

**Rubric-based holistic review: A promising route to equitable graduate admissions in physics**

Nicholas T. Young<sup>1,2,\*</sup> K. Tollefson<sup>1</sup>  
 Remco G. T. Zegers<sup>1,3,4</sup> and Marcos D. Caballero<sup>1,2,5,6,†</sup>

<sup>1</sup>*Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA*


<sup>2</sup>*Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, Michigan 48824, USA*

<sup>3</sup>*National Superconducting Cyclotron Laboratory, Michigan State University, East Lansing, Michigan 48824, USA*

<sup>4</sup>*Joint Institute for Nuclear Astrophysics, Michigan State University, East Lansing, Michigan 48824, USA*

<sup>5</sup>*Center for Computing in Science Education and Department of Physics, University of Oslo, N-0316 Oslo, Norway*

<sup>6</sup>*CREATE for STEM Institute, Michigan State University, East Lansing, Michigan 48824, USA*

 (Received 8 October 2021; revised 7 September 2022; accepted 20 October 2022; published 30 November 2022)

As systematic inequities in higher education and society have been brought to the forefront, graduate programs are interested in increasing the diversity of their applicants and enrollees. Yet, structures in place to evaluate applicants may not support such aims. One potential solution to support those aims is rubric-based holistic review. Starting in 2018, our physics department implemented a rubric-based holistic review process for all applicants to our graduate program. The rubric assessed applicants on 18 metrics covering their grades, test scores, research experiences, noncognitive competencies, and fit with the program. We then compared faculty's ratings of applicants by admission status, sex, and undergraduate program over a three-year period. We find that the rubric scores show statistically significant differences between admitted and nonadmitted students as hoped. We also find that differences in rubric scores based on sex or undergraduate program reflected known systematic inequities such as applicants from smaller and less prestigious undergraduate universities scoring lower on the physics GRE and women performing more volunteer work in academia. Our results then suggest rubric-based holistic review as a possible route to making graduate admissions in physics more equitable.

DOI: [10.1103/PhysRevPhysEducRes.18.020140](https://doi.org/10.1103/PhysRevPhysEducRes.18.020140)

**I. INTRODUCTION**

Female and Black, Latinx, and Indigenous scholars have been and are underrepresented at all levels of physics. The percentage of physics degrees awarded to women has stagnated at around 20% [1] while the percentage of physics degrees awarded to Black, Latinx, and Indigenous students has remained less than 10% despite these students making up a larger portion of the college population than in the past [2]. While there are numerous possibilities to address the systematic inequities these scholars face at all levels of academia that limit their participation [3–9], this paper will focus on graduate

admissions in physics. Specifically, if we treat graduate admissions as a four stage process similar to how O'Meara *et al.* treats faculty hiring as a four-stage process [10] consisting of framing the position and forming a committee, marketing, outreach, and recruitment, evaluating candidates, and making short lists and final decisions, the latter two fall within the scope of this paper.

While physics departments may be interested in increasing their diversity, the dominant processes of evaluating applicants for graduate school do not support such aims. Prior work has found that diversity considerations are often secondary when evaluating applicants and are discussed after many diverse candidates have already been cut from the applicant pool [11,12]. Therefore, increasing diversity and equity during the admissions process requires rethinking the process physics departments use to evaluate applicants.

One promising approach to rethinking the admissions process is holistic review, where a broad range of candidate qualities are considered [13]. In physics, the use of rubric-based review to facilitate such holistic reviews has been gaining traction through the Inclusive Graduate Education

\*ntyoung@umich.edu

†Corresponding author.  
caball14@msu.edu

*Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.*

Network [14]. Under their approach, applicants are rated according to a predefined rubric on both traditional metrics such as GPA and test scores as well as noncognitive skills such as showing instances of initiative and perseverance in their essays and recommendation letters. The use of a predefined rubric is claimed to ensure that each applicant is treated fairly and biases by reviewers are checked [15–17], and hence, it could make graduate admissions more equitable.

