International Journal of
STEM Education

REVIEW Open Access

Validity, acceptability, and procedural issues of selection methods for graduate study admissions in the fields of science, technology, engineering, and mathematics: a mapping review

Anastasia Kurysheva[1,2*], Harold V. M. van Rijen[1,2], Cecily Stolte[1,2] and Gönül Dilaver[1,2]

Abstract

This review presents the first comprehensive synthesis of available research on selection methods for STEM graduate study admissions. Ten categories of graduate selection methods emerged. Each category was critically appraised against the following evaluative quality principles: predictive validity and reliability, acceptability, procedural issues, and cost-effectiveness. The findings advance the field of graduate selective admissions by (a) detecting selection methods and study success dimensions that are specific for STEM admissions, (b) including research evidence both on cognitive and noncognitive selection methods, and (c) showing the importance of accounting for all four evaluative quality principles in practice. Overall, this synthesis allows admissions committees to choose which selection methods to use and which essential aspects of their implementation to account for.

Keywords Review, Graduate admissions, Selection methods, Predictive validity, Acceptability

Introduction

A high-quality student selection procedure for graduate level education is of utmost importance for programs, students, and society. Higher education has seen several influential policy developments over the past decades such as the introduction of the Bologna Process in 1999 in Europe and the increased internationalization of higher education across the globe (De Wit & Altbach, 2020). These policies contributed to rising international/cross-border and national (i.e., between higher education institutions within one country) student mobility (Okahana & Zhou, 2018; Payne, 2015). The knock-on effect of this mobility has created a growing diversity of graduate application files. Admissions committees are now faced with applicants from different higher education systems, potentially a variety of background fields, and varying levels of academic skills and proficiency in the language of instruction.

Furthermore, the problem of underrepresentation of students with certain backgrounds persists across the globe, including countries with well-developed higher education systems (Salmi & Bassett, 2014). As such, it is still more difficult for students with low socioeconomic status (SES), a migration background, those who are first-generation students, or students with disabilities to gain admissions into higher education programs compared to students with middle/high SES, no migration background, parents who hold academic degree, or students

*Correspondence:
Anastasia Kurysheva
a.kurysheva@umcutrecht.nl
[1] Center of Education and Training, University Medical Center Utrecht, HB-4.05, P.O. Box 85500, 3508 GA Utrecht, Netherlands
[2] Graduate School of Life Sciences, Biomedical Sciences department, Utrecht University, HB-4.05, P.O. Box 85500, 3508 GA Utrecht, Netherlands

without disabilities (Garaz & Torotcoi, 2017; Salmi & Bassett, 2014; Weedon, 2017). Students' application files are often conditioned by their background: For example, students with parents of low SES cannot typically show an impressive list of extracurricular activities on their resume in contrast to their peers with parents of high SES (Jayakumar & Page, 2021). Since these factors contribute to the inequality already at the entrance to higher education—at the undergraduate level (Zimdars, 2016), they may further exacerbate their effects in the selective graduate level of education, where there are even fewer places available. It is, therefore, often the case that a straightforward assessment of application files is not feasible because of the multifaceted nature of each application. Unsurprisingly, it is a complex task for admissions committees to evaluate the educational background and achievements of (inter)national students with diverse backgrounds. Regardless of described complexities, admissions decisions must be objective, fair, and transparent to ensure their adequate justification.

Evaluative quality principles

To facilitate the achievement of the overarching goals of objectivity, fairness, and transparency, four evaluative quality principles regarding student selection methods were recognized as essential (Patterson et al., 2016):

A) Effectiveness combines both (predictive) (incremental) validity and reliability. This principle encompasses several questions that should ideally be considered together: Does a selection method predict study success and to what extent? Even if a selection method does predict study success, does it provide additional value beyond other valid selection methods? Does the use of a selection method deliver consistent results across time, locations, and assessors?

B) Procedural issues of a selection method refer to any aspects that are important in the practical implementation of the method such as its limitations, the impact of its structure and format on its effectiveness, any biases that are naturally integrated into its design etc.

C) Acceptability refers to both the willingness to implement a selection method and the satisfaction of stakeholders from its usage. Relevant questions in this regard are: How widely is the selection method used across different disciplines, countries, and regions? To what extent are admissions committees willing to apply the method? Do they find it useful? Finally, how much do applicants favor the selection method?

D) Cost-effectiveness is a quality evaluative principle that refers to the financial impact of a selection method on educational programs and applicants. In other words, it refers to the questions: Who pays for its usage in the admissions process, and how much does it cost?

There is a striking lack of studies that synthesize research evidence on selection methods for graduate study admissions while accounting for all four evaluative quality principles. Instead, the existing reviews and meta-analyses address evidence for each selection method separately: standardized testing (Kuncel & Hezlett, 2007b, 2010; Kuncel et al., 2004, 2010), recommendation letters (Kuncel et al., 2014), personal statements (Murphy et al., 2009), and other various noncognitive measures (Kuncel et al., 2020; Kyllonen et al., 2005, 2011; Megginson, 2009). Moreover, these studies usually focus on predictive validity and rarely on procedural issues, with only limited or no attention to reliability, acceptability, and cost-effectiveness.

The only review to combine evidence on all available selection methods within one study and included the four evaluative quality principles (validity/reliability, procedural issues, acceptability, and cost-effectiveness) was conducted by Patterson et al. (2016). However, this review only focused on selection methods in medical education. For example, it does not present evidence on (nonmedical) standardized tests of academic aptitude, tests of language of instruction, or amount and quality of prior research experience. Therefore, its findings can only be partially generalized for graduate admissions.

The question that arises is which educational field (except medical education) has attracted enough high-quality research that (a) addresses the four evaluative quality principles and (b) allows admissions committees to use the findings in a wide range of graduate programs, therefore, enhancing the potential impact of this review? From the preliminary overview, we think that science, technology, engineering, and mathematics (STEM) fields meet these two conditions. STEM fields have been recognized worldwide as fundamental for finding solutions to urgent societal problems (Proudfoot & Hoffer, 2016). The efforts of certain countries to become leaders in STEM higher education and research (e.g., China; Kirby & van der Wende, 2019) are illustrative of how crucial the STEM fields are for economic growth and prosperity. Unsurprisingly, STEM disciplines have attracted a rising number of students, making research evidence on selection methods for STEM studies increasingly more relevant. Since there has been no synthesis of such evidence to date, we designed this review to address this gap.

The present review

The aim of this review is to present a comprehensive overview of research evidence on the existing selection methods in graduate admissions in STEM fields. The review focuses on evaluative quality principles of validity, reliability, procedural issues, acceptability, and cost-effectiveness. The term "graduate" refers to both master's and doctoral levels. That is, studies on both levels were collected for this review.

Research questions

What evidence is provided in research literature within STEM graduate admissions field on:

A) the extent to which different selection methods are valid and reliable?
B) procedural issues of the selection methods?
C) the extent to which different selection methods are accepted by stakeholders?
D) the extent to which different selection methods are cost-effective?

Methods

For this review, a systematic search was conducted and complemented with an expanded search of literature in reference lists of relevant books and articles.

Inclusion criteria for the literature review

The inclusion criteria for this review were: (1) the topic on selection methods in graduate admissions, (2) the graduate level of education (i.e., master's and/or PhD phase), (3) samples that include students from STEM disciplines, (4) studies addressing at least one of four evaluative quality principles of interest: validity/reliability, procedural issues, acceptability, and cost-effectiveness, (5) studies conducted in at least one of the Organization for Economic Co-operation and Development (OECD) countries, (6) studies published in English, (7) studies that went through a peer-review process, (8) studies conducted in the period between 2005 and June 2023.

The OECD countries were chosen because of their well-developed higher education systems as well as an expectation that the quality of research in these countries is comparable. The time frame was chosen in accordance with the changes in European higher education systems after the introduction of the Bologna Process (The Bologna Declaration, 1999). Countries joined the process in different subsequent years. Therefore, 2005 was chosen as a plausible cut-off moment to account for the fact that the first students, studying within the new system, could graduate. The same time frame was applied for the US research context.

We chose to review the literature, referring to master's and PhD levels together (that is, on a graduate level overall), because the training on both levels is advanced. Furthermore, many studies that were included in this review did not make a distinction between the two levels. We also considered different STEM majors or contexts (e.g., the European vs. the US contexts) together, because we aimed to detect overarching patterns in evaluative quality principles that would be applicable to a variety of majors and higher education contexts on a graduate level.

The literature search procedure

The literature search delivered 3244 potentially relevant items including duplicates. The main portion of the results was obtained via conducting a systematic search in the specialized databases (ERIC: $n = 1089$; PsycInfo: $n = 1112$; Medline: $n = 234$; Scopus: $n = 649$). The keywords of the systematic search can be found in Additional file 1: Table S1. The syntax for each database is available upon request. While we did not have the opportunity to carry out searches in all specific databases for each STEM education field (e.g., databases focusing on engineering education), we expect that the large educational data bases such as ERIC contain a substantial number of studies related to our topic in each of those fields. Next, the literature search was extended beyond the database approaches. Namely, the citations from relevant articles were examined ($n = 71$), and previously collected research literature was added ($n = 89$). The screening was conducted in two steps. In the first step, the titles and abstracts were scanned to remove duplicates and obviously irrelevant search results. In the second step, the full texts of remaining articles were obtained and examined. The full texts of four articles were not found even after contacting the authors and were not included in the final number.

Figure 1 presents a detailed flowchart of the steps undertaken. Two coders (the first and the third authors) conducted both steps of screenings. To ensure that the same papers were selected, both coders screened all papers at both steps according to the inclusion criteria. They used codes, such as "yes", "no", and "may be", with the later meaning that an article required a joint decision during the discussion. All papers were independently screened by the two coders during both steps. Although the agreement after the first screening was near complete (kappa = 0.88) and that of the second screening was strong (kappa = 0.70), there were papers with different codes (e.g., "yes" and "may be", or more rarely "yes" and

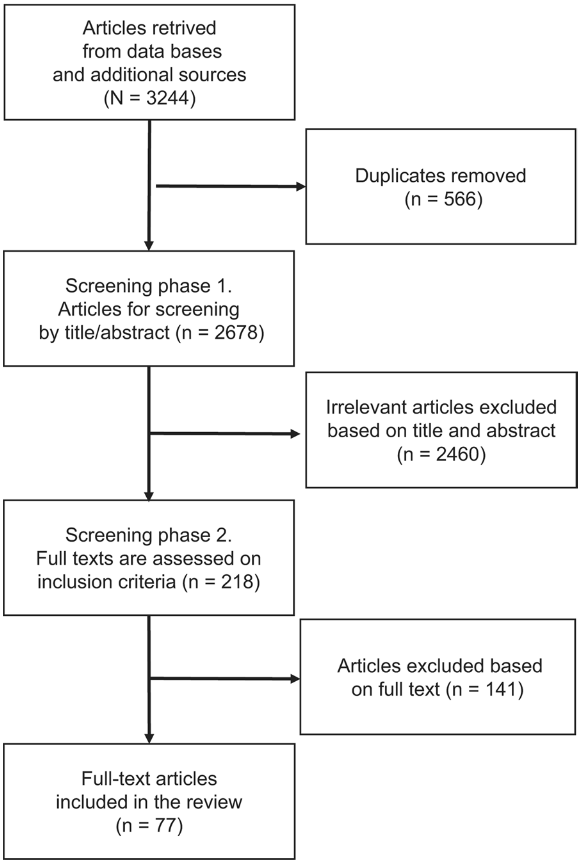
Articles retrived from data bases and additional sources (N = 3244)
Duplicates removed (n = 566)
Screening phase 1. Articles for screening by title/abstract (n = 2678)
Irrelevant articles excluded based on title and abstract (n = 2460)
Screening phase 2. Full texts are assessed on inclusion criteria (n = 218)
Articles excluded based on full text (n = 141)
Full-text articles included in the review (n = 77)

Fig. 1 Flowchart of articles' selection

“no”) or about which the coders had doubts (a code of “may be”). All such disagreements were resolved through discussion.

