of graduate programs' GRE score requirements;
- stronger strategic (i.e., critical) reasoning in terms of how, with ScoreSelect, there is nothing to lose in retesting if time and money are not prohibitive; and
- greater retesting (and stronger intentions to retest) because of a – g.

Study 2 Results

Our hypotheses received partial support. Survey results provided evidence for many of the theoretical determinants of changes in mean subgroup differences related to ScoreSelect. Table 1 shows the results of statistical significance tests: analysis of variance (ANOVA), for mean differences and chi-square for response frequency differences, run on data from the survey that appears in Appendix C.

Table 1 Survey Characteristics by Citizenship Subgroups

Survey item no.	Description of survey item or response option content	U.S. citizens	Chinese citizens	Non-U.S., non-Chinese citizens	Chinese citizens compared to U.S. citizens			Non-U.S., non-Chinese citizens compared to U.S. citizens		
		Mean (SD)	Mean (SD)	Mean (SD)	*F*	*p*		*F*	*p*	
1	Avg. Verbal Reasoning Score	155.5 (7.54)	151.8 (5.57)	151.6 (6.73)	22.45	**.000**		23.87	**.000**	
1	Avg. Quant. Reasoning Score	152.5 (7.88)	166.4 (3.21)	160 (6.42)	287.69	**.000**		80.28	**.000**	
1	Avg. Analytical Writing Score	4.1 (0.74)	3.2 (0.37)	3.6 (0.62)	132.79	**.000**		50.97	**.000**	
13	No. Hrs. Spent Prep. GRE[a]	62.8 (100.70)	301.6 (300.5)	167.3 (201.3)	313.26	**.000**		76.90	**.000**	
14	No. Times Taken GRE[a] Past 5 years	1.4 (0.75)	1.8 (0.85)	1.5 (0.73)	20.45	**.000**		0.00	.962	
		n (%)	*n* (%)	*n* (%)	χ^2	*p*	OR	χ^2	*p*	OR
6	Interested in a Doctorate Program	484 (42%)	43 (47%)	31 (35%)	0.89	.344	1.2	1.83	.177	0.7
7	Importance of Attend. Grad School	989 (78%)	60 (65%)	80 (86%) (99%)	7.25	**.007**	0.5	3.67	**.055**	1.8
8	Grad. Deg. Provides Opportunity	1,250 (98%)	90 (98%)	89 (96%)	0.10	.754	0.8	3.07	**.080**	0.4
8	Benefits of Deg. Outweighs Cost	1,108 (87%)	85 (92%)	79 (85%)	2.00	.157	1.8	0.46	.497	0.8
8	Grad. Deg. Increases Income	1,186 (94%)	81 (89%)	84 (91%)	3.81	**.051**	0.5	1.20	.274	0.7
8	Grad. Deg. Better Than Work Exper.	859 (68%)	61 (66%)	60 (65%)	0.11	.743	0.9	0.47	.493	0.9
8	Attending Grad. School Goal	1,211 (95%)	85 (93%)	91 (99%)	0.77	.379	0.7	2.51	.113	4.4
9	GRE[a] Scores Important for Admit	894 (70%)	75 (82%)	79 (85%)	5.45	**.020**	1.9	13.23	**.004**	2.4
10	GRE[a] Scores of Accepted Students Avail.	830 (65%)	74 (80%)	64 (70%)	9.01	**.003**	2.2	0.76	.384	1.2
11	Am Seeking Financial Assistance	785 (61%)	69 (75%)	71 (76%)	6.69	**.010**	1.9	8.18	**.004**	2.0
12	GRE[a] Scores Important for Aid	344 (44%)	44 (64%)	46 (65%)	10.12	**.001**	2.3	11.48	**.001**	2.4
15	Satisfied W/Verbal Reasoning Score	882 (69%)	46 (50%)	59 (63%)	14.38	**.000**	0.4	1.30	.254	0.8
15	Satisfied W/Quant. Reasoning Score	716 (56%)	72 (79%)	67 (72%)	18.22	**.000**	2.9	8.84	**.003**	2.0
15	Satisfied W/Analytic Writing Score	692 (55%)	29 (32%)	44 (47%)	18.30	**.000**	0.4	1.84	.175	0.7
16	Plan to Take GRE[a] Again	208 (16%)	17 (18%)	23 (25%)	0.30	.584	1.2	4.41	**.036**	1.7
17	Retesting: Believe Improve Score	162 (13%)	12 (14%)	17 (20%)	0.03	.855	1.1	3.21	**.073**	1.7
17	Retesting: Need Higher Scores	130 (11%)	11 (13%)	13 (16%)	0.33	.567	1.2	2.11	.147	1.6
17	Retesting: Nothing to Lose	37 (3%)	1 (1%)	8 (10%)	0.93	.334	0.4	10.08	**.001**	3.4
17	Retesting: Despite Anxiety Worth It	70 (6%)	1 (1%)	5 (7%)	3.00	**.083**	0.2	0.06	.804	1.1
17	Retesting: for Other Reasons	27 (2%)	3 (4%)	5 (7%)	0.57	.450	1.6	4.98	**.026**	2.9
18	Retesting Costs Too Much Money	281 (26%)	12 (16%)	16 (23%)	3.84	**.050**	0.5	0.38	.535	0.8
19	If I Study for VR Can Raise Scores	965 (76%)	87 (95%)	71 (77%)	16.69	**.000**	5.5	0.05	.816	1.1
19	I'm Not a Good Test Taker	479 (38%)	40 (43%)	27 (30%)	1.13	.287	1.3	2.24	.135	0.7

Table 1 Continued

		n (%)	n (%)	n (%)	χ^2	p	OR	χ^2	p	OR
19	If I Study for QR Can Raise Scores	1,056 (84%)	58 (64%)	74 (81%)	22.99	**.000**	0.3	0.32	.570	0.9
19	AW Scores Will Not Improve Matter	573 (45%)	31 (34%)	29 (32%)	4.44	**.035**	0.6	5.94	**.015**	0.6
19	Scores Will Improve Now I Know GRE[a]	810 (64%)	67 (73%)	67 (74%)	2.84	**.092**	1.5	3.35	**.067**	1.6
19	I Cannot Afford to Take GRE[a] Again	727 (57%)	54 (59%)	59 (64%)	0.06	.805	1.1	1.60	.205	1.3
22	Know Can Send No/Some/All Scores	732 (57%)	60 (65%)	52 (56%)	2.19	.139	1.4	0.07	.791	0.9
22	Know Can Send Only Highest GRE[a] Scores	471 (37%)	53 (58%)	48 (52%)	15.60	**.000**	2.3	7.99	**.005**	1.8
22	Know Can Send 4 Free Reports	946 (74%)	56 (61%)	70 (75%)	7.63	**.006**	0.5	0.06	.800	1.1
22	Know $25 Cost to Send Reports	765 (60%)	51 (55%)	67 (72%)	0.71	.399	0.8	5.35	**.021**	1.7
22	Know When Can Pick Any Scores	185 (14%)	16 (17%)	24 (26%)	0.58	.447	1.2	8.59	**.003**	2.1
24	Same Day: Free of Charge	986 (79%)	49 (58%)	70 (78%)	20.12	**.000**	0.4	0.04	.838	0.9
24	Same Day: Scores High Enough	344 (56%)	16 (31%)	32 (62%)	12.59	**.000**	0.3	0.54	.464	1.2
24	Same Day: Didn't Know Opt. Not Send	51 (16%)	3 (8%)	4 (17%)	1.88	.170	0.4	0.01	.936	1.0
24	Same Day: Didn't Know Opt. Later	62 (19%)	0 (0%)	6 (23%)	8.17	**.004**	0.0	0.28	.598	1.3
28	Not Sent Later: Could Not Afford	103 (11%)	2 (2%)	6 (7%)	5.54	**.019**	0.2	0.90	.343	0.7
28	Not Sent Later: Did Not Need To	339 (28%)	10 (11%)	11 (13%)	12.06	**.001**	0.3	9.91	**.002**	0.4
28	Not Sent Later: Did Not Know	31 (3%)	1 (1%)	0 (0%)	1.13	.288	0.4	2.72	**.099**	0.0
29	If Time/Money Not An Issue Nothing To Lose	601 (47%)	68 (74%)	67 (72%)	24.69	**.000**	3.2	21.58	**.000**	2.9
30	Knowing ScoreSelect-More Likely	469 (37%)	45 (48%)	50 (54%)	5.59	**.018**	1.7	11.53	**.001**	2.0
30	Knowing ScoreSelect-Less Likely Retake	26 (2%)	7 (8%)	6 (7%)	11.43	**.001**	4.0	7.64	**.006**	3.3
30	Knowing ScoreSelect-Neither	779 (61%)	39 (43%)	36 (39%)	11.71	**.001**	0.5	16.24	**.000**	0.4

Note. Full text of the survey items appears in Appendix C. Means and standard deviations are provided for numerical variables; frequencies (Ns); and percents are provided for categorical variables. Odds ratio (OR) indicates how many times more likely it is that an individual in the comparison group (i.e., Chinese citizen or non-U.S., non-Chinese citizen) reports a characteristic than an individual in the control group (U.S. citizen). VR = verbal reasoning; QR = quantitative reasoning; AW = analytical writing. Bolded values indicate $p < .10$. [a]Indicates GRE General Test.

Means or frequencies of U.S. citizens' responses are compared to the means or frequencies of Chinese citizens and then separately to the means or frequencies of non-U.S., non-Chinese citizens. The table also reports effect sizes (odds ratios, or ORs). For parsimony, we collapsed response categories on the survey so that chi-square analyses were run on 2 × 2 contingency tables. For example, for Question 7 ("How important is it to you to attend graduate or professional school?"), we collapsed *very important* with *important* and *not very important* with *not important at all*. We treated responses to multiple-response questions (e.g., Question 18, "Please check the statement(s) that best describe your reasons for not taking the GRE General Test again") as a dichotomous ("Yes" vs. "No") outcome for each citizenship group. There were many significance tests run and thus an increased familywise error rate. Moreover, the survey results cannot unequivocally establish causation. Nevertheless, we believed that the following analyses (especially where resulting p values would be especially low) would yield useful information and help frame future research on the determinants of score choice behaviors.