To our knowledge, however, few studies have examined how these rubrics work in practice and whether they fulfill such aims of equitable admissions. Therefore, the goal of this paper is to empirically examine claims of equitable admissions in the context of our department’s graduate program and its rubric. Based on the data we have access to, our paper addresses three questions related to rubric-based review:

1. How do faculty assign rubric scores to applicants and how do those differ between admitted and rejected applicants?
2. How do the scores assigned by faculty differ by applicant’s sex?
3. How do the scores assigned by faculty differ by the type of institution the applicant attended?

As Scherr *et al.* concluded in their study of graduate admissions practices in physics, many departments are unaware of what other departments do and hence, they might be willing to change their practices if they become aware of successful practices in use elsewhere [18]. Therefore, a secondary goal of this paper is to describe alternative admissions practices in physics and how departments may apply these alternative practices to their own admissions processes.

The rest of the paper is organized as follows. In Sec. II, we provide an overview of holistic review, rubric-based review, and evidence from other fields about their potential for success. In Sec. III, we describe how our department transitioned to rubric-based review, how we collected data relevant to evaluating our admissions process, and how we analyzed such data. In Sec. IV, we share results that suggest our rubric does support more equitable admissions practices and in Sec. V, we contextualize our results, answer our research questions, and examine how our choices as researchers may have affected the results. In Secs. VI and VII we examine the limitations of this study and suggest directions for future work. Finally, in Sec. VIII, we provide recommendations for departments interested in adopting rubric-based review.

For readers especially interested in implementing a rubric-based admissions process in their department, we direct the readers to Sec. III A for background on our admissions process, Sec. VIII for recommendations on how to implement rubric-based review, and Secs. II and III of the Supplemental Material [19] for our rubric and applicant statement prompts.

## II. BACKGROUND

### A. A typical admissions process in physics

When applying to a physics graduate program in the United States, an applicant will submit their undergraduate transcripts, general and physics GRE scores, multiple statements addressing their background, prior preparation, and research interests, and letters of recommendation. A group of physics faculty, the admissions committee, then reviews the applications and offers admission to some of the applicants.

Historically, there have been two main approaches for admitting students: emphasizing research or emphasizing grades [20]. More recent work however has tended to find that programs, including the one studied in this paper, emphasize grades and test scores over research, in terms of both what faculty say they do [21,22] and what faculty actually do [23,24].

Numerous potential equity issues emerge when admissions is focused around test scores and grades. First, there is evidence that GRE scores vary based on gender and race [25,26] and the type of undergraduate university the test-taker attends [27]. When combined with the practice of using cutoff scores, which Potvin *et al.* estimate at least 1 in 3 departments do despite the creators of the GRE and physics GRE recommending against it [21], applicants from underrepresented groups in physics may be more likely to not make the first cut.

Second, the tests themselves can be a financial burden for students [28]. The cost to take the general GRE is currently \$220 in most parts of the world (and up to \$231.30 in some regions) [29] and the cost to take the physics GRE is \$150 [30]. In addition, if the applicant applies to more than 4 programs, they must pay \$27 per school to send their scores. As Owens *et al.* notes, some students also need to travel to a testing center, which may incur travel or lodging costs [31].

Third, grades vary by applicants’ demographics and the type of university they attended. Whitcomb, Cwik, and Singh found that wealthier, continuing-generation, white students earned higher grades and that even the most privileged racially underrepresented students in physics earned lower grades than the least privileged white students [32] and other work has found that Black students receive lower grades than their Asian and white peers [33]. Additionally, grades are not standardized measures across universities, with students at private universities tending to be awarded slightly higher grades than their peers at public universities [34].

Further, evidence has not necessarily supported these metrics as useful predictors of who will earn their Ph.D. For example, Miller *et al.* found that while grade point averages were useful to some degree for predicting completion, the physics GRE had limited use [26]. More recent evidence suggests that the physics GRE and undergraduate grade

point average only have a relation to Ph.D. completion because they are related to graduate grade point average, which is then related to Ph.D. completion [35].

Given known issues with test scores and GPA, why do programs continue to emphasize them over the qualitative parts of the application? Perhaps the simplest answer is that comparing numbers is quick and convenient [23]. Quantitative data are often seen as more objective and true than qualitative data are [36] and therefore it might be perceived to be easier to rank candidates on these measures. The creators of the GRE push for such a view, saying GRE scores “provide a common, objective measure to help programs fairly compare applicants from different backgrounds” [37]. A more nuanced answer might be that qualitative parts of an application can contain substantial variability in what is addressed and these parts of an application can have their own inequities (see Woo *et al.* for an overview [38]).