In total, 77 articles met the inclusion criteria for this review. The distribution across the OECD countries is presented in Table 1. The distribution across STEM disciplines is presented in Table 2.

Table 1 Distribution of the studies included in the review across the OECD countries

OECD country	Number of studies included in this review
Not specified/across continents	19
Across Europe	2
Belgium	1
Canada	1
Mexico	1
The Netherlands	7
Puerto Rico (the unincorporated territory of the US)	1
Switzerland	2
Turkey	1
The US	42
Total number of articles	77

Table 2 Frequencies of STEM disciplines in studies reviewed

Discipline	Frequencies
Geosciences	2
Pharmacy	1
Natural Sciences overall	2
Technology overall	3
Life sciences overall	7
Mathematics	4
Chemistry	5
Computer Sciences	5
Biology	7
Biomedical sciences	8
Engineering	9
Physics	15
STEM disciplines overall	24

After the screening was completed, the graduate selection methods from 77 studies were assigned into ten categories: (1) prior grades, (2) standardized testing of academic abilities, (3) letters of recommendation, (4) interviews, (5) personal statements (i.e., motivation letters), (6) personality assessments, (7) intelligence assessments, (8) language proficiency, (9) prior research experience, and (10) various, rarely studied selection methods that do not fall under more common methods above (such as resumes, selectivity of prior higher education institution (HEI), former (type of) HEI, amount and quality of research experience, or composite scores). If one study addressed different methods or evaluative quality principles, that study was included in all respective categories. The numbers of papers cross-tabulated according to selection method and evaluative quality principle are presented in Additional file 1: Table S2. Additional file 1: Table S3 shows the main characteristics of studies, such as study design, country, field of study, and so forth. Additional file 1: Table S3 also includes the summary of the relevant findings per study.

Contributions of this review

The main contribution of this review is that it synthesizes high-quality research evidence across four evaluative quality principles, as proposed by Patterson et al. (2016), for both cognitive and noncognitive selection methods. No such synthesis has been conducted in the field of STEM graduate admissions (For an overview of the assessment of only noncognitive constructs in graduate education, one may consult the papers of de

Boer and Van Rijnsoever, 2022a; Kyllonen et al., 2005, 2011). Another strong aspect of this review is that it compares the findings of primary and secondary (i.e., reviews, meta-analyses) studies, wherever possible. This is important considering possible limitations of primary studies, such as range restriction and criteria unreliability, which can be accounted for in meta-analyses (Sedlacek, 2003). Overall, this review aims to provide a compilation of state-of-the-art research on selective graduate admissions in STEM fields of study.

Results

Additional file 1: Table S2 shows the numbers of articles on each selection method and evaluative quality principle. We note the overall lack of research on the topics of reliability and cost-effectiveness. Therefore, the evidence below is presented mostly on validity, acceptability, and procedural issues. When studies on reliability or cost-effectiveness are available, they are reported in the respective selection methods' categories.

Prior grades

Validity and reliability of prior grades

The research focused on exploring the predictive validity of different aspects of grade point average (GPA), such as undergraduate GPA (UGPA), the first-year GPA, and the last-year GPA. Findings and relevant references are presented in Table 3. Overall, it appears that UGPA is a valid predictor of graduate degree completion, student performance on introductory graduate courses, and graduate GPA (GGPA). However, UGPA is not valid for predicting research productivity (defined as number of published papers, presentations, and obtained grants) and passing qualifying exams. There is mixed evidence on predictive validity of UGPA toward time to graduate degree and faculty ratings.

Some single studies looked at UGPA in more detail. Namely, they disentangled UGPA on first-year UGPA and last-year GPA. A study that tried to predict graduate degree completion with first-year UGPA found no such relationship (DeClou, 2016). A study that explored the predictive validity of last-year UGPA found that last-year GPA is positively related to GGPA (Zimmermann et al., 2017a).

Table 3 Research evidence on validity of prior grades

Valid for the following dimensions of study success (References)	Exceptions or additional findings	Mixed /not sufficient evidence for the following dimensions of study success (References)	Not valid for the following dimensions of graduate success (References)
Undergraduate grade point average (UGPA)			
Graduate degree completion *Positive relationship* (Kurysheva et al., 2022a; Mendoza-Sanchez et al., 2022; Moneta-Koehler et al., 2017; Verostek et al., 2021; Wollast et al., 2018)	**Graduate degree completion** *No relation* (Cox et al., 2009)	**Time to graduate degree** *Negative relationship* (Howell et al., 2014; Kurysheva et al., 2022a ; Mendoza-Sanchez et al., 2022) *No relation* (Moneta-Koehler et al., 2017)	**Research productivity,** defined as number of published papers, presentations and obtained grants (Howell et al., 2014; Moneta-Koehler et al., 2017)
Performance on introductory courses (Moneta-Koehler et al., 2017; Park et al., 2018; Willcockson et al., 2009)		**Faculty ratings** *Positive relationship*: ratings (Moneta-Koehler et al., 2017) *No relation* (Howell et al., 2014)	**Passing a qualifying exam** (Burmeister et al., 2014; Moneta-Koehler et al., 2017)
Graduate grade point average (GGPA) (Bridgeman et al., 2009; Burton & Wang, 2005; Howell et al., 2014; Kurysheva et al., 2022a ; Kurysheva, van Ooijen-van der Linden et al., 2022; Moneta-Koehler et al., 2017; Verostek et al., 2021; Zimmermann et al., 2015)	**Graduate grade point average (GGPA)** The overall good predictive power of UGPA differs per field of study. UGPA has a stronger predictive validity toward GGPA in chemistry departments than it has in biology departments (Burton & Wang, 2005). It also depends on how much narrowed the range UGPA in a study is (Burmeister et al., 2014)		
The first-year undergraduate GPA			
		Graduate degree completion *No relationship* (DeClou, 2016)	
The last-year undergraduate GPA			
		GGPA *Positive relationship* (Zimmermann et al. 2015)	

We found one study that addressed the question of reliability estimates. The author calculated eight different reliability coefficients for fourth-year cumulative GPA at each higher education institution included in the study and then meta-analyzed them (Westrick, 2017). The study showed that the various reliability estimates ranged between 0.89 and 0.92. The author recommends using stratified alpha as a reliability coefficient for cumulative GPA, which works best with the multi-factor data, due to the variation in the processes involved in earning grades in the first-year and fourth-year courses (Westrick, 2017).

Procedural issues of prior grades

There are several procedural issues with using prior grades for admissions decisions. The first one is grade inflation—a practice of awarding higher grades than previously assigned for given levels of achievement (Merriam-Webster dictionary, n.d.): For example, teachers giving higher grades for positive student ratings (European Grade Conversion System [EGRACONS], 2020). In her observational study of top graduate research programs, Posselt (2014) indicated that grade inflation is a widespread phenomenon in highly selective universities. In such universities, students from underrepresented backgrounds are extremely lacking; therefore, setting a grade-threshold on a high level disproportionately excluded these students (Posselt, 2014).

The second one refers to differences in grading standards, which relates to the fact that one grade obtained at different institutions might reflect a different level of academic qualification. Grade conversion and grade distribution tables, which are developed to tackle these issues, are not without limitations. They can often be crude, and this can affect both selection decisions and research done on grades as predictors of graduate study success (see, e.g., Zimmermann et al., 2017a).

The third procedural issue relates to a possibility of cognitive biases of assessors to influence grading: This could be an origin of differences in prior grades observed between applicants with various socioeconomic status (SES), genders, and races (Woo et al., 2023). Finally, the relatedness, or fit, between undergraduate and graduate programs affects the predictive value of grades received during undergraduate studies: When the programs are related to a high extent, the relationship between undergraduate and graduate grades is stronger compared to a situation when the undergraduate and graduate programs are related to a low extent (de Boer & Rijnsoever, 2022b).

Acceptability of prior grades

Prior grades are a widely accepted selective admissions method (Boyette-Davis, 2018; MasterMind Europe, 2017). The largest weight in admissions decisions is given to grades on undergraduate courses that are closest in terms of content to the courses of a graduate program (Chari & Potvin, 2019). When explaining what the reasons are behind high acceptability of grades and even overestimation of their importance in graduate admissions by admissions committees, Posselt (2014) states that high conventual achievements, such as grades, are consistent with the identity of an elite intellectual community, which admissions committee members, implicitly or explicitly, refer themselves.

Standardized testing of academic abilities

Validity of standardized admissions tests of academic abilities

Among different standardized admissions tests, the ones which are typically required for selective admissions to graduate programs in STEM disciplines are the Graduate Record Examinations (GRE) General and GRE Subject. All but one study, which addressed validity of standardized tests, referred to these two GRE tests. The only exception was the standardized test EXANI-III, which is used in Mexico.

Validity of graduate standardized admissions tests has been a controversial topic in research, with some studies providing evidence for their weak-to-moderate predictive power toward graduate study success and others indicating the absence of predictive power (see Table 4). From Table 4, we can infer that the standardized test most often examined is the GRE General.

The GRE General is a positive predictor of first-year GGPA, GGPA, and faculty ratings. This is in line with the existing reviews and meta-analyses (Kuncel & Hezlett, 2007b, 2010; Kuncel et al., 2010). From the majority of primary studies, it appears that the GRE General does not predict graduate degree completion and research productivity defined as the number of publications.

The meta-analyses on the topic, however, found that after meta-analytical corrections for statistical artifacts in primary studies were applied (such as a correction for the restriction of range of a predictor), these two relationships (1) between the GRE General and degree completion and (2) between the GRE General and research productivity, although weak, were detected (Kuncel & Hezlett, 2007a, 2007b).

Finally, there was mixed or limited evidence for GRE General efficiency in prediction of time to graduate degree, performance on core program courses, qualifying exam, rate of progress, and thesis performance (see Table 4 for details).