Differences in Initial Mean Scores

Evidence suggested that higher initial GRE scores have no effect on ScoreSelect-related changes in mean subgroup differences. Specifically, where there was a practically significant ScoreSelect-associated change in subgroup differences, evidence suggested that Chinese and non-U.S., non-Chinese groups that benefit from the change had a higher initial score than U.S. citizens only on GRE-Q. In general, Chinese and non-U.S., non-Chinese examinees score higher on average on GRE-Q and lower on GRE-V and GRE-AW in comparison to U.S. citizens (ETS, 2013). Those findings replicated for the survey sample as well in which ANOVA results indicated statistically significant differences (p = .000 for survey Item 1 in Table 1). Therefore, evidence suggests that ScoreSelect-associated change in subgroup differences for GRE-V and GRE-AW for the benefit of non-U.S. citizens do not seem to be driven by capitalization on initial score advantages (i.e., Matthew effects). It is questionable whether that capitalization is occurring for GRE-Q either (where non-U.S. citizens saw a 5%–6% improvement in the admissions ratio relative to U.S. citizens; see Table A5), because practically significant improvement for non-U.S. citizens occurred only for ASRs and not free ones.

Differences in Financial Resources

Contrary to what one might assume given concerns that greater financial resources would unfairly advantage wealthier examinees in general (see Matthews, 2009), evidence did not show that greater disposable financial resources advantage U.S. citizens in terms of their having a greater ability than non-U.S. citizens to improve GRE scores via ScoreSelect and retesting. U.S. citizens generally have higher average incomes than citizens of most other countries, including China (World Bank, 2014). This fact was evidenced by the finding that more Chinese and non-U.S., non-Chinese citizen examinees sought financial assistance than did U.S. citizens (survey Item 11; ps = .010 and .004) suggesting that (a) U.S. citizen examinees in particular had greater disposable financial income than did non-U.S. citizens, (b) U.S. citizens were not as sensitive to financial issues, or (c) both. If U.S. citizens more frequently have disposable financial resources for ScoreSelect and retesting, the evidence suggests that those greater resources do not lead to greater use of ScoreSelect and retesting to improve GRE scores. Unexpectedly, U.S. citizens more often than Chinese citizens wanted to take advantage of free reports available on the test date (Item 24; p = .000); and especially compared to Chinese citizens, U.S. citizens more frequently reported an inability to afford the cost of sending test scores after the test date (Item 28; p = .019). In general, only a minority of all three citizen groups thought the retesting was too costly (16%–26%; see Item 18). There was no statistically or practically significant difference between U.S. and non-U.S., non-Chinese citizens in terms of their opinion about retesting being too expensive (Item 18; p = .535); for Chinese citizens, the difference was statistically significant (p = .050), but it was U.S. citizens (26%) who, compared to Chinese citizens (16%), more frequently thought retesting to be too expensive. When more directly queried if they themselves individually could afford to retake the GRE test because of its cost (Item 19), there was no statistically significant difference between the non-U.S. citizens groups and U.S. citizens (ps = .205 and .805).

Differences in Need for Achievement and Retesting Self-Efficacy

Several of the analyses about group differences in achievement motivation and retesting self-efficacy did not appear to explain any clear relationships of a score choice option to changes in mean group differences. U.S. citizens more frequently reported that attending graduate school was important or very important than did Chinese citizens (Item 7; 78% vs. 65%; $p = .007$), but U.S. citizens less frequently reported that attending graduate school was important or very important than did non-U.S., non-Chinese citizens (Item 7; 78% vs. 86%; $p = .055$). There were no statistically significant findings for group differences in the interest in obtaining a doctorate degree (for which GRE scores are often important) versus a nondoctorate degree (for which GRE scores are less frequently important); see Item 6; ps = .344 and .177. Although U.S. citizens in comparison to non-U.S., non-Chinese citizens more frequently thought that a graduate or professional degree provided better career opportunities (Item 8; $p = .080$), that difference was not practically significant (98% vs. 96%). Although U.S. citizens in comparison to Chinese citizens more frequently thought that a graduate or professional degree would increase income potential (Item 8; $p = .051$), that difference was not practically significant (94% vs. 89%). Citizenship groups viewed quite similarly cost–benefit analyses of graduate school attendance, the importance of graduate school attendance versus additional work experience, and attendance of graduate school as a personal goal (Item 8; $ps > .100$).

However, findings did suggest that the differences among citizenship groups' perceptions of the role of the GRE in reaching examinees' admission and funding goals are encouraging the use of ScoreSelect to achieve those goals. There were statistically and practically significant differences among groups in terms of the importance with which GRE scores are perceived for gaining admission and obtaining financial assistance. Chinese and non-U.S., non-Chinese citizens viewed the GRE as more important for achieving both purposes than did U.S. citizens (Items 9 and 12; $.001 \leq ps \leq .020$). Group differences in the self-reported number of hours spent preparing for the GRE (Item 13) provided behaviorally based support for that viewpoint ($p = .00$, with non-U.S. citizens devoting 1.8 to 4.8 times as many hours to GRE test preparation).[16]

There is some evidence that, under ScoreSelect, Chinese and non-U.S., non-Chinese examinees have higher retesting self-efficacy than U.S. citizens do, but this evidence is possibly complicated by existing differences in how citizenship groups generally perform on GRE subtests. There are no statistically significant citizenship group differences in testing self-efficacy in general (Item 19, "I'm Not a Good Test Taker"; ps = .287 and .135). Chinese and non-U.S., non-Chinese examinees more frequently experienced higher testing self-efficacy than did U.S. citizens once they had seen what the GRE test was like (Item 19; ps = .092 and .067). Overcoming test anxiety did not appear to be major part of retesting considerations for any of the citizenship groups (Item 17, with no more than 7% of any group indicating the contrary), but Chinese citizens considered test anxiety even less often than did U.S. citizens (Item 17; $p = .083$). Chinese and non-U.S., non-Chinese citizens thought more than did U.S. citizens that they could improve their scores in general (Item 19; ps = .092 and .067).

For specific GRE sections, the results more clearly showed group differences. This is possibly due to actual differences among citizenship groups in performance on specific GRE sections. Chinese citizens more often than U.S. citizens believed that they could raise their GRE-V and GRE-AW scores with further study (Item 19; ps = .000 and .035). Non-U.S., non-Chinese citizens believed more frequently than U.S. citizens that they could raise their GRE-AW scores with further study (Item 19; $p = .015$), but not so for GRE-V scores (Item 19; $p = .816$). U.S. citizens more often felt that they could raise their GRE-Q scores than did Chinese citizens (Item 19; $p = .000$) but not non-U.S., non-Chinese citizens (Item 19; $p = .570$). These findings for specific subtests might be based on realistic self-appraisals of how the groups' members generally perform on these subtests: If a group commonly scored lower on the subtest in question, then the subgroup felt that there was more opportunity for improvement, and vice-versa. Differences in subgroup responses provided some evidence for this finding (Item 15; with four out of six p values $\leq .003$), although two of the differences between U.S. citizens and non-U.S., non-Chinese citizens were not statistically significant (Item 15; ps = .254 and .175).

Differences in Knowledge of ScoreSelect's Existence and Graduate Schools' Score Requirements

In comparison to U.S. citizens, both Chinese citizens and non-U.S., non-Chinese citizens more frequently reported being familiar with ScoreSelect when they most recently took the GRE test (Item 20; ps = .000). Findings regarding group differences in familiarity with ScoreSelect's basic rules were mixed, with differences that were not always statistically or practically significant (see the five response options to Item 22 in Table 1). Compared to U.S. citizens, Chinese citizens most frequently reported knowledge of the option to send scores after a test date, but this finding was statistically significant for only one of the two items that measured this difference in knowledge (Item 24; $p = .004$; Item 28; $p = .288$). There was no statistically significant difference between U.S. citizens and either Chinese or non-U.S., non-Chinese citizens in terms of familiarity with the option not to send any scores on the test date (Item 24; ps = .170 and .936). Compared to U.S. citizens (who reported for themselves), Chinese citizens reported more frequently that the GRE scores of students accepted to the programs to which the Chinese citizens applied were available (Item 10; $p = .003$). This finding was not statistically significant for a comparison of non-U.S., non-Chinese citizens to U.S. citizens (Item 10; $p = .384$).

Differences in Strategic Reasoning

Both Chinese citizens and non-U.S., non-Chinese citizens were substantially more aware than were U.S. citizens that they had nothing to lose in retesting if time and money are not prohibitive (Item 29; *p*s = .000). Even after being told of the benefits of ScoreSelect, expressing equivalent desires to attend graduate school (including doctorate programs), and indicating less need for financial assistance, U.S. citizens were not nearly as likely as Chinese citizens or non-U.S., non-Chinese citizens to take the GRE test again (Item 30; *p*s = .018 and .001). We do not know whether this is due to persistently fewer U.S. citizens understanding ScoreSelect's benefits, lower motivation of U.S. citizens to change retesting behavior, U.S. citizens having greater financial concerns, or other reasons.

Differences in the Frequency of Retesting and Intentions to Retest

Taken as a whole, external data and survey results suggest that non-U.S. citizens generally retested (and intended to retest) more often than did U.S. citizens. However, results are not fully conclusive. From August 2011 to December 2013 (inclusive of the July 2012 – July 2013 data in this study), the percentages of Chinese citizens who retested (approximately 38%, $N = 41{,}557$) and non-U.S., non-Chinese citizens who retested (approximately 16%, $N = 56{,}255$) was larger than the percentage of U.S. citizens who retested (approximately 11%, $N = 91{,}183$) in a personal communication from F. Robin (September 10, 2014).[17] From the survey, Chinese citizens reported more retesting than did U.S. citizens (Item 14; $p = .000$). Responses to Item 15 (satisfaction with GRE scores) suggested that Chinese citizens were more dissatisfied with their GRE-V and GRE-AW scores than were U.S. citizens (*p*s = .000), but U.S. citizens were more dissatisfied with their GRE-Q scores than were Chinese citizens ($p = .000$). Although non-U.S., non-Chinese citizens reported more retesting than did U.S. citizens (Item 14; 1.5 times vs. 1.4 times), those findings were not statistically significant (Item 14; $p = .962$). Self-reports of intentions to retake the GRE test (Item 16) indicate that non-U.S., non-Chinese examinees more frequently intended to retest than did U.S. citizens ($p = .036$); such was the case for Chinese citizens (18% vs. 16%), but the difference was not statistically significant ($p = .584$).