One possible conclusion is then that all application materials have inequities, after all they are produced in an inequitable society, so what is the point of changing anything. We instead adopt a pragmatic view that some parts of the admissions process are more inequitable than others and, therefore, our goal is to develop methods to minimize or eliminate inequities to the best of our ability in an inequitable society.

## B. Holistic review

One possible approach to addressing inequities in the admissions process is holistic review, which Kent and McCarthy define “as the consideration of a broad range of candidate qualities including ‘noncognitive’ or personal attributes” [13]. Here, we will use holistic review to refer to the general process regardless of what tools or systems are used to conduct it. When talking about our department’s rubric-based process or similar processes, we will use rubric-based holistic review.

While the idea of holistic admissions is hardly new, its implementation is becoming more common due to both greater awareness that quantitative measures may not accurately predict success in graduate school [39–41] and institutions wanting to use the most predictive measures of success in their programs [13]. In addition, professional societies such as the American Astronomical Society (AAS) have called for programs to implement “evidence-based, systematic, holistic approaches” to graduate admissions [42].

Using holistic review has also been claimed to lead to beneficial outcomes for universities including increasing diversity and improving student outcomes (see Ref. [13]), though most of these studies have happened outside of physics and related fields. For example, Hawkins found that using holistic review increased diversity in a Doctor of Physical Therapy program [43] and in a literature review of predominantly medicine-related fields, Francis *et al.* found

that holistic review generally increased racial and ethnic diversity [44]. For science, technology, engineering, and mathematics (STEM) fields, Wilson *et al.* found that using holistic review in a biomedical science program resulted in applicant assessments that were independent of gender, race, and citizenship status [45] and Pacheco *et al.* found that using a composite score that included GPA, test scores, research experience, and publications was correlated with earning a university fellowship and a shorter completion time while applicant’s test scores and GPAs individually were not [46].

While holistic review shows promise, programs may have concerns about implementing it. For example, common concerns include limited faculty time to review applications, a lack of data correlating admissions criteria and student success, and limited resources to implement it [13]. In addition, there may be concerns that because the decisions can be more subjective than using a quantitative measure like a test score, there may be variability based on who reviews the application. However, a study of holistic admissions at the undergraduate level found that only 3% of reviews showed substantial variability in the overall score between reviewers [47], suggesting that in practice variability in the overall rating between reviewers is limited.

### 1. Noncognitive skills

Regardless of the specifics of a holistic review process, most approaches include some examination of the applicant’s noncognitive skills, which may also be referred to as soft skills, personality traits, character traits or socioemotional skills depending on the discipline or context [48]. While there are multiple definitions of these (see Ref. [49]) we adopt Roberts’ definition that noncognitive skills or personality traits are “the relatively enduring patterns of thoughts, feelings, and behaviors that reflect the tendency to respond in certain ways under certain circumstances” [49]. Often these have been operationalized as the big five, which are openness to experience, conscientiousness, extroversion, agreeableness, and neuroticism [48,50], though other categorizations exist. For example, in higher education admissions, Sedlacek proposed eight noncognitive traits, which he defines as things not measured by standardized tests: positive self-concept, realistic self-appraisal, understands and knows how to handle racism (the system), prefers long-range to short-term or immediate needs, availability of strong support person, successful leadership experience, demonstrated community service, and knowledge acquired in or about a field [51] while in a review of the noncognitive skills literature, researchers found 15 terms that are often categorized as noncognitive skills including attention, cognitive flexibility, conscientiousness, delay of gratification, effortful control, emotional reactivity, emotional regulation, executive function, impulsivity, inhibitory control, persistence,

self-control, self-regulation, temperament, and working memory [52]. Other researchers have instead argued that executive function and self-regulation are the overarching noncognitive skills and all others fall under these [53].