There is an indication that another standardized test, the GRE Subject in Physics, is predictive for faculty ratings, while its predictive value for graduate degree

Table 4 Research evidence on validity of standardized tests of academic abilities

Valid for the following dimensions of study success (References)	Exceptions or additional findings	Mixed /not sufficient evidence for the following dimensions of study success (References)	Not valid for the following dimensions of study success (References)	Exceptions or additional findings
GRE General				
First-year graduate GPA (Bridgeman et al., 2009; Burmeister et al., 2014; Moneta-Koehler et al., 2017)		**Time to graduate degree** *No relationship* (Hall et al., 2017; Mendoza-Sanchez et al., 2022; Moneta-Koehler et al., 2017; Petersen et al., 2018; Sealy et al., 2019) *Positive relationship*: the higher the GRE scores, the longer the time (Howell et al., 2014; Lorden et al., 2011)	**Graduate degree completion** (Cox et al., 2009; Lorden et al., 2011; Lott et al., 2009; Mendoza-Sanchez et al., 2022; Miller et al., 2019; Moneta-Koehler et al., 2017; Petersen et al., 2018)	When a relative GRE score is considered instead of its absolute score (i.e., compared to student's peers in a program), one study has found GRE General to be efficient in distinguishing students with lower and higher odd of attrition in one of the studies (Lott et al. 2009)
GGPA (Burton & Wang, 2005; Howell et al., 2014; Klieger et al., 2014; Moneta-Koehler et al., 2017; Zimmermann et al., 2017a, 2017b)	**GGPA** No relationship of GRE-Q and GRE-V to the GGPA in Physics (Verostek et al., 2021)	**Performance on core program courses** The findings are ambiguous: one study found that the correlation is positive (Burmeister et al., 2014), another one showed that while the GRE General contribution existed in univariate analysis, it did not hold in the adjusted model (Park et al., 2018). There is another study showing that only GRE-V, but not GRE-Q was positively related to a core course performance (Willcockson et al., 2009)	**Research productivity** defined as number of publications (Hall et al., 2017; Moneta-Koehler et al., 2017; Sealy et al., 2019)	One study is an exception, where GRE-Q was the only and very weak predictor of number of publications, explaining 5% of variance in the outcome (Howell et al., 2014)
Faculty ratings (Burmeister et al., 2014; Burton & Wang, 2005; Howell et al., 2014; Moneta-Koehler et al., 2017)	An exception is a study that found the negative relationship between the GRE General and faculty rankings: better ranked PhD students tended to have lower GRE scores from their faculty mentors (Sealy et al., 2019). Though the authors state that they had a wide span of the GRE scores in their data, the sample size in this study was extremely low: 28 students, and part of the analysis was conducted on the lower and upper quartiles of the GRE scores, lowering the sample size even more	**Qualifying exam** *No relationship* (Moneta-Koehler et al., 2017) *Positive relationship* (Burmeister et al., 2014)	**Conference presentations** (Moneta-Koehler et al., 2017)	
		Rate of progress GRE-Q is a weak predictor of rate of progress (Zimmermann et al., 2017a, 2017b)	**Obtaining grants, fellowships, or awards** (Moneta-Koehler et al., 2017; Sealy et al., 2019)	
		Thesis performance *No relationship* (Zimmermann et al., 2017a, 2017b)		

Table 4 (continued)

Valid for the following dimensions of study success (References)	Exceptions or additional findings	Mixed /not sufficient evidence for the following dimensions of study success (References)	Not valid for the following dimensions of study success (References)	Exceptions or additional findings
Graduate Record Examinations Subject (Physics)				
Graduate GPA (Verostek et al., 2021) **Faculty ratings** (Burmeister et al., 2014)			**Graduate degree completion** (Miller et al., 2019)	The study of Miller et al. (2019) was heavily criticized for several issues in research design and statistical approach (Weissman, 2020), thus suggesting that the finding must be regarded with a lot of caution
Other standardized tests (EXANI-III in Mexico)				
			Graduate degree completion (Álvarez-Montero et al., 2014)	

completion remains unclear. Two meta-analyses also found that the GRE Subject is a meaningful predictor of graduate study success (Kuncel & Hezlett, 2007b; Kuncel et al., 2010).

Procedural issues of standardized admissions tests of academic abilities

The primary studies showed a possibility of (1) adverse impact of the GRE on underrepresented groups (including ethnic minorities and females in STEM), which can be mitigated by applying a systematic and holistic approach in reviewing admissions files (Bleske-Rechek & Browne, 2014; Murphy, 2009; Posselt, 2014; Wilson et al., 2018, 2019), and (2) item position effects, which can be mitigated by allowing proper time limits for taking the test (Davey & Lee, 2011).

However, the reviews and meta-analyses on procedural issues refuted several common beliefs regarding standardized tests, such as: (1) the coaching effects, which were shown to be modest with one quarter of a standard deviation improvement in test performance (Hausknecht et al., 2007; Kuncel & Hezlett, 2007a, 2007b). Such an improvement refers primarily toward the GRE Analytical Writing section (GRE-A) (Powers, 2017). GRE Verbal Reasoning (GRE-V) and GRE Quantitative Reasoning (GRE-Q) were prone to coaching to a negligible extent in contrast to claims of commercial organizations that prepare test takers for standardized tests (Powers, 2017); (2) lack of predictive independence from SES, which was contested by demonstrating that even after controlling for SES, standardized test scores remained predictive of study success (Camara et al., 2013; Kuncel & Hezlett, 2010); (3) bias in testing. Some researchers state that bias in graduate testing is a myth, as, according to their findings, standardized tests appeared to predict graduate study success of both females and males equally (Fischer et al., 2013; Kuncel & Hezlett, 2007b) as well as ethnic groups (Kuncel & Hezlett, 2007b). The authors of these studies also indicated that the differences in performance between different groups might reflect societal problems, such as lack of family, social, environmental, peer, and financial support. They state that standardized tests simply expose the preexisting differences created by the above-mentioned societal problems (Camara et al., 2013; Kuncel & Hezlett, 2010); (4) negative effect of stereotype threat on standardized test performance: Test takers, who believe that their nonoptimal performance on standardized tests might confirm the stereotypes of their minority group's intellectual capacity, might perform worse because of that self-fulfilling prophecy (Garces, 2014).

Acceptability of standardized admissions tests of academic abilities

Acceptability by admissions committees In the US context, admissions committees—especially for research programs—actively use the GRE General and consider it to be a valuable contributor for their admissions decisions (Boyette-Davis, 2018; Chari & Potvin, 2019; Rock & Adler, 2014). Out of the three sections, GRE-V and GRE-Q are used most, while GRE-A is considered the least often (only around 35% of surveyed programs; Briihl & Wasieleski, 2007). When it comes to positioning GRE as a selection method, the GRE appeared less important than, for example, previous research experience, UGPA, and certain personal characteristics (e.g., critical thinking, work ethics; Boyette-Davis, 2018). However, the GRE had more weight in selection decisions for doctoral programs than for masters' programs (Chari & Potvin, 2019).

A survey among masters' programs in Europe showed that the results of standardized admissions tests are rarely used for elimination purposes (only around 5% masters' programs admitted such a practice), but higher scores, if present, do provide an advantage to students in one fourth of the programs (MasterMind Europe, 2017). However, Europe has seen a steady increase in GRE test takers (e.g., it increased from 12,243 in 2004 to 29,211 in 2013) since the introduction of the Bologna Process and the increasing internationalization of European graduate education (Payne, 2015). Test takers aiming to study STEM disciplines represented the largest group among all European GRE test takers (Payne, 2015).

Acceptability by applicants Applicants viewed the GRE as less important in graduate admissions than UGPA, recommendation letters, and work experience (Cline & Powers, 2014). Applicants coming from racial minority groups had more negative feelings about the GRE than white test takers (Cline & Powers, 2014). International students felt that the GRE is culturally biased (Mupinga & Mupinga, 2005). Applicants perceived publishing prompts from GRE-A positively (Powers, 2005) and desired to get additional information about their writing skills beyond their GRE-A score (Attali & Sinharay, 2015).

Cost-effectiveness of standardized admissions tests of academic abilities

One study looked at this evaluative quality principle. In their study, Klieger et al. (2014) provided an example of calculation of the benefits for one US doctoral program. They estimated the financial benefits of using the GRE for admissions and funding decisions as considerable, but obviously, the exact numbers will depend on a specific program and a number of GRE sections used for admissions decisions.

Letters of recommendation (LoRs)

Validity and reliability of letters of recommendation

The only primary study which examined predictive validity of LoRs for STEM disciplines (namely, the biomedical sciences) found that the scores on LoRs did not predict time to degree, but they were the most powerful predictor of first-author student publications (Hall et al., 2017). The review of Kuncel et al. (2014) showed that LoRs do not deliver incremental validity over standardized admissions tests and UGPA toward GGPA and faculty ratings but do deliver small incremental validity in prediction of degree completion (an outcome usually difficult to predict using other measures). The review of Megginson (2009) showed that narrative LoRs have minimal reliability and are prone to subjective interpretations.

Procedural issues of letters of recommendation

The primary studies that explored biases in narrative LoRs at the graduate level found evidence of: (1) gender and race biases (Biernat & Eidelman, 2007; Morgan et al., 2013); (2) bias arising from tone of LoRs (Posselt, 2018); (3) bias arising from admissions committees' members being (un)familiar with the LoR writer (Posselt, 2018); (4) bias in admissions committees' evaluations against underrepresented minority groups once applicants' names are visible (Morgan et al., 2013). Requiring admissions committees to elaborate on their evaluations of narrative LoRs reduces biases (Morgan et al., 2013).

Acceptability of letters of recommendation

Two primary studies explored the acceptability of LoRs. One study showed that LoRs are the second most valued selection method in admissions to doctoral programs in the US context, because they shed light on applicants' personal characteristics (Boyette-Davis, 2018). However, another study in the European context did not find that LoRs are given weight by admissions committees when they decide to reject or admit a student to a master's program (MasterMind Europe, 2017). In the latter study, more than a half (58.3%) of surveyed applicants reported that they had to provide an LoR within their application file.

Interviews

Validity of interviews

Evidence on validity of interviews in STEM graduate programs is limited to two studies. One focused on traditional interviews and the other on the highly structured and formalized form of interviews: multiple mini-interviews (MMIs). Traditional interviews do not allow to distinguish between most and least productive graduate students (in terms of their time to degree and number of first-author papers; Hall et al., 2017). However, MMIs allow to predict planning-related problematic study behavior (oude Egbrink & Schuwirth, 2016).

Procedural issues of interviews

No study addressed the procedural issues of interviews specifically in graduate admissions.

Acceptability of interviews

A survey among European masters' programs demonstrated that interviews are used in 22.6% of English-taught masters' programs across Europe (MasterMind Europe, 2017). Although it is not a widely used selection method, it is valued and regarded as a good practice by admissions committees. In addition, members of admissions committees reported that a poor interview is a reason for rejection in less than 5% of all cases. No studies were conducted on how favorable interviews are perceived by applicants to graduate programs.

Cost-effectiveness of interviews

Interviews can be expensive both for applicants and graduate school (Woo et al., 2023). Applicants may be required to travel and/or to take time off from their work for an interview. In addition, they usually take time to prepare for it. On the side of graduate schools, interviewing takes substantial time investment of admissions committees both for preparation and for conducting the interviews.

Personal statements (motivation letters)

Validity of personal statements

A meta-analysis on predictive validity of personal statements showed that they were weak predictors of grades and faculty ratings and when considered together with the UGPA and standardized admissions tests, they provided no incremental validity (Murphy et al., 2009).

Procedural issues of personal statements

Woo et al. (2023) bring attention to the fact that financial and social capitals are of great asset for richer students who seek help in writing personal statements. The same authors indicate that prior research has shown that men tend to use more acting and self-promotional tone in writing than females, which can have direct effects for creating biases in graduate admissions toward men (Woo et al., 2023).

Acceptability of personal statements

Personal statements are used frequently (MasterMind Europe, 2017) and are required from international applicants almost twice as often as from internal applicants (i.e., those, who obtained a bachelor's degree at the same institution; MasterMind Europe, 2017). Personal statements are used to assess students' motivation, make inferences about personal qualities, previous academic background, and cognitive ability (Kurysheva et al., 2019), provide information on whether a student's background will contribute to the diversity of the student body (Posselt, 2014).