Combined Discussion and Conclusions

Our investigation of the possible effects of a score choice option on mean subgroup differences in test scores often led to results inconsistent with our expectations. ScoreSelect is associated with relatively few practically significant changes in mean group differences (Tables A1 – A5). The mean score differences that we did observe did not increase to a practically significant extent in favor of initially higher performing groups, although with ScoreSelect, non-U.S. citizens would somewhat increase their representation in graduate school where admission is at least moderately selective, top-down, and based solely on GRE-Q. Results suggest that non-U.S. citizens (in particular Chinese citizens) used ScoreSelect to close score gaps on GRE-V and GRE-AW, subtests on which their earlier performance was unsurprisingly lower than that of U.S. citizens on average. Differences in citizenship-based patterns of ScoreSelect use (Figures 1 and 2; Tables B1 and B2) lent additional support to the view that ScoreSelect is contributing to changes in how the citizenship groups are performing relative to each other.

Notwithstanding the foregoing, we believe that the impact of ScoreSelect on the admissions rates of the demographic groups studied is likely to be even less than the few, small, and generally undetectable effects described in this article. How decision-making operates in many real-world contexts is likely to dilute the impact simulated in this article. For example, graduate and professional programs are not always very selective, and to the extent that they are selective, they consider admissions information in addition to GRE scores (e.g., letters of recommendation) and employ selection methods other than pure top-down selection (e.g., holistic or impressionistic evaluation; see, e.g., Walpole et al., 2002). Thisarticle cannot simulate all possible admissions systems or even any single admissions system perfectly. Rather, this article illustrates the impact of ScoreSelect on diversity in a so-called worst-case scenario.

In an attempt to better understand why ScoreSelect might contribute to the few practically significant changes that it does, the online surveys that we developed, administered, and analyzed (see Appendix C) presented mixed findings that partly support our expectations. Results were strong enough to suggest that the determinants of changes in mean group differences (and thus admission and funding rates) related to ScoreSelect and retesting may include the following:

- non-U.S. citizens' more common belief that the GRE is important to gaining admissions and obtaining funding;
- non-U.S. citizens performing worse than U.S. citizens on GRE-V and GRE-AW (i.e., on the majority of the three GRE subtests), possibly leading to a greater motivation to retest to improve GRE scores;
- non-U.S. citizens' more common strategic reasoning about how retesting along with ScoreSelect can be used to maximize the probability of obtaining favorable admissions and funding decisions; and
- non-U.S. citizens' greater retesting (possibly due to the foregoing factors).

Chinese citizens' greater knowledge of graduate programs' GRE score requirements may also have played a role in changing mean group differences in scores, but that finding is based on just a single survey item. Most of the analyses regarding group differences in achievement motivation and retesting self-efficacy did not reveal clear patterns relating the score choice option to changes in mean group score differences. If differences in groups' knowledge of ScoreSelect's existence and basic rules contributed to changes in mean score differences, the results of our survey analyses did not clearly bear out that finding. Because non-U.S. citizens do not in general have higher initial mean scores on GRE-AW and GRE-V than do U.S. citizens, we could not conclude that Matthew effects were a primary driver of changes in group score differences. Greater disposable financial resources are not a clear driver either: In fact, where we expected U.S. citizens to have greater resources to take advantage of ScoreSelect and retesting, we instead found that where there were any differences in groups' perspectives on affordability, U.S. citizens were more concerned about cost.

Further investigation of the impact of score choice options, and its determinants, is warranted. ScoreSelect (and score choice in general) is relatively new, and findings for Study 1 might change in the future. In addition, some of the sample sizes in Study 2 were not particularly large, and some of the results in Study 2 were inconsistent. The number of items in the survey was small, and one usually measures constructs such as motivation, self-efficacy, and test anxiety with at least several items. Furthermore, Study 2 did not analyze linkages between determinants. Therefore, future studies should use structural equation models (e.g., path analysis) to better understand the determinants of score choice outcomes. It is unclear if the results in either study would replicate for other populations (e.g., applicants to college, job applicants). Also, the impacts of a score choice system and their determinants may depend on the specific features of the system (e.g., cost, specific score reporting rules). We also recommend studies on the possible impact of a score choice option on the validity of scores. Given the lack of published research on score choice options, we believe that our studies raise important questions and provide a valuable framework for future analyses.

Notes

1 Consistent with Nunnally (1978, p. 190), *measurement error* discussed in this article includes both error that is purely random and generally beyond human control (e.g., examinee illness) as well as systematic error (e.g., group differences in prior knowledge about the kinds of items on a test).

2 This article uses the terms *higher scoring group, initially higher scoring group, higher performing group*, and *initially higher performing group* synonymously. Similarly, it uses the terms *lower scoring group, initially lower scoring group, lower performing group*, and *initially lower performing group* synonymously.

3 We hypothesize that ScoreSelect will be associated with practically and statistically significant changes in mean score differences between subgroups for reasons detailed later in this article, but we were unable to locate previous studies regarding the effects of score choice options on such differences and to hypothesize the specific size of the expected changes. An investigation of the effects of retesting on subgroup score differences was conducted; however, even if one can reasonably separate out the effects of score selection from any interaction effects of score selection with retesting, the generalizability of prior research findings for retesting effects is questionable given that the effects of retesting may be specific to (a) the nature of the examinee population in question, (b) the nature of the test in question, (c) the use of the test scores, and (d) specific features of the retesting system. Given the foregoing limitations, we do not offer a specific, expected magnitude of change in score differences.

4 Examinees do not see previous scores when making this decision.

5 Each subgroup mean is based on (a) taking an average of scores in each score report belonging to a member of the subgroup and then (b) averaging across those score report averages. Where a report contains just a single set of scores, those scores were included the same as the average for a report with multiple sets of scores.

[6] Due to the considerable sample sizes here, one might find statistically significant differences between *d* values (or nonoverlapping 95% confidence intervals for the *No SS* and *SS d* values reported in Tables A1–A4) when the impacts (differences in admission rates) are not practically significant. Given the very large sample sizes, the reverse (i.e., practical significance without statistical significance) was extremely unlikely.

[7] Top-down selection refers to admission of the applicant with the highest score first, of the applicant with the second highest score second, and so forth. Top-down selection can be thought of as the most stringent version of cut score usage that is most likely to reveal any potential effects of ScoreSelect on admissions. In order to precisely communicate any possible effects of ScoreSelect on admissions, this article simulates admissions based on GRE-V scores only, GRE-Q scores only, and GRE-AW scores only.

[8] This condition is hypothetical, because the scores analyzed were reportable during a time period when ScoreSelect was available (between July 1, 2012, and July 2013). Therefore, during this period, at least some examinees reported fewer than all reportable scores (see Tables B1 and B2).

[9] We discontinued an approach in which we compared scores reported before the introduction of ScoreSelect to scores reported after the introduction of ScoreSelect. We needed to use scores from only the current version of the GRE test in order to avoid confounding effects from ScoreSelect with effects based on the version of the GRE test. By using the *Hypothetical All versus Actually Sent* approach instead, there were fewer scores from the previous version of the GRE test that we had to omit. Also, the Hypothetical All versus Actually Sent approach avoided potential idiosyncrasies associated with the transition to a new version of the GRE test not long before the introduction of ScoreSelect (e.g., a possibly atypical examinee population around the time of, and due to, the transition).

[10] One reviewer asked whether there were any notable differences in ScoreSelect usage and/or survey responses (discussed below) among different U.S. citizen subgroups (e.g., women/men, White examinees/racial minority subgroups). Given that, the ultimate purpose of these studies was a practical one, to detect and explain possible impacts on admissions rates—impacts which we eventually determined did not occur outside of citizenship comparisons—we concluded that investigating the answers to these questions for U.S. citizen subgroups was beyond the scope of this article.

[11] At least in theory, a score choice option can directly or indirectly influence a score reporting decision where only one set of scores exists. For example, an examinee might choose not to report the examinee's only available set of scores with the perspective that scores will improve upon retesting and only the higher scores would be reported under a score choice option.

[12] The term derives from the following passage in the Book of Matthew: “For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken even that which he hath.” (*King James Bible*, Matthew 25:29, n.d.). Sometimes the term *fan-spread* has been used to describe a differential growth pattern that a Matthew effect produces (see Walberg & Tsai, 1983, whose use of the terms *Matthew effect* and fan-spread predates Stanovich's popularization of the term Matthew effect in 1986).

[13] One reviewer pointed out that these findings might reflect a gender-by-motivation interaction. Except for students who think they have a relatively high probability of receiving a National Merit Scholarship via high PSAT scores, examinees might generally perceive the PSAT/NMSQT as low stakes. Male examinees may be more likely to discount a low stakes test than females, whereas both genders might be relatively motivated to perform well on the SAT given its role in admissions. If the finding for PSAT reflects a gender-by-motivation interaction, then it would not be anticipated to arise in two Graduate Management Admission Test (GMAT) administrations that one might expect examinees to perceive as relatively high stakes given the GMAT's role in admissions.

[14] We do not know the extent to which these general differences in resources replicate specifically for GRE examinees or graduate school students. We could not find in the literature any direct comparisons of the financial resources of (a) Chinese citizen examinees (or graduate students) or non-U.S., non-Chinese citizens examinees (or graduate students) to the financial resources of (b) U.S. citizen examinees (or graduate students). Sources we found excluded U.S. citizens from analyses, combined graduate students with undergraduate students, and/or were limited to samples of prospective graduate students who had paid for educational consulting services. Also, graduate school costs for non-U.S. citizens (Chinese or otherwise) may differ from those for U.S. citizens, but we were unable to locate sources that provided comparison statistics.