In terms of their utility, noncognitive skills have been found to be predictive or correlated with academic success, though these studies have happened outside of the context of physics. At the undergraduate level, noncognitive skills in isolation and in concert with test scores have been found to be more predictive of success and graduation than test scores alone [54–56]. Likewise, at the graduate and professional levels, noncognitive skills have been found to be correlated with GPA and class rank [57,58], clinical performance [59], and overall success in programs [60,61] but were not found to be associated with doing well on a licensing exam [62]. Of the individual noncognitive skills, conscientiousness has been found to be most strongly and consistently associated with academic success [63].

In addition to their benefits related to academic success, noncognitive skills can be useful for promoting equity in admissions. For example, including noncognitive skills can increase diversity without harming validity [64,65] as noncognitive measures have been shown to be just as valid for majoritized and minoritized groups [64,66,67]. While including noncognitive skills as part of admissions may seem like a hard ask of faculty, many faculty already acknowledge the usefulness of noncognitive skills in graduate school [64], including in physics [68].

Yet, a pressing concern is how to measure such noncognitive skills accurately. While applicant self-reports or recommender ratings are typical approaches, such methods may result in inflated or skewed ratings [64]. A recent study suggests that even sharing descriptions of noncognitive skills and why they are useful for predicting later success can artificially inflate judgments [69]. Thus, how best to measure such skills is still an active area of inquiry [70].

## 2. Rubric-based review

One promising approach to implementing holistic review is rubric-based review. Under this approach, applicants are evaluated based on a set of predefined criteria. By pre-selecting criteria, what is required for admission is clear to reviewers and provides a structure to assess all applicants [15,23]. This explicitness has been shown to enhance both validity and reliability [38,71,72].

In addition, rubrics can help make the admissions process more equitable [23]. By explicitly laying out the review criteria and what is required to achieve each level of the rubric, all applicants can be judged fairly and individual reviewer's expectations can be mitigated [67]. From research into other areas of academic hiring, we know that gender and racial biases exist in the hiring process, including in physics [73,74]. Specifically in graduate admissions, faculty, including astronomy and physics faculty, have been documented showing preferences to

applicants with similar backgrounds as themselves or within the same research subfield of their discipline [23]. Thus, rubrics offer a possible route to counter those biases. Indeed, a recent study in admissions for a psychiatry residency program found that using rubric-based holistic review led to more underrepresented applicants receiving an offer to interview compared to the traditional approach [75] while a recent study of grade-school writing found that teachers rated writing attributed to a Black author lower than when it was attributed to a white author but did not find the effect when the teachers were instructed to use a clearly defined rubric [76].

As rubric-based approaches to admission are still relatively new, best practices are still in development. Yet, a few recommendations do exist [15]. First, criteria should be selected before reviewing any applications with individual programs deciding what qualities are critical for success in their program [42]. Second, rubrics should be coarse grained in that there are fewer possible scores for each construct such as low, medium, or high instead of 1–10 to limit disagreements over scores [67]. Third, each level of the rubric should be clearly defined so that a reviewer can easily determine which score an applicant should get on each construct. These levels should be picked so that each possible score will be received by many applicants [15]. Finally, these criteria and levels should allow for diverse forms of excellence to be counted as achievements so that applicants with nontraditional markers of excellence are not excluded [77]. For example, assessing an applicant's research abilities should go beyond their number of publications or number of years working in a research lab and could instead focus on what the applicant accomplished in the lab or what skills they have.

While rubric-based approaches have received little research in physics, they have been successfully incorporated into larger physics graduate program initiatives. Two of the most well-known initiatives are the Fisk-Vanderbilt program, which graduates one of the largest classes of Black Ph.D. physicists in the nation [78], and the APS Bridge Program, which has successfully admitted and retained graduate students of color at rates higher than the national average [2]. Even though rubrics in admission were one of many changes made, these programs suggest that rubric-based review has promise.