In most cases, personal statements did not serve as a reason for failure in the admissions process, according to members of admissions committees (MasterMind Europe, 2017).

Intelligence assessments

Validity of intelligence assessments

Intelligence assessments are significantly correlated with academic performance (defined as grades, results of educational tests, and procedural and declarative knowledge; Poropat, 2009; Schneider & Preckel, 2017).

Procedural issues of intelligence assessments

Practical utility of intelligence as a predictor of study success is usually reduced, because it overlaps significantly with measures of prior performance (e.g., grades; Poropat, 2009).

Acceptability of intelligence assessments

In a cross-sectional study on the samples of students in the life sciences and natural sciences, it was shown that admissions criteria related to intelligence play a moderately important role in admissions decisions along with several other admissions criteria (Kurysheva et al., 2019). However, those admissions committees participating in the study did not apply specific intelligence assessments in their programs; the inferences on student intelligence were made from other selection methods rather than specific intelligence testing (Kurysheva et al., 2019).

Personality assessments

Validity of personality assessments

The most common personality assessment is based on the five-factor model named the "Big Five". It distinguishes five primary factors of personality (Goldberg, 1993): (1) conscientiousness, and it is one of the most stable findings both from individual and meta-analytical studies that conscientiousness is a medium-to-large predictor of study success (Butter & Born, 2012; Poropat, 2009; Schneider & Preckel, 2017; Trapmann et al., 2007; Walsh, 2020); (2) agreeableness, with mixed findings regarding its predictive value; (3) openness to experience, also has mixed findings, (4) neuroticism with no significant relation to study success, (5) extraversion with no significant relation to study success (Poropat, 2009; Trapmann et al., 2007).

Other personal traits, not explicitly included in the Big Five, were also examined: (1) grit (defined as determination to achieve long-term goals), which does not explain additional variance in study success beyond conscientiousness (Walsh, 2020); (2) emotional intelligence, which has a weak-to-moderate effect on study success (Schneider & Preckel, 2017); (3) need for cognition (defined as an inclination to value activities that include effortful cognition), which has a weak-to-moderate effect on study success (Schneider & Preckel, 2017); (4) conscientiousness related to time management, so-called ecological conscientiousness, which is valid beyond the conventional Big Five in predicting Ph.D. performance criteria such as research progress, meeting deadlines, and probability to obtain a Ph.D. degree on time (Butter & Born, 2012).

Procedural issues of personality assessments

Two procedural issues of personality assessments are referred to in the context of graduate admissions: applicant faking and their coachability (Kyllonen et al., 2005). They arise from the fact that personality assessments are typically based on self-reports.

Acceptability of personality assessments

While graduate admissions committees regard personality assessment important to consider in principle (Kyllonen et al., 2005), they do not report to use them extensively (Boyette-Davis, 2018; MasterMind Europe, 2017).

Language proficiency assessments

Validity of language proficiency

The available evidence on validity of different language assessments toward different dimensions of study success is presented in Table 5.

Procedural issues of language proficiency assessments

No studies were detected that examined procedural issues of language proficiency assessments, specifically for graduate admissions.

Acceptability of language proficiency assessments

Four relevant aspects are worth noting: (1) In the European context, English language assessments were required mostly from foreign applicants to masters' programs, although internal applicants are sometimes expected to submit them as well (MasterMind Europe, 2017); (2) Perceived importance of language proficiency by faculty members depended on a discipline: In humanities, for example, the importance is higher than in science disciplines (Lee & Greene, 2007); (3) Admissions

Table 5 Research evidence on validity of language assessments

Valid for the following dimensions of study success (References)	Exceptions or additional findings	Mixed /not sufficient evidence for the following dimensions of study success (References)	Not valid for the following dimensions of study success (References)
Test of English as a Foreign Language (TOEFL)			
Graduate GPA *Small positive relationship* (Cho & Bridgeman, 2012; Zimmermann et al., 2017a, 2017b)		**First-year GPA** *Positive relationship* (Burmeister et al. 2014). Some studies find the incremental value of TOEFL (Cho & Bridgeman, 2012; Zimmermann et al., 2017a, 2017b)	
		Course average; Faculty ratings *Positive relationship* (Burmeister et al., 2014)	
The Computerized Enhanced ESL Placement Test (CEEPT)			
		First semester academic performance *Mixed findings* (Lee & Greene, 2007)	
A scale that considers the nature of the previous language use			
		Completing a PhD degree in a foreign HEI *Positive relationship* (Mathews, 2007)	

committees usually limit the usage of language proficiencies assessments by checking whether the institutional cutoff score was met. Faculty members often expressed dissatisfaction with the language proficiency of admitted students, because some of them think that the cutoffs reflect not adequate but only minimal required language proficiency (Ginther & Elder, 2014); (4) Test takers do not seem to perceive TOEFL scores as a good indication of one's language abilities (Mathews, 2007).

Prior research experience

Validity of prior research experience

Prior research experience has been shown predictive for research skills performance (Gilmore et al., 2015), master's and doctoral degree completion (Cox et al., 2009; Kurysheva et al., 2022a), GGPA (Kurysheva et al., 2022a; Kurysheva et al., 2022b), faculty ratings (Weiner, 2014), time to degree (Kurysheva et al., 2022a), but not for introductory graduate biomedical course (Park et al., 2018), graduate student productivity (Hall et al., 2017), time to degree (Hall et al., 2017). A meta-analysis showed that research experience during undergraduate studies, defined as a dichotomy "present" or "absent", is unrelated to graduate study success (Miller et al., 2021).

Procedural issues of prior research experience

No studies examined procedural issues of prior research experience specifically in graduate admissions. However, there are concerns raised regarding usage of undergraduate research experience as a selection criterion as it might undermine diversity (Miller et al., 2021) or dilute the education mission of graduate curriculum (Kurysheva et al., 2022a).

Acceptability of prior research experience

It appears that prior research experience is a valued component in graduate admissions (Boyette-Davis, 2018; Chari & Potvin, 2019). However, the extent of its importance depends on whether it is applied to a master's or a doctoral program level (Chari & Potvin, 2019). The extent of importance of prior research experience also depends on what aspects are available for review. For example, simply having a basic level of research experience is significantly more important than having publications or conference participation records (Boyette-Davis, 2018).

Various graduate selection methods

In this category, the selection methods were collected that did not fall in previously reviewed categories: undergraduate institution selectivity, type of prior degree (bachelor's or master's), type of prior higher education institution, a rubric based on or a composite score of different selection methods, rate of progress, duration of prior studies and other specific assessment instruments.

Validity of various graduate selection methods

Undergraduate institution selectivity appears to have a positive relation to performance during the first semester of graduate studies (Moneta-Koehler et al., 2017; Park et al., 2018). Having a prior graduate degree increases the chances of graduate study success

(Willcockson et al., 2009). The last four sub tables of Additional file 1: Table S.3 (S3.26–S3.29) provide details into the findings of single studies on validation of all selection methods, which fell in this category.

Procedural issues of various graduate selection methods

Due to the scarcity of validation studies of the selection methods in this category, the procedural issues remain underexamined. One study addressed academic pedigree as a procedural issue of undergraduate institution selectivity (Posselt, 2018). Academic pedigree is the belief that higher rank of prior HEI signifies stronger student performance potential. In case of academic pedigree, the grades might be interpreted within the context of how rigorous the student's curriculum was at a prior HEI. However, it appears that the selectivity and reputation of prior HEI are not clearly stated but somewhat hidden selection methods (Posselt, 2018). Posselt (2018) underscored that "privileging elite academic pedigrees in graduate admissions preserves racial and socioeconomic inequities that many institutions say they wish to reduce" (p. 497).

Acceptability of various graduate selection methods

Acceptability of selection methods in this category varies. The decisive factors in admissions by graduate admissions committees are as follows: certain undergraduate courses, type of prior academic background, type of prior education institution (Chari & Potvin, 2019).

Other selection methods, even if required, were not given substantial weight in selection decisions (Boyette-Davis, 2018; MasterMind Europe, 2017). Among them are extracurricular activities, teaching experience, quantitative skills, work experience, curriculum vitae (CV), photographs, essays, time management skills, understanding social relevance of research, evidence of integrity. Applicants seem to accept well selection

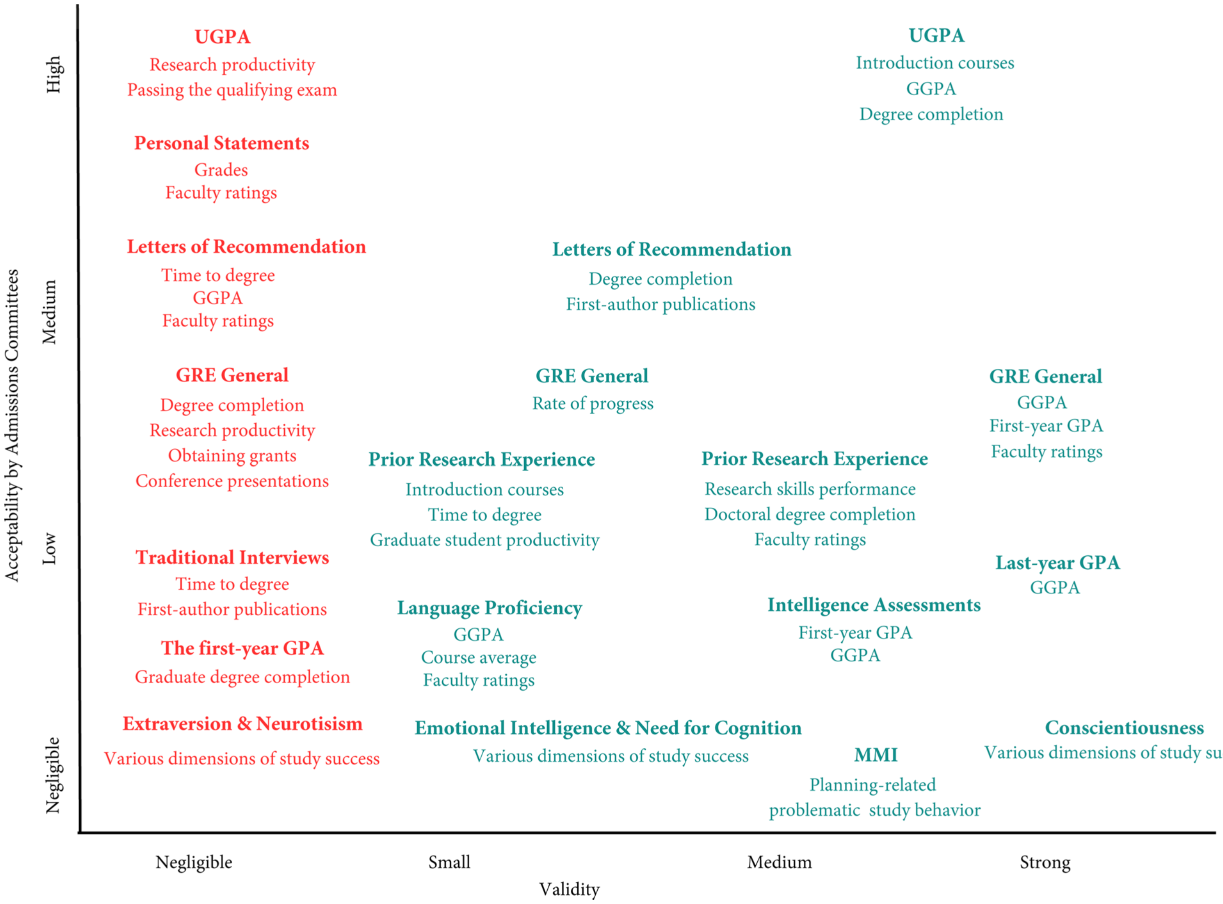

Fig. 2 Summary of the findings on two evaluative quality principles: validity and acceptability by admissions committees *Note*. The location of selection criteria (in a larger font) and the respective dimensions of study success (in a smaller font) are approximations based on the findings of the review. The colors refer to the *X*-axis: Red is used for selection methods that are invalid toward respective dimensions of study success. Green is used for selection methods that are valid toward respective dimensions of study success

methods that consist of different scales, even if the scales concern questions ranging from scientific knowledge to motivation (van Os, 2007).