[15] There is some evidence that a substantial percentage of citizens of China and India who seek master's degrees believe that there are more quality higher education opportunities outside of their home countries than within them (see Chang et al., 2014, but note that the respondents were limited to those who had paid for educational consulting services). Given that many U.S. graduate programs use the GRE test in making admissions and funding decisions and that non-U.S. academic credentials may be harder for U.S. institutions to interpret (see Walpole et al., 2002), this evidence about perceptions of quality higher educational opportunities—if it generalizes to GRE examinees—might suggest a greater motivation of non-U.S. citizen examinees to achieve and to report higher GRE scores.

[16] One reviewer suggested the possibility that some citizens of China view strong GRE performance as a way to obtain an assistantship that would increase the probability of receiving a visa to study in the United States.

17 The numbers are approximations, because the values cited here for Chinese citizens (mainland China/the People's Republic of China) include values for Hong Kong, Taiwan, and South Korea. The vast majority of this aggregated group consists of citizens of mainland China, however.

References

Bowen, W. G., Kurzweil, M. A., & Tobin, E. M. (with Pichler, S. C.). (2005). *Equity and excellence in American higher education*. University of Virginia Press.

Chang, L., Schulmann, P., & Lu, Z. (2014). *Bridging the digital divide: Segmenting and recruiting international millennial students*. (Report 06). World Education News & Reviews. https://knowledge.wes.org/rs/worldeducationservice/images/RAS-Paper-06-Bridging-the-Digital-Divide-Segmenting-and-Recruiting-International-Millenial-Students.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, F., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. U.S. Government Printing Office.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin Company.

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology*, *89*(2), 220–230. https://doi.org/10.1037/0021-9010.89.2.220

DeNavas-Walt, C., Proctor, B. D., & Smith, J. C. (2013). *Income, poverty, and health insurance coverage in the United States: 2012* (Current Population Reports P60-245). U.S. Department of Commerce, Economics and Statistics Administration & U.S. Census Bureau. https://www2.census.gov/library/publications/2013/demo/p60-245/p60-245.pdf

ETS. (2013). *A snapshot of the individuals who took the GRE revised General Test: August 2011–June 2012*. https://www.ets.org/research/policy_research_reports/publications/report/2013/jork

Furchtgott, E. (1999). *Aging and human motivation*. Springer https://doi.org/10.1007/978-1-4757-4463-7

Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). *The effect of computer-based tests on racial/ethnic, gender, and language group* (GRE Board Report No. GREB-96-21P). ETS. https://doi.org/10.1002/j.2333-8504.2000.tb01831.x

Goodman, W. (2011, September 29). *Flash musical: College exam interrupted by variation on "Les Miserables."*. http://www.cbsnews.com/news/flash-musical-college-exam-interrupted-by-variation-on-les-miserables/

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger Publishers.

King James Bible, Matthew 25:29, n.d.

Klieger, D., Bochenek, J., Ezzo, C., Holtzman, S., Cline, F. & Olivera-Aguilar, M. (2022). Using third-party evaluations to assess socioemotional skills in graduate and professional school admissions. *International Journal of Testing*, *22*(1), 72–99, https://doi.org/10.1080/15305058.2021.2019748

Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*(2), 435–447. https://doi.org/10.3102/00028312021002435

Lyu, C. F., & Lawrence, I. (1998). *Test-taking patterns and average score gains for the SAT* (Unpublished Statistical College Board Report No. SR-98-05). ETS.

Matthews, J. (2009, January 16). Misguided colleges skewer Score Choice. *Washington Post*. http://www.washingtonpost.com/wp-dyn/content/article/2009/01/16/AR2009011601123.html

Nathan, J. S., & Camara, W. (1998). *Score change when retaking the SAT® I: Reasoning test*. *Research notes RN-05*. T he College Board.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Proctor, T. P., & YoungKoung, R. K. (2010). *Score change for 2007 PSAT/NMSQT® Test-takers: An analysis of score changes for PSAT/NMSQT test-takers who also took the 2008 PSAT/NMSQT test or a spring 2008 SAT® test* (Research Notes RN-41). The College Board.

Rudner, L. M. (2005). *Examinees retaking the Graduate Management Admission Test™*. Research Reports No. 05-01. Graduate Management Admission Council.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*(3), 707–721. https://doi.org/10.1111/j.1744-6570.1997.tb00711.x

Sackett, P. R., & Ryan, A. M. (2012). Concerns about generalizing stereotype threat research findings to operational high-stakes testing. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, processes, and application* (pp. 249–263). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199732449.003.0016

Sackett, P. R., & Wilk, S. L. (1994) Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*(11), 929–954. https://doi.org/10.1037/0003-066X.49.11.929

Salthouse, T. A. (2004). Localizing age-related individual differences in a hierarchical structure. *Intelligence*, *32*(6), 541–561. https://doi.org/10.1016/j.intell.2004.07.003

Schleicher, D. J., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2010). If at first you don't succeed, try, try again: Understanding race, age, and gender differences in retesting score improvement. *Journal of Applied Psychology*, *95*(4), 603–617. https://doi.org/10.1037/a0018920

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. https://doi.org/10.1598/RRQ.21.4.1

Stricker, L. J. (1998). *Inquiring about examinees' ethnicity and sex: Effects on AP® Calculus AB Examination performance* (Research Report No. RR 98–05; College Board Report No. 98–01). ETS and College Board. https://doi.org/10.1002/j.2333-8504.1998.tb01754.x

Stricker, L. J., & Ward, W. C. (1998). *Inquiring about examinees' ethnicity and sex: Effects on computerized placement tests performance* (Research Report No. RR 98-09; College Board Report No. 98-02). ETS and College Board. https://doi.org/10.1002/j.2333-8504.1998.tb01758.x

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test taker's ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology*, *34*(4) 665–693. https://doi.org/10.1111/j.1559-1816.2004.tb02564.x

Trapani, C. S. (2013). *Multilevel modeling of cognitive ability in highly functioning adults* [Unpublished doctoral dissertation]. Fordham University.

Van Iddekinge, C. H., Morgeson, F. P., Schleicher, D. J., & Campion, M. A. (2011). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology*, *96*(5), 941–955. https://doi.org/10.1037/a0023562

Walberg, H. J., & Tsai, S-L. (1983). Matthew effects in education. *American Educational Research Journal*, *20*(3), 359–373. https://doi.org/10.3102/00028312020003359

Walker, M. E., & Bridgeman, B. (2008). *Stereotype threat spillover and SAT® scores* (Research Report No. RR-08-28; College Board Research Report No. 2008-2). ETS and The College Board. https://doi.org/10.1002/j.2333-8504.2008.tb02114.x

Walpole, M., Burton, N. W., Kanyi, K., & Jackenthal, A. (2002). *Selecting successful graduate students: In-depth interviews with GRE users* (Research Report No. RR-02-08). ETS. https://doi.org/10.1002/j.2333-8504.2002.tb01875.x

Walters, A. M., Lee, S., & Trapani, C. (2004). *Stereotype threat*, *the test-center environment*, *and performance on the GRE General Test* (Research Report No. RR-04-37; College Board Report No 01-03R). ETS and College Board. https://doi.org/10.1002/j.2333-8504.2004.tb01964.x

Wigfield, A., Eccles, J. S., Roeser, R. W., & Schiefele, U. (2008). Development of achievement motivation. In W. Damon & R. M. Lerner (Eds.). *Child and adolescent development: An advanced course*. John Wiley & Sons Inc. https://doi.org/10.1002/9780470147658.chpsy0315

World Bank. (2014). *GNI per capita, Atlas method (current US$)*. http://data.worldbank.org/indicator/NY.GNP.PCAP.CD

Appendix A

Point Estimates, Sample Sizes, and Focal Analyses

Table A1 Effect Size Differences Between Non-SS and SS Users by Gender and Age Groups for GRE Verbal, Quantitative, and Analytical Writing

				Sample size			
		Effect size (Cohen's *d*)		No SS		SS	
GRE section	SRT	No SS	SS	Male	Female	Male	Female
V	ASR	−.36	−.35	181,390	283,892	181,345	283,849
	Free	−.32	−.32	234,746	396,300	234,746	396,300
Q	ASR	−.55	−.54	181,441	283,926	181,397	283,884
	Free	−.51	−.50	234,746	396,300	234,746	396,300
AW	ASR	−.08	−.08	181,127	283,619	181,048	283,530
	Free	−.06	−.06	234,085	395,433	234,006	395,247
				Age			
				15–24	25–39	15–24	25–39
V[a]	ASR	.16	.16	365,830	118,114	365,805	118,063
	Free	.13	.12	442,290	199,547	442,290	199,547
Q	ASR	−.20	−.20	365,855	118,168	365,830	118,119
	Free	−.22	−.22	442,290	199,547	442,290	199,547
AW	ASR	−.04	−.04	365,550	117,867	365,469	117,791
	Free	−.10	−.10	441,562	198,751	441,412	198,650
				Age			
				15–24	40+	15–24	40+
V[a]	ASR	−.08	−.09	365,830	7,808	365,805	7,796
	Free	.00	−.01	442,290	27,920	442,290	27,920
Q	ASR	−.91	−.91	365,855	7,820	365,830	7,808
	Free	−.77	−.77	442,290	27,920	442,290	27,920
AW	ASR	−.51	−.50	365,550	7,774	365,469	7,762
	Free	−.46	−.46	441,562	27,805	441,412	27,778

Note. SRT = Score report type; SS = ScoreSelect; ASR = additional score report; Free = free score report; V = verbal reasoning; Q = quantitative reasoning; AW = analytical writing. Except where indicated with superscript a, the initially higher scoring group appears first under sample size. A Cohen's *d* less than 0 indicates a higher mean score for the initially higher scoring group. In general, the lower and upper bounds for 95% confidence intervals around the *d* values would be, after rounding, the *d* values ± .01. Samples compared are solely U.S. citizens unless otherwise indicated with superscript b. Sample size represents number of score reports. Sample sizes sometimes are slightly smaller for the ScoreSelect condition, because some examinees who could have sent score reports (and thus whose reports were counted toward the hypothetical No ScoreSelect condition) decided not to send any score reports (and thus did not have any reports counted toward the real-world ScoreSelect condition).