For a more in-depth review about equitable admissions practices in STEM doctoral programs, we refer the reader to Roberts *et al.* [16]

## III. METHODS

### A. Our rubric and applicant evaluation process

In 2018, the Department of Physics and Astronomy at Michigan State University introduced a rubric-based approach to evaluate applications to the graduate program in physics, informed by the Council of Graduate Schools'



2016 report on Holistic Review in Graduate Admissions [13]. The main goal was to improve the identification of strong candidates for the program and to make the selection more equitable, thereby increasing the participation of students from underrepresented groups in the department. In preparation for the introduction of the rubric, Casey Miller and Julie Posselt, the Inclusive Practice Hub Director and Research Hub Director, respectively, of the National Science Foundation supported Inclusive Graduate Education Network, led a workshop with faculty who served at that time in the Graduate Recruiting Committee. This workshop resulted in a selection of five rubric categories, which each had several subcategories. Applicants are ranked with a score of either 0, 1, or 2, corresponding to low, medium, or high, for each subcategory, based on defined criteria for each score. The subcategory scores are then averaged per category and category scores summed (with weights as given below) to calculate the overall score. The categories, with subcategories in parenthesis, are

- Academic preparation, with a weight of 25% (physics coursework, math coursework, other coursework, and academic recognition and honors)
- Research, with a weight of 25% (variety and duration, quality of work, technical skills, and research disposition)
- Noncognitive competencies, with a weight of 25% (achievement orientation, conscientiousness, initiative, and perseverance)
- Fit with program, with a weight of 15% (fit with research programs of the department, fit to research programs of specific faculty, (prior) commitment to participation in the department or school community, and advocacy for and/or contributions to a diverse, equitable, and inclusive physics community)
- GRE scores, with a weight of 10% (general GRE scores, and physics GRE scores) [79]

The full rubric, which part of the applicant's file is used to evaluate each category, and the prompts for the statements applicants are evaluated on are included in the Supplemental Material [19].

The choice of these categories and subcategories was based on the discussions in the workshop and advice from the workshop leaders, and included considerations based on experiences during previous recruiting cycles. Another consideration for the choice of the categories is a reasonably close alignment with criteria used at MSU for awarding fellowship packages to students. Therefore, the rubric scoring can also be used for selecting nominations for university fellowships. This is important because fellowship nominations are due shortly after the application deadline (January 1).

Applications for the graduate program are submitted to MSU's central application system. All folders with a complete or near-complete application package are reviewed. The applications are divided up into several

groups, which each are reviewed by different members of the graduate recruiting committee. This committee has a rotating membership with representation from faculty in all major research directions present in the department. Committee members are instructed about the use of the rubric and provided with the criteria. As part of the review process, they also sort students by their interest in research area(s). The results from the rubric scoring are compiled by the Graduate Program Director. Students whose folders are near complete, but have a ranking for which an offer is not impossible, are contacted and asked to provide the missing information. If that additional information is provided, the rubric scoring is updated.

Subsequently, the spreadsheet is used by committee representatives from each major research area in the department to make a list of students they would like to make an offer to for a position in that specific research area. The number of students who are made an offer to depends on openings available per research area, the number of teaching assistant slots available, and the historical acceptance rates for each research area. Therefore, offers are not made strictly based on the rubric score or a cutoff value and instead, rubric scores are used as a guide for making offers. Minor changes to weight factors used in the rubric do not significantly change the rubric scores or affect admissions decisions.

Typically, the process results in a list of offers that will be made and a wait list for additional offers that can be made if recruiting targets are not met in the initial round of offers. In this stage of the recruiting process, the match to available positions is revisited as committee members from specific research areas are better aware than general faculty members about the recruiting needs for that year. In spite of the instructions and criteria provided to reviewers, the scoring is still somewhat subject to differences in reviewing styles and interpretation of the criteria. This is, for example, apparent in the comparison of average summed scores per reviewer. Therefore, this second stage of the review process also allows for another comparison of applications based on the rubric by a few faculty members in each research area. Because of these reasons, the list of students whom an offer will be made to, or who are put on a wait list, quite closely follows the original rubric scoring but modifications do occur.

The whole process is organized and overseen by the Graduate Program Director with support from the Graduate Program Secretary. The Graduate Program Director also serves as the point of contact for questions about the use and interpretation of the rubric, reviews applications of likely candidates, and leads the selection of nominations for fellowships.