Cost-effectiveness of various graduate selection methods

Using a total score of a rubric that combines different selection methods substantially increases the admissions rate of underrepresented students without increasing the time investment of admissions committees (Young et al., 2023).

Discussion

This study, which focuses on the available research between 2005 and 2023, is the first review on both cognitive and noncognitive selection methods in graduate education and focuses on STEM disciplines. Studies dedicated to reliability and cost-effectiveness of graduate selection methods were rarely conducted during the examined time span. Therefore, the review's focus was on integrating research evidence on the three evaluative quality principles of predictive validity, acceptability, and procedural issues.

Summary: key findings

Figure 2 provides a visualization of the selection methods located according to the extent of their predictive validity and acceptability by admissions committees. The dimensions of acceptability by applicants or procedural issues are not depicted, because this would require a third and fourth dimensions which would make the figure more difficult to interpret.

The key findings of this review relate to three main evaluative quality principles we examined. The first key finding is that the predictive validity of applied selection methods varies substantially. The medium-to-strong predictors of several graduate study success dimensions are (1) prior grades (including UGPA), (2) GRE General, (3) intelligence assessments, and (4) the personality trait conscientiousness. The following selection methods are also valid, but to a lesser extent: (1) letters of recommendation, (2) tests on language proficiency, (3) personality aspects such as emotional intelligence and need for cognition, (4) undergraduate research experience (when defined as a grade for undergraduate thesis, duration of research project, but not as dichotomous absence or presence of research experience), and (5) MMI (based on limited amount of research). The selection methods in graduate admissions with lack of predictive validity were also detected: (1) personal statements, (2) traditional interviews, and (3) two personal traits (extraversion and neuroticism). This review highlights that the specific selection methods (e.g., the GRE General and UGPA) would appear valid toward certain dimensions of study success (e.g., GGPA) but not the others (e.g., research productivity).

The second key finding shows that the main procedural issues of selection methods are admissions biases, faking, coaching effects, item position effects, test preparation, and stereotype threat. While for some of the methods, the procedural issues constitute a prominent research debate (e.g., a debate on biases involved in implementation of the GRE), the procedural issues of others have not been adequately addressed (e.g., imperfections of grade conversion).

The third finding is that some invalid selection methods are widely accepted by admissions committees, while a similar method with a more structured format and with preliminary indications for validity does not appear to be widespread in STEM admissions. For example, personal statements appear to have negligible validity, especially in the presence of other selection methods but are still widely used (see Fig. 2).

Some evidence from outside of STEM graduate admissions

It is important to note that there is profound research on procedural issues and acceptability of selection methods outside of graduate admissions, namely, in undergraduate admissions and personnel selection. They were not included in results, because they did not fulfill inclusion criteria for this review. However, they are worth mentioning here in the discussion section, because it is unlikely that the procedural issues of the same selection method such as biases, faking, or coaching would be heavily determined by the education level. The following two subsections (procedural issues and acceptability) will, therefore, be dedicated to the outline of those procedural issues and acceptability of some selection methods that received little attention in graduate admissions but were investigated in undergraduate admissions and personnel selection.

Procedural issues

Procedural issues of (traditional) interviews. The current review did not detect studies on procedural issues of interviews in graduate STEM admissions. However, the findings from undergraduate, graduate nonSTEM, and personnel selection research are as follows.

The first procedural issue is susceptibility of interviews to biases toward gender, disability status, and ethnicity. Biases during interviews might come into play at different moments starting from so-called rapport building (a "small chat" aimed at helping applicants to feel comfortable), through the interview itself, and during the evaluation stage after the interview has ended

(Levashina et al., 2014). Reducing bias and increasing validity and reliability of interviews is possible through introducing structure and different formats of interview: for example, phone or video interviews are more adaptable for structuring than face-to-face interviews (Levashina et al., 2014).

The second procedural issue is susceptibility of interviews to subjective interpretations of student "soft variables", such as motivation. A study on a sample of students in a selective college in the Netherlands demonstrated that scores on interviews contribute little to prediction of study success but create risk of subjective interpretations. For example, many of the students whom the interviewers indicated were at risk of expulsion finished their first year successfully (Reumer & van der Wende, 2010). The authors note that "interviews provide extra guidance to both the student and the institution as to whether the student is choosing the right study program (and not so much as whether he is able to complete it successfully)" (Reumer & van der Wende, 2010, p. 20).

The third procedural issue of interviews is faking by applicants, defined as "the conscious distortions of answers to the interview questions to obtain a better score on the interview and/or otherwise create favorable perceptions" (Levashina & Campion, 2007, p. 1639). Among undergraduate job applicants, the estimates of faking, understood in the above-defined broad sense, are as high as 90%, and the estimates of faking that is closer to lying range from 28 to 75% (Levashina & Campion, 2007).

The fourth procedural issue is impression management strategy used by some applicants (e.g., constant smiling), which contributes to admissions committees' perception of these applicants as "glowing" and having "a very nice personality" (Posselt, 2016, p. 144). The fifth procedural issue of interviews is that they provoke a broader actual evaluation of applicants than is formally communicated. For example, it has been shown that sometimes admissions committees' distrust language skills of certain groups of international applicants, and therefore, they use the interview as an additional language check, while proclaiming that they want to assess applicants' knowledge on the subject (Posselt, 2016).

The fifth procedural issue is susceptibility of interviews to weight bias. It was shown that applicants with higher body mass index (BMI) were admitted to a graduate psychology program less frequently than students with lower BMI, and this difference is especially prominent for female applicants (Burmeister et al., 2013).

Procedural issues of personal statements. In the literature outside of STEM graduate selection, namely, in the medical education programs, the biases of gender, age, socioeconomic class, country of origin, and ethnicity were shown to be present in admissions committees' evaluations of personal statements (for the description, see the review of Kuncel et al., 2020).

Procedural issues of personality assessments. Similar to findings in graduate admissions, researchers who conducted studies in undergraduate and personnel selection show that the major procedural issue appears to be faking (Birkeland et al., 2006; König et al., 2017; Pavlov et al., 2019). The extent of faking depends on personality dimension under examination, type of test, aimed position (Birkeland et al., 2006), and situation stakes (Pavlov et al., 2019). However, there are approaches, where supervisors of students are asked to report on their personality, and while the supervisors also tend to fake when reporting on the personality of their students, the extent of their faking is smaller (König et al., 2017).

Acceptability

In personnel selection, a review was conducted on how favorable different selection methods are rated by job applicants. From the review, it appears that the most preferred methods are work sample and interviews; overall favorably evaluated selection methods are resumes, cognitive tests, references, and personality assessments. The least preferred are honesty tests, personal contacts, and graphology (Anderson et al., 2010). Each selection method was assessed on several acceptability scales. For example, perceived scientific validity of LoRs is low, but their interpersonal warmth is high. In contrast to LoRs, intelligence assessments are perceived high on scientific validity and respectful of privacy but low on interpersonal warmth (Anderson et al., 2010). Interestingly, when it comes to structure of interviews, both applicants and interviewers perceive structured interviews less positively than unstructured interviews (Levashina et al., 2014). Similar to interviews, applicants perceive personality assessments favorably, especially the dimension "opportunity to perform" (Anderson et al., 2010).

Graduate selection methods as a distinct area for research

This review maps research evidence on selection methods used specifically at the graduate level. Several selection instruments that are used in admissions to professional schools such as medical school (e.g., situational judgment tests, MMIs, and selection centers) are not used in graduate STEM admissions. What are the potential reasons for this difference? The most obvious difference is that admissions to professional schools are directed toward detecting certain skills and traits of applicants to predict key competencies which are different from those of STEM researchers. The frameworks have been developed that define key competencies in medical profession (e.g., the Canadian Medical Education

Directives for Specialists). They specify the knowledge, skills, abilities, and other characteristics (KSCAOs), related to competent performance within certain healthcare professions (for example, see Kerrin et al., 2018). Like medical education, graduate STEM education is also confronted with the question of which KSCAOs define an engineer or a researcher in STEM fields. A more general question would be even broader: whether a person is a researcher or a professional or not—and if not, why not? Does this have to do with academic freedom of researchers (Vrielink et al., 2011) and their roles as producers of critical knowledge, contributors to expansive learning, and organizers of a space for dialogue (Miettinen, 2004)? Do the existing selection instruments reviewed in this study adequately capture prerequisites for competent performance on researchers' roles? Are there any other selection methods that have potential to do this better? This review might, therefore, be regarded only as one of the first steps toward getting closer to answering such questions.

Implications for research and practice

Implications for research. This review has revealed significant gaps in the existing research, with an extremely low number of papers examining certain selection methods that appear to demonstrate medium and strong validity in graduate education. For example, the validity of MMIs, last-year GPA, and prior research experience have all been investigated in single studies, and the results are promising. To draw more meaningful conclusions, researchers in the field of student selection may wish to study the validity and other evaluative quality principles of these methods across a range of student populations and disciplines.

Implications for practice. From our review, it appeared that the selection methods that have no predictive value in graduate student selection are (1) personal statements; (2) traditional interviews; (3) narrative recommendation letters. Therefore, it is advised to avoid these instruments when making admissions decisions. This, however, does not mean that these instruments cannot be used for other purposes. For example, personal statements may be used for encouraging students to reflect on their motivation for a specific program and getting acquainted with it through exploration of the program's curriculum, internship opportunities, and career perspectives (Wouters et al., 2014).

The variety of selection methods which practitioners should consider including in their selective admissions to research masters' programs in STEM are as follows: (1) undergraduate grade point average (UGPA), (2) GRE General, (3) standardized language tests, such as TOEFL.

With additional caution, the following methods could be considered: (1) prior research experience (for admissions to research graduate programs); (2) GPA for the last year of a bachelor's program; (3) standardized recommendation letters; (4) multiply mini-interviews; (5) standardized certified intelligence assessments; (6) assessments of (ecological) conscientiousness.

Inclusion of each of these selection methods should be guided by understanding which dimensions of study success these selection methods are capable of predicting, whether a selection method is accepted (and to what extent) by admissions committees and applicants, and whether the admissions committees are aware of the correct usage of a selection method.

Future directions

The methodological approach toward researching selective admissions

In most of the primary studies reviewed, the regression approach was used. While it is a widely accepted type of analysis in this field, it is limited, because the findings on amount of explained variance are usually hard to interpret. Moreover, the findings based on the regression approach do not allow one to set the cutoff scores. Future research would benefit from applying other methodologies. For example, Bridgeman et al. (2009) offer a method that divides students within a department into quartiles based on a selection method of interest and a dimension of study success. The methodology that allows (under certain conditions) the establishment of cutoff scores for selective admissions methods is the Signal Detection Theory (van Ooijen-van der Linden, 2017). Finally, future research approaches toward selection methods should account for a multilevel and dynamic nature of student selection (Patterson et al., 2018) as well as the importance of other evaluative quality principles of selection methods not addressed in this review, such as practicality/administrative convenience, ease of interpretation, and so forth (see for the full list Patterson and Ferguson, 2010).