Table A2 Effect Size Differences Between Non-SS and SS Users by Race/Ethnic Groups for GRE Verbal, Quantitative, and Analytical Writing

				Sample size			
		Effect size (Cohen's *d*)		No SS		SS	
GRE section	SRT	No SS	SS	White	Comparison	White	Comparison
White/Black or African American							
V	ASR	−.75	−.75	353,349	22,729	353,288	22,727
	Free	−.86	−.84	460,006	52,498	460,006	52,498
Q	ASR	−.81	−.81	353,378	22,737	353,319	22,735
	Free	−.86	−.85	460,006	52,498	460,006	52,498
AW	ASR	−.73	−.73	352,953	22,693	352,842	22,687
	Free	−.82	−.82	458,970	52,350	458,792	52,318

Table A2 Continued

				Sample size			
		Effect size (Cohen's *d*)		No SS		SS	
GRE section	SRT	No SS	SS	White	Comparison	White	Comparison
White/Latino							
V	ASR	−.36	−.36	353,349	29,388	353,288	29,375
	Free	−.45	−.44	460,006	48,469	460,006	48,469
Q	ASR	−.36	−.36	353,378	29,426	353,319	29,413
	Free	−.41	−.41	460,006	48,469	460,006	48,469
AW	ASR	−.38	−.38	352,953	29,364	352,842	29,346
	Free	−.44	−.44	458,970	48,317	458,792	48,294
White/Mexican, Mexican American, or Chicano							
V	ASR	−.36	−.36	353,349	10,018	353,288	10,017
	Free	−.48	−.47	460,006	18,229	460,006	18,229
Q	ASR	−.35	−.35	353,378	10,019	353,319	10,018
	Free	−.43	−.43	460,006	18,229	460,006	18,229
AW	ASR	−.34	−.34	352,953	10,012	352,842	10,011
	Free	−.42	−.42	458,970	18,184	458,792	18,176
White/Puerto Rican							
V	ASR	−53	−.53	353,349	3,900	353,288	3,900
	Free	−.56	−.55	460,006	6,017	460,006	6,017
Q	ASR	−.49	−.49	353,378	3,905	353,319	3,905
	Free	−.50	−.49	460,006	6,017	460,006	6,017
AW	ASR	−.65	−.65	352,953	3,898	352,842	3,897
	Free	−.69	−.70	458,970	6,001	458,792	5,996
White/Other Hispanic or Latin American							
V	ASR	−.32	−.32	353,349	15,470	353,288	15,458
	Free	−.40	−.39	460,006	24,223	460,006	24,223
Q	ASR	−.34	−.34	353,378	15,502	353,319	15,490
	Free	−.38	−.37	460,006	24,223	460,006	24,223
AW	ASR	−.34	−.34	352,953	15,454	352,842	15,438
	Free	−.39	−.39	458,970	24,132	458,792	24,122
White/Asian or Asian American							
V	ASR	−.09	−.08	353,349	43,112	353,288	43,112
	Free	−.15	−.14	460,006	44,223	460,006	44,223
Q[a]	ASR	.48	.48	353,378	43,114	353,319	43,114
	Free	.36	.36	460,006	44,223	460,006	44,223
AW	ASR	−.08	−.08	352,953	43,091	352,842	43,086
	Free	−.07	−.07	458,970	44,149	458,792	44,135
White/American Indian or Alaskan Native							
V	ASR	−.26	−.26	353,349	2,029	353,288	2,029
	Free	−.38	−.37	460,006	3,657	460,006	3,657
Q	ASR	−.23	−.24	353,378	2,029	353,319	2,029
	Free	−.36	−.36	460,006	3,657	460,006	3,657
AW	ASR	−.34	−.34	352,953	2,028	352,842	2,028
	Free	−.43	−.43	458,970	3,655	458,792	3,650
White/Hawaiian or Pacific Islander							
V	ASR	−.36	−.35	353,349	1,389	353,288	1,389
	Free	−.43	−.42	460,006	2,169	460,006	2,169
Q	ASR	−.13	−.13	353,378	1,389	353,319	1,389
	Free	−.25	−.25	460,006	2,169	460,006	2,169
AW	ASR	−.15	−.14	352,953	1,389	352,842	1,389
	Free	−.26	−.26	458,970	2,169	458,792	2,169

Table A2 Continued

		Effect size (Cohen's *d*)		Sample size			
				No SS		SS	
GRE section	SRT	No SS	SS	White	Comparison	White	Comparison
White/Other							
V	ASR	.09	.09	353,349	23,483	353,288	23,471
	Free	.06	.06	460,006	27,815	460,006	27,815
Q	ASR	.03	.03	353,378	23,490	353,319	23,478
	Free	−.04	−.04	460,006	27,815	460,006	27,815
AW	ASR	.02	.02	352,953	23,441	352,842	23,427
	Free	−.02	−.02	458,970	27,733	458,792	27,729

Note. SRT = Score report type; SS = ScoreSelect; ASR = additional score report; Free = free score report; V = verbal reasoning; Q = quantitative reasoning; AW = analytical writing. Except where indicated with superscript a, the initially higher scoring group appears first under sample size. A Cohen's *d* less than 0 indicates a higher mean score for the initially higher scoring group. In general, the lower and upper bounds for 95% confidence intervals around the *d* values would be, after rounding, the *d* values ± .01. Samples compared are solely U.S. citizens unless otherwise indicated with superscript b. Sample size represents number of score reports. Sample sizes sometimes are slightly smaller for the ScoreSelect condition, because some examinees who could have sent score reports (and thus whose reports were counted toward the hypothetical No ScoreSelect condition) decided not to send any score reports (and thus did not have any reports counted toward the real-world ScoreSelect condition).

Table A3 Effect Size Differences Between Non-SS and SS Users by Parental Educational Attainment for GRE Verbal, Quantitative, and Analytical Writing

		Effect size (Cohen's *d*)		Sample size			
				No SS		SS	
GRE section	SRT	No SS	SS	Some graduate, graduate, or professional	Comparison	Some graduate, graduate, or professional	Comparison
Some graduate, graduate, or professional/some high school, grade school, or less							
V	ASR	−.88	−.87	206,081	7,803	206,032	7,790
	Free	−.92	−.90	231,125	15,548	231,125	15,548
Q	ASR	−.65	−.65	206,126	7,813	206,077	7,800
	Free	−.73	−.72	231,125	15,548	231,125	15,548
AW	ASR	−.73	−.73	205,918	7,787	205,838	7,769
	Free	−.76	−.76	230,736	15,514	230,645	15,504
Some graduate, graduate, or professional/high school equivalent, business or trade school, some college							
V	ASR	−.59	−.58	206,081	72,141	206,032	72,132
	Free	−.58	−.57	231,125	135,577	231,125	135,577
Q	ASR	−.53	−.53	206,126	72,161	206,077	72,152
	Free	−.52	−.52	231,125	135,577	231,125	135,577
AW	ASR	−.44	−.43	205,918	72,161	205,838	72,008
	Free	−.45	−.45	230,736	135,577	230,645	135,104
Some graduate, graduate, or professional/associate or bachelor's							
V	ASR	−.39	−.39	206,081	131,183	206,032	131,167
	Free	−.36	−.35	231,125	192,393	231,125	192,393
Q	ASR	−.30	−.30	206,126	131,186	206,077	131,172
	Free	−.26	−.25	231,125	192,393	231,125	192,393
AW	ASR	−.27	−.27	205,918	131,055	205,838	131,026
	Free	−.25	−.25	230,736	191,965	230,645	191,895

Note. SRT = Score report type; SS = ScoreSelect; ASR = additional score report; Free = free score report; V = verbal reasoning; Q = quantitative reasoning; AW = analytical writing. Except where indicated with superscript a, the initially higher scoring group appears first under sample size. A Cohen's *d* less than 0 indicates a higher mean score for the initially higher scoring group. In general, the lower and upper bounds for 95% confidence intervals around the *d* values would be, after rounding, the *d* values ± .01. Samples compared are solely U.S. citizens unless otherwise indicated with superscript b. Sample size represents number of score reports. Sample sizes sometimes are slightly smaller for the ScoreSelect condition, because some examinees who could have sent score reports (and thus whose reports were counted toward the hypothetical No ScoreSelect condition) decided not to send any score reports (and thus did not have any reports counted toward the real-world ScoreSelect condition).

Table A4 Effect Size Differences Between non-SS and SS Users by Citizenship Subgroups for GRE Verbal, Quantitative, and Analytical Writing

				Sample size			
		Effect size (Cohen's *d*)		No SS		SS	
GRE section	SRT	No SS	SS	U.S. citizen	Comparison	U.S. citizen	Comparison
[b]Non U.S. citizens							
V	ASR	−.81	−.74	489,842	486,190	489,754	485,953
	Free	−.74	−.71	669,440	369,437	669,440	369,437
Q[a]	ASR	1.01	1.02	489,933	486,214	489,847	485,965
	Free	.69	.69	669,440	369,437	669,440	369,437
AW	ASR	−1.30	−1.28	489,281	485,530	489,112	485,176
	Free	−1.08	−1.07	667,804	367,671	667,526	367,503
[b]Chinese citizens							
V	ASR	−.83	−.70	489,842	187,031	489,754	186,823
	Free	−.60	−.50	669,440	39,160	669,440	39,160
Q[a]	ASR	1.55	1.57	489,933	187,031	489,847	186,823
	Free	1.60	1.61	669,440	39,160	669,440	39,160
AW	ASR	−1.52	−1.48	489,281	187,024	489,112	186,807
	Free	−1.18	−1.16	667,804	39,148	667,526	39,146
[b]Non-U.S., non-Chinese citizens							
V	ASR	−.79	−.76	489,842	299,159	489,754	299,130
	Free	−.76	−.74	669,440	330,277	669,440	330,277
Q[a]	ASR	.69	.71	489,933	299,183	489,847	299,142
	Free	.58	.59	669,440	330,277	669,440	330,277
AW	ASR	−1.13	−1.12	489,281	298,506	489,112	298,369
	Free	−1.05	−1.05	667,804	328,523	667,526	328,357

Note. SRT = score report type; SS = ScoreSelect; ASR = additional score report; Free = free score report; V = Verbal Reasoning; Q = Quantitative Reasoning; AW = Analytical Writing. Except where indicated with superscript a, the initially higher scoring group appears first under sample size. A Cohen's *d* less than 0 indicates a higher mean score for the initially higher scoring group. In general, the lower and upper bounds for 95% confidence intervals around the *d* values would be, after rounding, the *d* values ± .01. Samples compared are solely U.S. citizens unless otherwise indicated with superscript b. Sample size represents number of score reports. Sample sizes sometimes are slightly smaller for the ScoreSelect condition, because some examinees who could have sent score reports (and thus whose reports were counted toward the hypothetical No ScoreSelect condition) decided not to send any score reports (and thus did not have any reports counted toward the real-world ScoreSelect condition).

Table A5 How Many Times More Often the Initially Higher Scoring Group Would Be Admitted Compared to the Initially Lower Scoring Group Based on Graduate Program Selectivity

			Highly selective			Selective			Nonselective		
	GRE section	SRT	No SS	SS	Δ	No SS	SS	Δ	No SS	SS	Δ
Male/Female											
	V	ASR	1.98	1.95	−1%	1.39	1.38	−1%	1.09	1.09	0%
		Free	1.85	1.81	−2%	1.34	1.33	−1%	1.08	1.08	0%
	Q	ASR	2.99	2.94	−2%	1.72	1.71	−1%	1.17	1.17	0%
		Free	2.74	2.68	−2%	1.64	1.62	−1%	1.16	1.15	0%
	AW	ASR	1.16	1.15	−1%	1.07	1.07	0%	1.02	1.02	0%
		Free	1.12	1.12	0%	1.05	1.05	0%	1.01	1.01	0%
15–24/25–39											
	V[a]	ASR	1.35	1.33	−1%	1.15	1.14	−1%	1.04	1.03	0%
		Free	1.26	1.25	−1%	1.11	1.11	−1%	1.03	1.03	0%
	Q	ASR	1.45	1.45	0%	1.19	1.19	0%	1.05	1.05	0%
		Free	1.50	1.50	0%	1.21	1.21	0%	1.05	1.05	0%
	AW	ASR	1.07	1.08	1%	1.03	1.03	0%	1.01	1.01	0%
		Free	1.20	1.20	0%	1.09	1.09	0%	1.02	1.02	0%

Table A5 Continued

			Highly selective			Selective			Nonselective		
	GRE section	SRT	No SS	SS	Δ	No SS	SS	Δ	No SS	SS	Δ
15–24/40+											
	V[a]	ASR	0.86	0.85	−1%	0.93	0.93	−1%	0.98	0.98	0%
		Free	1.00	0.98	−2%	1.00	0.99	−1%	1.00	1.00	0%
	Q	ASR	7.00	7.09	1%	2.75	2.77	1%	1.39	1.40	0%
		Free	5.00	4.93	−1%	2.27	2.25	−1%	1.30	1.29	0%
	AW	ASR	2.71	2.69	−1%	1.63	1.63	0%	1.15	1.15	0%
		Free	2.47	2.46	0%	1.55	1.55	0%	1.13	1.13	0%
White/Black or African American											
	V	ASR	4.73	4.72	0%	2.20	2.20	0%	1.28	1.28	0%
		Free	6.16	5.95	−3%	2.55	2.50	−2%	1.35	1.34	−1%
	Q	ASR	5.43	5.42	0%	2.38	2.38	0%	1.32	1.32	0%
		Free	6.17	5.98	−3%	2.56	2.51	−2%	1.36	1.35	−1%
	AW	ASR	4.46	4.48	0%	2.13	2.14	0%	1.27	1.27	0%
		Free	5.56	5.55	0%	2.41	2.41	0%	1.33	1.32	0%
White/Latino											
	V	ASR	1.99	1.99	0%	1.39	1.39	0%	1.10	1.10	0%
		Free	2.39	2.36	−1%	1.53	1.52	−1%	1.13	1.13	0%
	Q	ASR	1.98	1.97	−1%	1.39	1.39	0%	1.10	1.09	0%
		Free	2.22	2.19	−1%	1.47	1.46	−1%	1.12	1.11	0%
	AW	ASR	2.07	2.07	0%	1.42	1.42	0%	1.10	1.10	0%
		Free	2.33	2.33	0%	1.51	1.51	0%	1.12	1.12	0%
White/Mexican, Mexican American, or Chicano											
	V	ASR	1.98	1.98	0%	1.39	1.39	0%	1.09	1.09	0%
		Free	2.55	2.51	−1%	1.58	1.57	−1%	1.14	1.14	0%
	Q	ASR	1.96	1.93	−1%	1.38	1.37	−1%	1.09	1.09	0%
		Free	2.31	2.28	−1%	1.50	1.49	−1%	1.12	1.12	0%
	AW	ASR	1.90	1.91	0%	1.36	1.36	0%	1.09	1.09	0%
		Free	2.26	2.26	0%	1.48	1.49	0%	1.12	1.12	0%
White/Puerto Rican											
	V	ASR	2.88	2.88	0%	1.69	1.69	0%	1.16	1.16	0%
		Free	3.02	2.98	−2%	1.73	1.72	−1%	1.18	1.17	0%
	Q	ASR	2.61	2.59	−1%	1.60	1.59	0%	1.15	1.14	0%
		Free	2.67	2.62	−2%	1.62	1.60	−1%	1.15	1.15	0%
	AW	ASR	3.77	3.75	0%	1.95	1.94	0%	1.22	1.22	0%
		Free	4.15	4.18	1%	2.05	2.06	0%	1.25	1.25	0%
White/Other Hispanic or Latin American											
	V	ASR	1.83	1.83	0%	1.34	1.34	0%	1.08	1.08	0%
		Free	2.15	2.12	−1%	1.45	1.44	−1%	1.11	1.11	0%
	Q	ASR	1.90	1.90	0%	1.36	1.36	0%	1.09	1.09	0%
		Free	2.07	2.04	−1%	1.42	1.41	−1%	1.10	1.10	0%
	AW	ASR	1.90	1.90	0%	1.36	1.36	0%	1.09	1.09	0%
		Free	2.10	2.10	0%	1.43	1.43	0%	1.10	1.10	0%
White/Asian or Asian American											
	V	ASR	1.16	1.16	−1%	1.07	1.07	0%	1.02	1.02	0%
		Free	1.30	1.29	−1%	1.13	1.12	−1%	1.03	1.03	0%
	Q[a]	ASR	2.55	2.54	0%	1.58	1.58	0%	1.14	1.14	0%
		Free	2.00	1.97	−1%	1.40	1.39	−1%	1.10	1.09	0%
	AW	ASR	1.15	1.15	0%	1.07	1.07	0%	1.02	1.02	0%
		Free	1.13	1.13	0%	1.06	1.06	0%	1.01	1.01	0%
White/American Indian or Alaskan Native											
	V	ASR	1.63	1.63	0%	1.26	1.26	0%	1.06	1.06	0%
		Free	2.05	2.02	−1%	1.41	1.40	−1%	1.10	1.10	0%
	Q	ASR	1.54	1.56	1%	1.23	1.23	0%	1.06	1.06	0%
		Free	1.99	1.98	−1%	1.39	1.39	0%	1.10	1.09	0%
	AW	ASR	1.89	1.90	1%	1.36	1.36	0%	1.09	1.09	0%
		Free	2.28	2.28	0%	1.49	1.49	0%	1.12	1.12	0%