Future directions in practice of selective admissions

Research evidence on selection methods has advanced significantly in recent years. In some national and institutional contexts, the research findings are actively being translated into practice (e.g., Council of Graduate Schools, 2021). However, along with that, "today's faculty choose students on the basis of an array of perceptions that only sometimes have a strong evidentiary basis" (Posselt, 2016, p. 176). Therefore, professionalization of admissions staff and formation of communities of good admissions practices are required. Even despite certain gaps in research, already existing evidence allows significant progress toward the evidence-based policy on selective admissions for graduate schools across the world.

In addition to professionalization of admissions staff, it is important to consider monitoring and evaluation of the admissions process: Is there a closed-loop control of the admissions process? Are the selection methods scrutinized adequately in accreditation? Is there sufficient reporting on the chosen admissions process and selection methods applied in the HEI to higher levels? Ultimately, the answers to these questions reflect the extent of accountability of admissions committees for the soundness of their admissions practices. Accountability would imply reporting on data on each selection round to higher levels within HEI's organization. Institutional research, in turn, could have a role in analyzing emerging patterns, testing these against relevant models, and giving warning signals when substantial deviations occur. This would contribute to an adaptive admissions process that could eventually lead to fairer and more objective graduate admissions (Zimmermann et al., 2017a, 2017b).

Selective admissions and societal responsibility

Considering increasing numbers of applications and capacity limitations at research universities, evidence-based student selection is increasingly recognized as a socially significant practice which should diminish rather than enhance inequality. Failing to meet requirements of fairness, objectiveness, and transparency primarily leads to missed opportunities for capable students and a HEI, the inability of a HEI to justify the selection decisions, jeopardizing the diversity of the student body, infringement of students' rights on equal access to higher education, and the loss of time and efforts both by students and institutions. In extreme cases, abandoning quality requirements toward selective admissions process might lead to appearances of criminal bribing schemes (e.g., the 2019 college admissions bribery scandal in the US). Designing a sound admissions process for graduate level education is, therefore, a necessary step for preventing these issues from arising or to cease their existence entirely. Finally, student selection has become an increasingly politicized societal topic, where advocacy groups and politicians are actively participating. In some countries, the alternatives to selective admissions are discussed, such as re-introducing the (weighted) lottery system in the Netherlands as a more neutral solution (The national government of the Netherlands, 2021). However, there is some critique of its effect on equal access, because a weighted lottery is based on selection criteria as well (Council of State of the Netherlands, 2021).

Limitations

Drawing conclusions from a large number of papers inevitably brings a risk of losing the nuances of each study (see Additional file 1: Table S3 for more details). It also means that the samples of studies on predictive validity of graduate selection methods in several instances included not only STEM students but also students from other disciplines. Even if the strength of the relationship between a selection method and various dimensions of graduate study success is diluted by inclusion of students from other disciplines, it is unlikely that the direction of relationship would be the opposite. From this, however, an advantage appeared that the findings of this review to a certain extent are generalizable to other academic disciplines within graduate levels of education.

Another limitation is that our inferences on the effects sizes (negligible, small, medium, and strong effect sizes) were based on the interpretations of the studies' authors. To refine the estimations of the effect sizes, the meta-analyses on reviewed selection methods would be required. Such goals were outside the scope of this review; however, the indications that this review provides are robust enough to answer the main question on whether a selection method is valid in principle.

Furthermore, most studies on the topic were carried out in the US, which has inevitably influenced this review. Therefore, practitioners and policymakers outside the US should account for this unintentional bias when referring to the results and conclusions of this review. However, we think that the cultural/geographical bias may have mainly impacted the results and conclusions related to acceptability of selection methods as it addresses individuals' perceptions, which are more easily affected by culture. On the other hand, we think that (a) validity and (b) procedural issues of selection methods are much less affected by cultural/geographical bias, because these evaluative quality principles relate to (a) the predictive power toward uniformed dimensions of study success and (b) concerns involved in using certain selection methods. For example, a common concern regarding richer applicants having more financial possibilities than poorer applicants to be coached on standardized testing is relevant in any country.

Finally, the reviewed literature on acceptability of selection methods often contained evidence from admissions committees' self-reports. Their reports could have been (un)consciously biased to a certain extent if they did not want to report, for example, the usage of invalid yet favored selection methods. Therefore, the observational ethnographic studies, like the one of Posselt (2016), gain special importance in this area of research: The observation might be a more appropriate method to detect "hidden" selection criteria and group dynamics within an admissions committee, because these concealed processes are influential toward admissions decisions.

Conclusion

The main aim of this review was to collect, map, synthesize, and critically analyze the available research evidence on graduate selection methods with a focus on STEM disciplines. The results of the systematic search of research literature were categorized according to a type of selection method and core evaluative quality principles (predictive validity, acceptability, and procedural issues). Ten categories of graduate selection methods emerged. It was found that the predictive validity of prior grades, GRE General, intelligence assessments, and conscientiousness toward several study success dimensions is of medium-to-strong extent. Letters of recommendation, tests on language proficiency, emotional intelligence, and need for cognition are valid as well, but of weak-to-medium extent. Based on the limited evidence, it also appears that prior research experience, multiple mini-interviews, and selectivity of prior institution might have significant relationships with certain dimensions of graduate study success. Personal statements, traditional interviews, and personal traits such as extroversion and neuroticism are invalid predictors of graduate study success.

When choosing the selection methods to be applied in the admissions process, policy makers and admissions committees should use only valid instruments. They should also be aware of typical applicant reactions toward these methods as well as procedural issues, such as possible adverse effects toward certain groups, susceptibility for biases, faking, coaching, and stereotype threat. The admissions committees are advised (1) to completely exclude invalid selection instruments from their admissions requirements, (2) to define the dimensions of study success that are most important for their program, (3) to use those selection methods that showed predictive validity toward these predefined study success dimensions, accounting for applicant reactions and procedural issues of each of those methods, and (4) to ensure the accountability of the admissions process by reporting on data on each selection round to higher levels within HEI's organization, which should in turn conduct further analysis and regular evaluations of admissions processes.

Abbreviations

CEEPT	The Computerized Enhanced ESL Placement Test
CV	Curriculum Vitae
ERIC	Education Resources Information Center
EXANI	Examen Nacional de Ingreso al Posgrado
GPA	Grade Point Average
GRE General	Graduate Record Examinations General Test
GRE-Q	The Quantitative Reasoning measure of the GRE General Test
GRE-V	The Verbal Reasoning measure of the GRE General Test
GRE-A	The Analytical Writing measure of the GRE General Test
HEI	Higher Education Institution
IELTS	International English Language Testing System
OECD	Organization for Economic Co-operation and Development
PhD	A Doctor of Philosophy
LoR	Letters of Recommendation
MMI	Multiple Mini-Interview
STEM	Science, Technology, Engineering, and Math
TOEFL	Test of English as a Foreign Language
UGPA	Undergraduate Grade Point Average

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40594-023-00445-4.

Additional file 1: Table S1. Key words used in the search in literature data bases. **Table S2.** Number of articles relating to each selection method and evaluative quality principle under consideration. **Table S3.** Summary of the relevant findings for each selection method.

Acknowledgements
The authors would like to thank Prof. Dr. Marijk van der Wende for her supervision of the overarching PhD project focused on graduate selective admissions, where this study was part of. The authors are grateful to Adam Frick and Dr. Christine Merie Fox for copy editing.

Author contributions

AK, HvR, and GD developed the original concept and defined the inclusion criteria. AK outlined, conducted the systematic search of literature, collected and coded the articles, wrote the original draft, and conducted a revision. AK and SC worked together on the coding framework. SC acted as the second coder and helped with the calculation of inter-rater agreement. All authors worked together to revise the article prior to submission. All authors read and approved the final manuscript.

Funding
This study received internal institutional funding.

Availability of data and materials

The set of articles used during the current review are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 10 January 2023 Accepted: 2 August 2023
Published online: 07 September 2023

References

Studies that met inclusion criteria and were included in the "Results" section of this review are marked with an asterisk.

*Álvarez-Montero, F., Mojardin-Heraldez, A., & Audelo-Lopez, C. (2014). Criteria and instruments for doctoral program admissions. *Electronic Journal of Research in Educational Psychology, 12*(3), 853–866. https://www.researchgate.net/publication/266200011_Criteria_and_instruments_for_doctoral_program_admission

Anderson, N., Salgado, J. F., & Hülsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment, 18*(3), 291–304. https://doi.org/10.1111/j.1468-2389.2010.00512.x

*Attali, Y., & Sinharay, S. (2015). Automated trait scores for GRE ® writing tasks: Automated trait scores for GRE ® writing tasks. (Report No. RR-15-15). *Educational Testing Service*. https://doi.org/10.1002/ets2.12062

*Biernat, M., & Eidelman, S. (2007). Translating subjective language in letters of recommendation: The case of the sexist professor. *European Journal of Social Psychology, 37*(6), 1149–1175. https://doi.org/10.1002/ejsp.432

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures: Job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

*Bleske-Rechek, A., & Browne, K. (2014). Trends in GRE scores and graduate enrollments by gender and ethnicity. *Intelligence, 46*, 25–34. https://doi.org/10.1016/j.intell.2014.05.005

*Boyette-Davis, J. (2018). A data-based assessment of research-doctorate programs in the United States. *The Journal of Undergraduate Neuroscience Education, 17*(1), A54–A58. https://doi.org/10.17226/12994

*Bridgeman, B., Burton, N., & Cline, F. (2009). A note on presenting what predictive validity numbers mean. *Applied Measurement in Education, 22*(2), 109–119. https://doi.org/10.1080/08957340902754577

*Briihl, D. S., & Wasieleski, D. T. (2007). The GRE analytical writing test: Description and utilization. *Teaching of Psychology, 34*(3), 191–193. https://doi.org/10.1080/00986280701498632

Burmeister, J. M., Kiefner, A. E., Carels, R. A., & Musher-Eizenman, D. R. (2013). Weight bias in graduate school admissions. *Obesity, 21*(5), 918–920. https://doi.org/10.1002/oby.20171

*Burmeister, J., McSpadden, E., Rakowski, J., Nalichowski, A., Yudelev, M., & Snyder, M. (2014). Correlation of admissions statistics to graduate student success in medical physics. *Journal of Applied Clinical Medical Physics, 15*(1), 375–385. https://doi.org/10.1120/jacmp.v15i1.4451

Burton, N. W., & Wang, M. (2005). Predicting long-term success in graduate school: A collaborative validity study. (Report No. 99-14R. ETS RR-05-03). Educational Testing Service. http://grad.uga.edu/wpcontent/uploads/2017/09/GRE_Research_Report.pdf

*Butter, R., & Born, MPh. (2012). Enhancing criterion-related validity through bottom-up contextualization of personality inventories: The construction of an ecological conscientiousness scale for PhD candidates. *Human Performance, 25*(4), 303–317. https://doi.org/10.1080/08959285.2012.703730

*Camara, W., Packman S., & Wiley A. (2013). College, graduate, and professional school admissions testing. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R., Kuncel, S. P., Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Testing and assessment in school psychology and education* (Vol 3, pp. 297–318). American Psychological Association. https://doi.org/10.1037/14049-014