Table A5 Continued

			Highly selective			Selective			Nonselective		
	GRE section	SRT	No SS	SS	Δ	No SS	SS	Δ	No SS	SS	Δ
White/Hawaiian or Pacific Islander											
	V	ASR	1.99	1.96	−2%	1.39	1.38	−1%	1.10	1.09	0%
		Free	2.28	2.26	−1%	1.49	1.48	−1%	1.12	1.12	0%
	Q	ASR	1.27	1.27	0%	1.12	1.12	0%	1.03	1.03	0%
		Free	1.58	1.58	0%	1.24	1.24	0%	1.06	1.06	0%
	AW	ASR	1.31	1.28	−2%	1.13	1.12	−1%	1.03	1.03	0%
		Free	1.62	1.63	0%	1.26	1.26	0%	1.06	1.06	0%
White/Other											
	V	ASR	0.85	0.86	1%	0.93	0.93	0%	0.98	0.98	0%
		Free	0.90	0.91	0%	0.95	0.96	0%	0.99	0.99	0%
	Q	ASR	0.95	0.95	0%	0.98	0.98	0%	0.99	0.99	0%
		Free	1.08	1.08	0%	1.04	1.04	0%	1.01	1.01	0%
	AW	ASR	0.96	0.96	0%	0.98	0.98	0%	1.00	1.00	0%
		Free	1.03	1.03	0%	1.01	1.01	0%	1.00	1.00	0%
Some graduate, graduate, or professional/some high school, grade school, or less											
	V	ASR	6.51	6.43	−1%	2.64	2.62	−1%	1.37	1.37	0%
		Free	7.24	6.93	−4%	2.80	2.73	−2%	1.40	1.39	−1%
	Q	ASR	3.77	3.78	0%	1.95	1.95	0%	1.22	1.22	0%
		Free	4.48	4.39	−2%	2.14	2.12	−1%	1.27	1.26	0%
	AW	ASR	4.56	4.55	0%	2.16	2.16	0%	1.27	1.27	0%
		Free	4.90	4.90	0%	2.25	2.25	0%	1.29	1.29	0%
Some graduate, graduate professional/high school equivalent, business or trade school, some college											
	V	ASR	3.23	3.22	0%	1.79	1.79	0%	1.19	1.19	0%
		Free	3.16	3.10	−2%	1.77	1.75	−1%	1.18	1.18	0%
	Q	ASR	2.87	2.87	0%	1.68	1.68	0%	1.16	1.16	0%
		Free	2.79	2.76	−1%	1.66	1.65	0%	1.16	1.16	0%
	AW	ASR	2.33	2.32	−1%	1.51	1.51	0%	1.12	1.12	0%
		Free	2.38	2.39	0%	1.53	1.53	0%	1.13	1.13	0%
Some graduate, graduate, or professional/associate or bachelor's											
	V	ASR	2.13	2.12	0%	1.44	1.44	0%	1.11	1.11	0%
		Free	1.98	1.95	−1%	1.39	1.38	−1%	1.10	1.09	0%
	Q	ASR	1.77	1.77	0%	1.31	1.31	0%	1.08	1.08	0%
		Free	1.61	1.60	−1%	1.25	1.25	0%	1.06	1.06	0%
	AW	ASR	1.65	1.64	−1%	1.27	1.26	0%	1.07	1.07	0%
		Free	1.58	1.58	0%	1.24	1.24	0%	1.06	1.06	0%
^{b}U.S. citizens/Non-U.S. citizens											
	V	ASR	5.48	4.63	−16%^{c}	2.39	2.18	−9%^{c}	1.32	1.27	−3%
		Free	4.63	4.31	−7%^{c}	2.18	2.09	−4%	1.27	1.26	−1%
	Q^{a}	ASR	9.12	9.36	3%	3.20	3.25	2%	1.48	1.49	1%
		Free	4.11	4.11	0%	2.04	2.04	0%	1.24	1.24	0%
	AW	ASR	20.30	19.19	−6%^{c}	5.17	5.00	−3%	1.83	1.80	−2%
		Free	10.99	10.70	−3%	3.57	3.50	−2%	1.55	1.54	−1%
^{b}U.S. citizens/Chinese citizens											
	V	ASR	5.76	4.21	−27%^{c}	2.46	2.07	−16%^{c}	1.33	1.25	−6%^{c}
		Free	3.34	2.67	−20%^{c}	1.82	1.62	−11%^{c}	1.20	1.15	−4%
	Q^{a}	ASR	43.18	45.97	6%^{c}	8.25	8.59	4%	2.28	2.33	2%
		Free	50.54	52.17	3%	9.12	9.31	2%	2.40	2.42	1%
	AW	ASR	39.33	34.77	−12%^{c}	7.80	7.20	−7%^{c}	2.22	2.14	−4%
		Free	14.46	13.68	−5%^{c}	4.20	4.06	−3%	1.67	1.64	−1%
^{b}U.S. citizens/Non-U.S., non-Chinese citizens											
	V	ASR	5.22	4.85	−7%^{c}	2.33	2.24	−4%	1.31	1.29	−2%
		Free	4.85	4.63	−5%^{c}	2.24	2.18	−3%	1.29	1.27	−1%
	Q^{a}	ASR	4.11	4.31	5%^{c}	2.04	2.09	3%	1.24	1.26	1%
		Free	3.19	3.26	2%	1.78	1.80	1%	1.19	1.19	0%
	AW	ASR	12.59	12.25	−3%	3.87	3.81	−2%	1.61	1.60	−1%
		Free	10.14	10.14	0%	3.40	3.40	0%	1.52	1.52	0%

Note. ASR = additional score report; Free = free score report; SRT = score report type; SS = ScoreSelect; Δ = % change from No SS to SS (based on values of No SS and SS taken to more digits than what is reportable in the table); V = Verbal Reasoning; Q = Quantitative Reasoning; AW = Analytical Writing; Highly selective = top 10% admitted; Selective = top 50% admitted; Nonselective = top 90% admitted. Except where indicated with superscript a, the initially higher scoring group appears first. Samples compared are solely U.S. citizens unless otherwise indicated with superscript b. Ratios are presented as initially higher scoring group compared to initially lower scoring group. Superscript c indicates practical difference observed. Sample sizes are identical to those for the same subgroup comparisons in Tables A1 – A4 and have been omitted here to conserve space.

Appendix B

ScoreSelect Use for Additional Score Reports and Free Reports

Table B1 ScoreSelect Use: Additional Score Reports

	U.S. citizens		Chinese citizens		Non-U.S., non-Chinese	
Available – Sent	%	*N*	%	*N*	%	*N*
1 – 0	0	77	0	271	0	45
1 – 1	84	360,029	57	97,665	79	216,747
2 – 0	0	0	0	20	0	0
2 – 1	9	38,700	30	51,529	15	41,108
2 – 2	6	26,774	4	6,094	3	8,166
3 – 1	1	2,679	7	11,671	2	5,971
3 – 2	0	847	1	1,249	0	1,298
3 – 3	0	997	0	462	0	317
4 – 1	0	133	1	1,554	0	820
4 – 2	0	63	0	206	0	256
4 – 3	0	23	0	47	0	30
4 – 4	0	21	0	29	0	20
5 – 1	0	12	0	166	0	107
5 – 2	0	0	0	8	0	83
5 – 3	0	0	0	0	0	8
5 – 4	0	1	0	0	0	0
5 – 5	0	1	0	0	0	4
6 – 1	0	0	0	0	0	0
6 – 6	0	0	0	0	0	0

Note. *Available* (the value to the left of the dash) indicates the number of sets of scores that could have been reported on a score report. *Sent* (the value to the right of the dash) indicates the number of sets of scores that were reported on a score report given a certain number of available score sets. % = percentage of the citizen population that made the reporting decision described by Available – Sent. *N* = number of score reports. Note that a ScoreSelect decision is directly made when Available (the value to the left of the dash) equals 2 or greater. The data in this table are illustrated in Figure 1.

Table B2 ScoreSelect Use: Free Reports

	U.S. citizens		Chinese citizens		Non-U.S., non-Chinese	
Available – Sent	%	*N*	%	*N*	%	*N*
1 – 0	0	0	0	0	0	0
1 – 1	89	594,237	64	25,164	86	282,814
2 – 0	0	0	0	0	0	0
2 – 1	5	35,643	25	9,808	10	31,816
2 – 2	5	35,296	4	1,650	3	10,300
3 – 1	0	2,199	5	1,997	1	3,567
3 – 2	0	0	0	0	0	0
3 – 3	0	1,947	0	124	0	945
4 – 1	0	215	1	356	0	532
4 – 2	0	0	0	0	0	0
4 – 3	0	0	0	0	0	0
4 – 4	0	150	0	15	0	144
5 – 1	0	18	0	46	0	96
5 – 2	0	0	0	0	0	0
5 – 3	0	0	0	0	0	0
5 – 4	0	0	0	0	0	0
5 – 5	0	27	0	0	0	25
6 – 1	0	1	0	0	0	1
6 – 6	0	5	0	0	0	0

Note. *Available* (the value to the left of the dash) indicates the number of sets of scores that could have been reported on a score report. *Sent* (the value to the right of the dash) indicates the number of sets of scores that were reported on a score report given a certain number of available score sets. % = percentage of the citizen population that made the reporting decision described by Available – Sent. *N* = number of score reports. Note that a ScoreSelect decision is directly made when Available (the value to the left of the dash) equals 2 or greater. The data in this table are illustrated in Figure 2.

Appendix C

Survey

Dear GRE Test Taker,

You are being asked to participate in a research study that is being conducted by Dr. David Klieger, who is a researcher at Educational Testing Service. The purpose of this research is to learn more about how GRE test-taking and score reporting behaviors are related to students' educational and career goals; your perspectives will support the GRE Program's ongoing commitment to fairness and effectively providing information about the score reporting options available to test takers.

Study participants will be asked to take a short online survey that takes approximately 15 minutes to complete. The survey includes questions about why you took the GRE General Test, whether you chose to take the test more than once, your score reporting decisions, and your goals for the future. Once you have completed this survey, we will send you an online gift card worth $15 as a token of our appreciation.

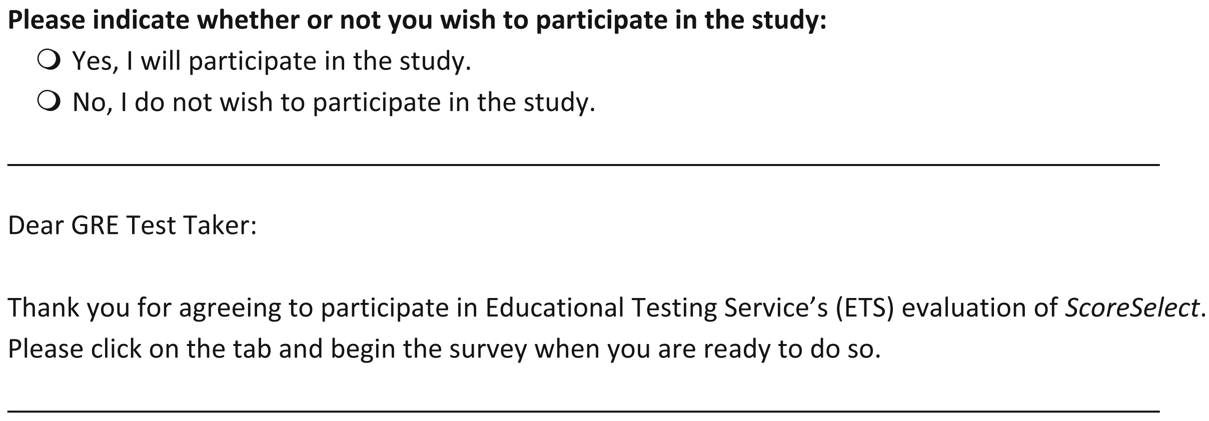
Participation in this study is voluntary. You may choose not to participate, and you may withdraw at any time during the study procedures without any penalty to you. In addition, you may choose not to answer any questions with which you are uncomfortable.

This research is confidential. ETS will keep your survey data confidential by limiting individuals' access to the information and by keeping the information in a secure location. The research team at ETS is the only party that will be allowed to see the data, except as may be required by law.

All data will be kept under lock and key for a period of 5 years. After that time, these data will be erased from all computer files and any hard copies of this information will be shredded. There are no foreseeable risks to participation in this study.

If you have any questions about the study procedures, you may contact Megan Schramm-Possinger.

Megan Schramm-Possinger, Ph.D.,
Educational Testing Service,
Turnbull Hall,
Rosedale Road,
Princeton, NJ 08541, USA.