*Chari, D., & Potvin, G. (2019). Admissions practices in terminal master's degree-granting physics departments: A comparative analysis. *Physical Review Physics Education Research, 15*(1), Article 010104. https://doi.org/10.1103/PhysRevPhysEducRes.15.010104

*Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing, 29*(3), 421–442. https://doi.org/10.1177/0265532211430368

*Cline, F., & Powers, D. (2014). Test-taker perceptions of the role of the GRE® General test in graduate admissions: Preliminary findings. In *The Research Foundation for the GRE revised general test: A compendium of studies* (p. 6.1.1–6.1.6). Educational Testing Service. https://www.ets.org/s/research/pdf/gre_compendium.pdf

Council of Graduate Schools. (2021). *CGS best practices programs in graduate admissions and enrollment management*. https://cgsnet.org/admissions-and-recruitment

Council of State of the Netherlands [Raad van State]. (2021). Amendment of the Higher Education and Scientific Research Act in relationship to the addition of decentralized draw as a selection method for higher education programs with fixed capacity [Wijziging van de wet op het hoger onderwijs en wetenschappelijk onderzoek in verband met het toevoegen van decentrale loting als selectiemethode voor opleidingen met capaciteitsfixus in het hoger onderwijs]. (W05.20.0508/I). https://www.raadvanstate.nl/@123920/w05-20-0508/

*Cox, G. W., Hughes, W. E., Jr., Etzkorn, L. H., & Weisskopf, M. E. (2009). Predicting computer science PhD completion: A case study. *IEEE Transactions on Education, 52*(1), 137–143. https://doi.org/10.1109/TE.2008.921458

*Davey, T., & Lee, Y.-H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised General test*. (Report No. GREB-08–01). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02262.x

De Boer, T., & Van Rijnsoever, F. (2022a). In search of valid non-cognitive student selection criteria. *Assessment & Evaluation in Higher Education, 47*(5), 783–800. https://doi.org/10.1080/02602938.2021.1958142

*De Boer, T., & Van Rijnsoever, F. J. (2022b). One field too far? Higher cognitive relatedness between bachelor and master leads to better predictive validity of bachelor grades during admission. Assessment & Evaluation in Higher Education. https://doi.org/10.1080/02602938.2022.2158453

DeClou, L. (2016). Who stays and for how long: examining attrition in Canadian graduate programs. *Canadian Journal of Higher Education, 46*(4), 174–198.

De Wit, H., & Altbach, P. G. (2020). Internationalization in higher education: global trends and recommendations for its future. *Policy Reviews in Higher Education, 5*(1), 28–46. https://doi.org/10.1080/23322969.2020.1820898

European Grade Conversion System. (2020). *Grade conversion–an introduction*. http://egracons.eu/page/about-egracons-project-and-tool

*Fischer, F. T., Schult, J., & Hell, B. (2013). Sex-specific differential prediction of college admissions tests: A meta-analysis. *Journal of Educational Psychology, 105*(2), 478–488. https://doi.org/10.1037/a0031956

Garaz, S., & Torotcoi, S. (2017). Increasing access to higher education and the reproduction of social inequalities: The case of Roma university students in Eastern and Southeastern Europe. *European Education, 49*(1), 10–35. https://doi.org/10.1080/10564934.2017.1280334

*Garces, L. M. (2014). Aligning diversity, quality, and equity: The implications of legal and public policy developments for promoting racial diversity in graduate studies. *American Journal of Education, 120*(4), 457–480. https://doi.org/10.1086/676909

*Gilmore, J., Vieyra, M., Timmerman, B., Feldon, D., & Maher, M. (2015). The relationship between undergraduate research participation and subsequent research performance of early career STEM graduate students. *The Journal of Higher Education, 86*(6), 834–863. https://doi.org/10.1353/jhe.2015.0031

*Ginther, A., & Elder, C. (2014). A comparative investigation into understandings and uses of the TOEFL iBT® test, the international English language testing service (academic) test, and the Pearson test of English for graduate admissions in the United States and Australia: A case study: An investigation into test score understandings and uses. (Report No. RR– 14-44). *Educational Testing Service*. https://doi.org/10.1002/ets2.12037

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*(1), 26–34. https://doi.org/10.1037/0003-066X.48.1.26

*Hall, J. D., O'Connell, A. B., & Cook, J. G. (2017). Predictors of student productivity in biomedical graduate school applications. *PLoS ONE, 12*(1), e0169121. https://doi.org/10.1371/journal.pone.0169121

*Hausknecht, J. P., Halpert, J. A., Paolo, N. T. D., & Gerrard, M. O. M. (2007). Retesting in selection: A meta-analysis of practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373–385.

Howell, L. L., Sorenson, C. D., & Jones, M. R. (2014). Are undergraduate GPA and general GRE percentiles valid predictors of student performance in an engineering graduate program? *International Journal of Engineering Education, 30*(5), 1145–1165. https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=2342&context=facpub

Jayakumar, U. M., & Page, S. E. (2021). Cultural capital and opportunities for exceptionalism: Bias in university admissions. *The Journal of Higher Education, 92*(7), 1109–1139. https://doi.org/10.1080/00221546.2021.1912554

Kerrin, M., Mossop, L., Morley, E., Fleming, G., & Flaxman, C. (2018). Role analysis: The foundation for selection systems. In F. Patterson & L. Zibarras (Eds.), *Selection and recruitment in the healthcare professions: Research, theory and practice* (pp. 139–165). Palgrave Macmillan.

Kirby, W., & van der Wende, M. (2019). The New Silk Road: Implications for higher education in China and the West? *Cambridge Journal of Regions, Economy and Society, 12*(1), 127–144. https://doi.org/10.1093/cjres/rsy034

*Klieger, D. M., Cline, F. A., Holtzman, S. L., Minsky, J. L., & Lorenz, F. (2014). New perspectives on the validity of the GRE ® General test for predicting graduate school grades: New perspectives for predicting graduate school grades (Report No. RR– 14-26). *Educational Testing Service*. https://doi.org/10.1002/ets2.12026

König, C. J., Steiner Thommen, L. A., Wittwer, A.-M., & Kleinmann, M. (2017). Are observer ratings of applicants' personality also faked? Yes, but less than self-reports. *International Journal of Selection and Assessment, 25*(2), 183–192. https://doi.org/10.1111/ijsa.12171

*Kuncel, N. R., & Hezlett, S. A. (2007a). The utility of standardized tests: Response. *Science, 316*, 1696–1697. https://doi.org/10.1126/science.316.5832.1694b

*Kuncel, N. R., & Hezlett, S. A. (2007b). Standardized tests predict graduate students' success. *Science, 315*(5815), 1080–1081. https://doi.org/10.1126/science.1136618

*Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science, 19*(6), 339–345. https://doi.org/10.1177/0963721410389459

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*(1), 148–161. https://doi.org/10.1037/0022-3514.86.1.148

*Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment, 22*(1), 101–107. https://doi.org/10.1111/ijsa.12060

*Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the Graduate Record Examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement, 70*(2), 340–352. https://doi.org/10.1177/0013164409344508

Kuncel, N., Tran, K., & Zhang, S. H. (2020). Measuring student character: Modernizing predictors of academic success. In M. E. Oliveri & C. Wendler (Eds.), *Higher education admissions practices: An international perspective* (pp. 276–302). Cambridge University Press.

*Kurysheva, A., Koning, N., Fox, C. M., van Rijen, H. V., & Dilaver, G. (2022). Once the best student always the best student? Predicting graduate study success, using undergraduate academic indicators. Evidence from research masters' programs in the Netherlands. *International Journal of Selection and Assessment, 30*(4), 1–17. https://doi.org/10.1111/ijsa.12397

*Kurysheva, A., van Ooijen-van der Linden, L., van der Smagt, M. J., & Dilaver, G. (2022). The added value of signal detection theory as a method in evidence-informed decision-making in higher education: A demonstration. *Frontiers in Education, 7*, Article 906611. https://doi.org/10.3389/feduc.2022.906611

*Kurysheva, A., van Rijen, H. V., & Dilaver, G. (2019). How do admission committees select? Do applicants know how they select? Selection criteria and transparency at a Dutch University. *Tertiary Education and Management, 25*, 367–388. https://doi.org/10.1007/s11233-019-09050-z

*Kyllonen, P., Walters, A. M., & Kaufman, J. C. (2005). Noncognitive constructs and their assessment in graduate education: A review. *Educational Assessment, 10*(3), 153–184. https://doi.org/10.1207/s15326977ea1003_2

Kyllonen, P. C., Walters, A. M., & Kaufman, J. C. (2011). The role of noncognitive constructs and other background variables in graduate education. (Report No. GREB-00-11). *Educational Testing Service*. https://doi.org/10.1002/j.2333-8504.2011.tb02248.x

*Lee, Y.-J., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research, 1*(4), 366–389. https://doi.org/10.1177/1558689807306148

Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: Development and validation of an interview faking behavior scale. *Journal of Applied Psychology, 92*(6), 1638–1656. https://doi.org/10.1037/0021-9010.92.6.1638

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology, 67*(1), 241–293. https://doi.org/10.1111/peps.12052

*Lorden, J. F., Ed, Kuh, C. V., Ed, & Voytuk, J. A., Ed. (2011). Research-doctorate programs in the biomedical sciences: Selected findings from the NRC assessment. The National Academies Collection: Reports funded by National Institutes of Health. https://doi.org/10.17226/13213

*Lott, J. L. I., Gardner, S., & Powers, D. A. (2009). Doctoral student attrition in the STEM fields: An exploratory event history analysis. *Journal of College Student Retention: Research, Theory and Practice, 11*(2), 247–266. https://doi.org/10.2190/CS.11.2.e

*MasterMind Europe. (2017). *Admissions to English-taught programs (ETPs) at master's level in Europe–Procedures, regulations, success rates and challenges for diverse applicants*. ACA, StudyPortals, and Vrije Universiteit Amsterdam. http://mastermindeurope.eu/wp-content/uploads/2017/01/Report-2-Admissions-to-ETPs.pdf

*Mathews, J. (2007). Predicting international students' academic success… may not always be enough: Assessing Turkey's foreign study scholarship program. *Higher Education, 53*(5), 645–673. https://doi.org/10.1007/s10734-005-2290-x

*Megginson, L. (2009). Noncognitive constructs in graduate admissions: An integrative review of available instruments. *Nurse Educator, 34*(6), 254–261. https://doi.org/10.1097/NNE.0b013e3181bc7465

*Mendoza-Sanchez, I., deGruyter, J. N., Savage, N. T., & Polymenis, M. (2022). Undergraduate GPA predicts biochemistry PhD completion and is associated with time to degree. *CBE—Life Sciences Education, 21*(2), ar19. https://doi.org/10.1187/cbe.21-07-0189

Merriam-Webster dictionary. (n.d.). Grade inflation. In *Merriam-Webster.com dictionary*. Retrieved February 18, 2022, from https://www.merriam-webster.com/dictionary/grade%20inflation

Miettinen, R. (2004). The roles of the researcher in developmentally-oriented research. In T. Kontinen (Ed.), *Development intervention. Actor and activity perspectives* (pp. 105–121). University of Helsinki, Center for Activity Theory and Developmental Work Research and Institute for Development Studies.

*Miller, E. M. (2019). Promoting student success in statistics courses by tapping diverse cognitive abilities. *Teaching of Psychology, 46*(2), 140–145. https://doi.org/10.1177/0098628319834198

*Miller, A., Crede, M., & Sotola, L. K. (2021). Should research experience be used for selection into graduate school: A discussion and meta-analytic synthesis of the available evidence. *International Journal of Selection and Assessment, 29*(1), 19–28. https://doi.org/10.1111/ijsa.12312

*Moneta-Koehler, L., Brown, A. M., Petrie, K. A., Evans, B. J., & Chalkley, R. (2017). The Limitations of the GRE in predicting success in biomedical graduate school. *PLoS ONE, 12*(1), Article e0166742. https://doi.org/10.1371/journal.pone.0166742

*Morgan, W. B., Elder, K. B., & King, E. B. (2013). The emergence and reduction of bias in letters of recommendation: Bias in letters of recommendation. *Journal of Applied Social Psychology, 43*(11), 2297–2306. https://doi.org/10.1111/jasp.12179

*Mupinga, E. E., & Mupinga, D. M. (2005). Perceptions of international students toward GRE. *College Student Journal, 39*(2), 402–409.

*Murphy, K. R. (2009). How a broader definition of the criterion domain changes our thinking about adverse impact. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 137–160). Routledge. https://doi.org/10.4324/9780203848418

Murphy, S. C., Klieger, D. M., Borneman, M. J., & Kuncel, N. R. (2009). The predictive power of personal statements in admissions: A meta-analysis and cautionary tale. *College & University, 84*(4), 83–88.

Okahana, H., & Zhou, E. (2018). *International graduate applications and enrollment: Fall 2017* (pp. 1–24). Washington, DC: Council of Graduate Schools. https://cgsnet.org/Data-Insights/

*oude Egbrink, M. G. A., & Schuwirth, L. W. T. (2016). Narrative information obtained during student selection predicts problematic study behavior. *Medical Teacher, 38*(8), 844–849. https://doi.org/10.3109/0142159X.2015.1132410

*Park, H.-Y., Berkowitz, O., Symes, K., & Dasgupta, S. (2018). The art and science of selecting graduate students in the biomedical sciences: Performance in doctoral study of the foundational sciences. *PLoS ONE, 13*(4), Article e0193901. https://doi.org/10.1371/journal.pone.0193901

Patterson, F., & Ferguson, E. (2010). Selection for medical education and training. In T. Swanwick (Ed.), *Understanding medical education* (pp. 352–365). Wiley-Blackwell. https://doi.org/10.1002/9781444320282.ch24

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? *A Systematic Review. Medical Education, 50*(1), 36–60. https://doi.org/10.1111/medu.12817

Patterson, F., Roberts, C., Hanson, M. D., Hampe, W., Eva, K., Ponnamperuma, G., Magzoub, M., Tekian, A., & Cleland, J. (2018). 2018 Ottawa consensus statement: Selection and recruitment to the healthcare professions. *Medical Teacher, 40*(11), 1–9. https://doi.org/10.1080/0142159X.2018.1498589

Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). Effects of applicant faking on forced-choice and Likert scores. *Organizational Research Methods, 22*(3), 710–739. https://doi.org/10.1177/1094428117753683

*Payne, D. (2015). A common yardstick for graduate education in Europe. *Journal of the European Higher Education Area, 2*, 21–48.

*Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338. https://doi.org/10.1037/a0014996

*Posselt, J. R. (2014). Toward inclusive excellence in graduate education: Constructing merit and diversity in PhD admissions. *American Journal of Education, 120*(4), 481–514. https://doi.org/10.1086/676910

Posselt, J. R. (2016). *Inside graduate admissions: Merit, diversity, and faculty gatekeeping*. Harvard University Press.

*Posselt, J. R. (2018). Trust Networks: A new perspective on pedigree and the ambiguities of admissions. *The Review of Higher Education, 41*(4), 497–521. https://doi.org/10.1353/rhe.2018.0023

*Powers, D. E. (2005). Effects of preexamination disclosure of essay prompts for the GRE analytical writing assessment. (Report No. 01-07R). *Educational Testing Service*. https://doi.org/10.1002/j.2333-8504.2005.tb01978.x

*Powers, D. E. (2017). Understanding the impact of special preparation for admissions tests. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment. The methodological, psychological and policy contributions of ETS* (pp. 553–564). Springer Open. https://doi.org/10.1007/978-3-319-58689-2

Proudfoot, S., & Hoffer, T. B. (2016). Science and engineering labor force in the US. In L. Gokhberg, N. Shmatko, & L. Auriol (Eds.), *The science and technology labor force* (pp. 77–120). Springer International Publishing. https://doi.org/10.1007/978-3-319-27210-8

Reumer, C., & van der Wende, M. (2010). Excellence and diversity: The emergence of selective admissions policies in Dutch higher education. A case study on Amsterdam University College. *Research & Occasional Paper Series: CSHE.15.10*, 1–28. https://cshe.berkeley.edu/publications/excellence-and-diversity-emergence-selective-admission-policies-dutch-higher-0

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

Salmi, J., & Bassett, R. M. (2014). The equity imperative in tertiary education: Promoting fairness and efficiency. *International Review of Education, 60*(3), 361–377. https://doi.org/10.1007/s11159-013-9391-z

*Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin, 143*(6), 565–600. https://doi.org/10.1037/bul0000098

*Sealy, L., Saunders, C., Blume, J., & Chalkley, R. (2019). The GRE over the entire range of scores lacks predictive ability for PhD outcomes in the biomedical sciences. *PloS One, 14*(3), e0201634. https://doi.org/10.1371/journal.pone.0201634

Sedlacek, W. E. (2003). Alternative admissions and scholarship selection measures in higher education. *Measurement and Evaluation in Counseling and Development, 35*(4), 263–272. https://doi.org/10.1080/07481756.2003.12069072

The Bologna Declaration. Joint declaration of the European Ministers of Education, June 19, 1999. http://www.ehea.info/page-ministerial-conference-bologna-1999

The national government of the Netherlands [Rijksoverheid]. (2021). *Loten voor studie zorgt voor kansengelijkheid* [*Lots for study ensures equality of opportunity*]. https://www.rijksoverheid.nl/actueel/nieuws/2021/03/12/loten-voor-studie-zorgt-voor-kansengelijkheid

*Trapmann, S., Hell, B., Hirn, J.-O.W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Zeitschrift Für Psychologie, 215*(2), 132–151. https://doi.org/10.1027/0044-3409.215.2.132

van Ooijen-van der Lindenvan der Smagt, L. M. J., Woertman, L., & te Pas, S. F. (2017). Signal detection theory as a tool for successful student selection. *Assessment & Evaluation in Higher Education, 42*(8), 1193–1207. https://doi.org/10.1080/02602938.2016.1241860

*van Os, W. (2007). Selection to the master's phase at the binary divide, a Dutch case study. *Tertiary Education and Management, 13*(2), 127–140. https://doi.org/10.1080/13583880701238365

*Verostek, M., Miller, C. W., & Zwickl, B. (2021). Analyzing admissions metrics as predictors of graduate GPA and whether graduate GPA mediates Ph. D. completion. *Physical Review Physics Education Research, 17*(2), 020115. https://doi.org/10.1103/PhysRevPhysEducRes.17.020115

Vrielink, J., Lemmens, P., & Parmentier, S. (2011). Academic freedom as a fundamental right. *Procedia-Social and Behavioral Sciences, 13*, 117–141. https://doi.org/10.1016/j.sbspro.2011.03.009

*Walsh, M. J. (2020). Online doctoral student grade point average, conscientiousness, and grit: A moderation analysis. *Journal of Educators Online, 17*(1). Advance online publication. https://www.thejeo.com/

Weedon, E. (2017). The construction of under-representation in UK and Swedish higher education: Implications for disabled students. *Education, Citizenship and Social Justice, 12*(1), 75–88. https://doi.org/10.1177/1746197916683470

*Weiner, O. D. (2014). How should we be selecting our graduate students? *Molecular Biology of the Cell, 25*(4), 429–430. https://doi.org/10.1091/mbc.e13-11-0646

*Weissman, M. B. (2020). Do GRE scores help predict getting a physics Ph.D.? A comment on a paper by Miller et al. *Science Advances, 6*(23), Article eaax3787. https://doi.org/10.1126/sciadv.aax3787

*Westrick, P. A. (2017). Reliability estimates for undergraduate grade point average. *Educational Assessment, 22*(4), 231–252. https://doi.org/10.1080/10627197.2017.1381554

*Willcockson, I. U., Johnson, C. W., Hersh, W., & Bernstam, E. V. (2009). Predictors of student success in graduate biomedical informatics training: Introductory course and program success. *Journal of the American Medical Informatics Association, 16*(6), 837–846. https://doi.org/10.1197/jamia.M2895

*Wilson, M. A., DePass, A. L., & Bean, A. J. (2018). Institutional interventions that remove barriers to recruit and retain diverse biomedical PhD students. *CBE—Life Sciences Education, 17*(2), Article 17:ar27. https://doi.org/10.1187/cbe.17-09-0210

*Wilson, M. A., Odem, M. A., Walters, T., DePass, A. L., & Bean, A. J. (2019). A model for holistic review in graduate admissions that decouples the GRE from race, ethnicity, and gender. *CBE—Life Sciences Education, 18*(1), Article 18: ae7. https://doi.org/10.1187/cbe.18-06-0103

*Wollast, R., Boudrenghien, G., Van der Linden, N., Galand, B., Roland, N., Devos, C., De Clercq, M., Klein, O., Azzi, A., & Frenay, M. (2018). Who are the doctoral students who drop out? Factors associated with the rate of doctoral degree completion in universities. *International Journal of Higher Education, 7*(4), 143–156. https://doi.org/10.5430/ijhe.v7n4p143

*Woo, S. E., LeBreton, J. M., Keith, M. G., & Tay, L. (2023). Bias, fairness, and validity in graduate-school admissions: A psychometric perspective. *Perspectives on Psychological Science, 18*(1), 3–31. https://doi.org/10.1177/17456916211055374

Wouters, A., Bakker, A.H., van Wijk, I.J. et al. (2014). A qualitative analysis of statements on motivation of applicants for medical school. *BMC Medical Education, 14*, 200. https://doi.org/10.1186/1472-6920-14-200

*Young, N. T., Tollefson, K., & Caballero, M. D. (2023). Making graduate admissions in physics more equitable. *Physics Today, 76*(7), 40–45. https://doi.org/10.1063/PT.3.5271

Zimdars, A. M. (2016). *Meritocracy and the university: Selective admission in England and the United States*. Bloomsbury Publishing.

*Zimmermann, J., Heinimann, H. R., Brodersen, K. H., & Buhmann, J. M. (2015). A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. *Journal of Educational Data Mining, 7*(3), 151–176. https://doi.org/10.5281/zenodo.3554733

*Zimmermann, J., von Davier, A. A., Buhmann, J. M., & Heinimann, H. R. (2017a). Validity of GRE General test scores and TOEFL scores for graduate admissions to a technical university in Western Europe. *European Journal of Engineering Education, 43*(1), 144–165. https://doi.org/10.1080/03043797.2017.1343277

Zimmermann, J., von Davier, A., & Heinimann, H. R. (2017b). Adaptive admissions process for effective and fair graduate admissions. *International Journal of Educational Management, 31*(4), 540–558. https://doi.org/10.1108/IJEM-06-2015-0080

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.