Please indicate whether or not you wish to participate in the study:

❍ Yes, I will participate in the study.

❍ No, I do not wish to participate in the study.

Dear GRE Test Taker:

Thank you for agreeing to participate in Educational Testing Service's (ETS) evaluation of *ScoreSelect*. Please click on the tab and begin the survey when you are ready to do so.

BACKGROUND

Last Name: ____

First Name: ____

Middle Name: (optional) ____

Date of birth: _______________________

1. What are your most recent GRE General Test scores?

Verbal Reasoning _______________________

Quantitative Reasoning _______________________

Analytical Writing _______________________

2. What is the highest degree you have earned?

○ Bachelors (B.A. or B.S.)

○ Master's (M.A. or M.S.)

○ M.B.A.

○ Doctorate (Ph.D., Ed.D.)

○ Professional (J.D., M.D.)

○ Other (please specify) ____________________________

3. What is your current educational status?

○ I am currently enrolled in an undergraduate degree program.

○ I am currently enrolled in a graduate degree program; I began the program in the year 20 ___.

○ I am not currently enrolled in an undergraduate, graduate, or professional degree program.

4. What are your plans to attend graduate school?

○ I plan to enroll in a graduate degree program starting in the year 20 ______.

○ I plan to attend a graduate degree program, but I'm not sure when I will enroll.

○ I am not sure if I will attend a graduate degree program.

○ I do not plan to enroll in a graduate degree program.

5. Please further describe your educational status:

○ I plan to enroll in a graduate degree program starting in the year 20 ___.

○ I plan to attend a graduate degree program, but I'm not sure when I will enroll.

○ I am not sure if I will attend a graduate degree program.

○ I was enrolled in a graduate degree program and completed my degree in the year 20 _____.

○ I was enrolled in a graduate degree program but did not complete a degree.

○ I have not enrolled in a graduate or professional degree program, and I do not plan to pursue a graduate or professional degree.

Destination: **Not going to grad school** (Set in 5 (I have not enrolled in a graduate or professional degree program, and I do not plan to pursue a graduate or professional degree.))

6. Which of the following best describes the degree program you want to attend, are currently enrolled in, or were enrolled in?

○ Nondegree graduate study
○ Master's (M.A. or M.S.)
○ M.B.A.
○ Post-Baccalaureate or post-Master's Degree Certificate Program
○ Doctorate (Ph.D., Ed.D.)
○ Professional (JD, MD)
○ Postdoctoral study

7. How important is it to you to attend graduate or professional school?

○ Very important
○ Important
○ Not very important
○ Not important at all

8. Please indicate your level of agreement with the following statements:

	Strongly agree	Agree	Disagree	Strongly disagree
A graduate or professional degree provides me with better career opportunities.	○	○	○	○
The benefit of a graduate or professional education outweighs the cost.	○	○	○	○
A graduate or professional degree will increase my income potential.	○	○	○	○
In my career, attending graduate or professional school is more important than work experience.	○	○	○	○
Attending graduate school is an important personal goal of mine.	○	○	○	○

9. How important do you think your scores on the GRE General Test are or were for gaining admittance to your graduate program?

❍ Very important
❍ Important
❍ Not very important
❍ Not important at all

10. When you applied to your graduate degree program(s), was information about the GRE General Test scores of students who were accepted available to you? (e.g., 50% of students in the Doctoral Program in Political Science scored above 550 in Verbal Reasoning, etc.)

❍ Yes, for each graduate degree program I applied to
❍ Yes, for some, but not all of the graduate degree programs I applied to
❍ No, not for any of the graduate degree programs I applied to
❍ I'm not sure

11. Did you seek financial assistance such as a TA, GA, or fellowship to pay for your graduate education?

❍ Yes
❍ No

12. How important do you think your scores on the GRE are or were for receiving financial support for your graduate education?

❍ Very important
❍ Important
❍ Not very important
❍ Not important at all

13. About how much time did you spend preparing for your most recent GRE test administration?

Number of hours: ______

14. How many times have you taken the GRE General Test in the past 5 years?

Number of times taken: ______

15. How satisfied are you or were you with your highest test scores in the Verbal, Quantitative, and Analytical Writing sections of the GRE General Test?

	Very satisfied	Somewhat satisfied	Somewhat dissatisfied	Very dissatisfied
Verbal Reasoning	○	○	○	○
Quantitative Reasoning	○	○	○	○
Analytical Writing	○	○	○	○

16. Do you plan to take the GRE General Test again?

○ Yes
○ No

17. Please check the statement(s) that best describe your reasons for taking the GRE General Test again. (Check all that apply)

□ I believe I can improve my scores.
□ I need higher scores to get into my graduate program.
□ If my scores go down, I don't have to send them, so I have nothing to lose.
□ Testing makes me anxious, but improving my scores would be worth it.
□ For other reasons.

18. Please check the statement(s) that best describe your reasons for not taking the GRE General Test again. (Check all that apply.)

□ I have already been accepted to, have attended, or am attending graduate school and do not need to take the GRE General Test again.
□ My scores are high enough for admittance into the graduate program of my choice.
□ Taking the GRE General Test again is not likely to improve my scores.
□ Taking the GRE General Test again costs too much money.
□ I don't have time.
□ Testing makes me anxious and improving my score is not worth it.
□ I plan to apply to a program that does not require GRE General Test scores for admittance.
□ For other reasons.

19. To what extent do you agree or disagree with the following statements? If you are currently matriculated in a graduate program, then please refer to the time prior to your admittance when answering this question.

	Strongly agree	Agree	Disagree	Strongly disagree
If I study for the Verbal Reasoning section of the GRE General Test, then I can raise my scores.	○	○	○	○
I am just not a good test taker.	○	○	○	○

If I study for the Quantitative Reasoning section of the GRE General Test, then I can raise my scores.	○	○	○	○
My score on the Analytical Writing section of the GRE General Test will not improve no matter how much I study.	○	○	○	○
My scores are likely to improve now that I know what the GRE General Test is like.	○	○	○	○
I cannot afford to take the GRE General Test again – it is too expensive.	○	○	○	○

After July 2012, ETS began offering ScoreSelect for the GRE General and Subject Tests. ScoreSelect allows test takers to make decisions about sending out their GRE test scores. The questions in this survey are specifically about only ScoreSelect for the GRE General Test, which assesses verbal reasoning, quantitative reasoning, and analytical writing skills.

20. How familiar were you with GRE ScoreSelect when you most recently took the GRE?

○ Very familiar
○ Somewhat familiar
○ Not very familiar
○ Not at all familiar

21. Where did you learn about the GRE ScoreSelect? (Check all that apply)

□ From my parents, siblings, or other family member
□ From a friend or fellow student
□ From a professor
□ From the ETS website
□ From the GRE test center, on the day of the test
□ From a test preparation program, booklet, or guide
□ From a career center or adviser at school
□ From a GRE registration booklet
□ Other: ___________________________

22. Which of the following ScoreSelect options were you aware of when you most recently took the GRE General Test? (Check all that apply.)

□ With ScoreSelect, you can choose to send none, some, or all of your sets of GRE General Test scores.

□ With ScoreSelect you have the option of sending only your highest set of GRE General Test scores, and no others.

□ At the test center on test day you can choose to send GRE General Test score reports to up to 4 schools for free.

□ After test day, the fee for sending your GRE General Test score reports is $25 per school.

□ The option to pick and choose among <u>any</u> of your sets of GRE General Test scores is only available after you leave the test center.

The next set of questions ask about the last time you took the GRE General Test.

23. At the testing center on the day of the test, did you choose to send any GRE General Test scores out?

○ Yes
○ No
○ I don't recall

---

24. Please select the statement(s) that best describe your reason(s) for sending your GRE General Test scores out at the testing center on the day of the test. (Select all that apply.)

□ I wanted to take advantage of the free-of-charge reports.
□ My scores were high enough to get me into my graduate program.
□ I didn’t know I could opt not to send scores.
□ I didn’t know that I could send my scores out after test day.

---

25. Please select the statement(s) that best describe your reasons for not sending your GRE General Test scores out at the testing center on the day of the test. (Check all that apply.)

□ I planned to take the test again.
□ My score(s) were too low.
□ I was not sure which schools I was going to apply to.
□ I wanted more time before deciding which of my scores to send.
□ The GRE General Test is not an admission requirement of my graduate programs.

26*. After test day*, did you send any GRE General Test scores out?

○ Yes
○ No
○ I don't recall

---

27. Please select the statement(s) that best describe your reason(s) for sending your GRE General Test scores out *after test day*. (Check all that apply.)

□ I had not sent any scores on the day of the test.
□ I chose my best scores and sent only those.
□ I sent all of my scores (because I was not aware that I could select).
□ I wanted to send additional scores.

---

28. Please select the statement(s) that best describe your reasons for not sending your GRE General Test scores out *after test day*. (Check all that apply.)

□ I couldn't afford the fee for sending each new set of scores.
□ I did not need to send any additional scores out.
□ I did not know I could send out my scores after the day of the test.

29. Did you know the following?: If time and money were not an issue, you have nothing to lose by retaking the GRE General Test. Why? If you don't perform well, you can use ScoreSelect to omit sets of scores from your score reports.

○ No, I did not know that.
○ Yes, I was aware of that.

30. Knowing you can select which scores to send to schools, would you be:

○ More likely to take the GRE General Test again
○ Less likely to take the GRE General Test again
○ Neither more nor less likely to take the GRE General Test again.

Thank you for sharing your perspectives with us. We appreciate your participation and we wish you the best as a graduate student or professional in your chosen field.

Please provide your name and email address so we can send you a gift card as a token of our appreciation.

First name: ____________________________
Last name: ____________________________
Email address: ____________________________

Suggested citation:

Klieger, D. M., Kotloff, L. J., Belur, V., & Schramm-Possinger, M. E., Holtzman, S. L., & Bunde, H. (2022). *Studies of possible effects of* GRE® ScoreSelect® *on subgroup differences in* GRE® *General Test scores* (Research Report No. RR-22-13). ETS. https://doi.org/10.1002/ets2.12356

Action Editor: Brent Bridgeman

Reviewers: GRE TAC and GRE Board

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/