Based on informal communications, the overall response from faculty who served in the recruiting committee and used the rubric has been positive, as it provides clear guidance for the review process and reduces the impact of

different reviewing styles and biases to what are the most important skills applicants to a physics graduate program should have. On average, the time spent by individual committee members on reviewing the folders has not increased and takes between 15 and 30 for a complete application. Faculty reviewers have provided feedback that it would be better if applicants are first sorted by research area so that the review is done by several faculty from the relevant research areas in the first step. Given the large number of applications and the limitations of the current software used to manage applications, this could not easily be accomplished in the past. MSU is implementing new software for managing and reviewing applications, which will make presorting of applications by research area possible, leading to a considerable increase in the efficiency of the process.

### B. Participants and data collection

Data for this study come from compiled records from applicants to our physics graduate program for Fall 2018, 2019, and 2020. Most admissions decisions for Fall 2020 had already been made before coronavirus accommodations took effect, suggesting at most minimal effects on our data.

When applying to the university, applicants submit a general university application, transcripts, test scores, a personal statement, an academic or research statement, and letters of recommendation to a central system. As the current admissions system does not allow for records to be compiled across applicants, two researchers manually extracted relevant information for this study. The researchers independently extracted data from the first 20 applications and then compared results to ensure they were interpreting the applications the same and agreeing on any conventions for reporting the data. Afterwards, the researchers independently went through the rest of the applications. Through this process, the researchers collected the applicant's demographics, grade point average, GRE scores, degrees earned or in progress, and previous institutions attended. Any information missing from the applications or entered into the application on a nonstandard scale (e.g., a GPA on a non 4.0 scale or a GRE score outside of the current scoring range) was treated as missing data for the analysis.

As rubric scores are determined by faculty and are not part of the materials applicants submit, aggregated scores were then matched with individual applicants using the applicant IDs. Through this process, we collected data on 826 applicants, 511 of which were domestic applicants.

### C. Analysis

Because of different application requirements and availability of institutional data for international and domestic students, we only include domestic students in our study. In addition, we only include applications sufficiently

complete that faculty were able to rate and were included in the Graduate Program Director compiled records, leaving us with 321 domestic applicants for this study. Applicants with missing information are often contacted to obtain the missing information. However, any evaluations happening after the initial review are not included in our data and we cannot make any conclusions on why applicants without complete files initially were later admitted or not admitted. Overall, only 18% of those without sufficiently complete files initially were later offered admission.

For our analysis, we were interested in how faculty rate applicants and hence, we computed the fraction of applicants in each level (low, medium, and high) of the rubric. In some cases ( $< 5\%$ ), faculty used a rating that was in between levels (e.g., low-medium). Because of this, we performed all subsequent analyses by first rounding up (so low-medium would become medium) and then repeating the analysis by rounding down.

First, we computed the fraction of applicants in each level of the rubric for all applicants, all admitted applicants, and all nonadmitted applicants.

Second, we compared applicants based on demographics by comparing the fraction of applicants in each bin of the rubric. While gender would be more appropriate, the application system only asks applicants about their sex and allows them to choose male or female. Thus we were only able to compare faculty ratings of males and females. We acknowledge that females is not the correct term to use, but as being female does not automatically imply being a woman, we do not believe it is appropriate to assume that someone marking female as their sex is necessarily a woman.

In terms of race, the application system does not allow applicants to enter their race or ethnicity, so we are unable to compare applicants of different races.

Finally, we compared applicants from different undergraduate backgrounds because prior work suggests the applicant's background may influence faculty's perceptions of them. For example, faculty may prefer applicants with similar backgrounds as themselves [23] and may interpret grade point averages in the context of the applicant's undergraduate program, with high GPAs from more prestigious universities carrying more "weight" than a high GPA from a lesser known school [80]. In addition, graduate admissions in physics have been characterized as "risk averse" where faculty prefer to admit applicants who are likely to complete their program rather than take chances on someone who might not and thus, prefer applicants who compare favorably to previously successful students according to some metric [18,23]. As students from smaller programs may be viewed as higher risk if previous students from that program struggled [80], it is possible faculty may be less likely to admit students from smaller undergraduate schools.

TABLE I. Percent of missing data by rubric construct.

Rubric construct	Percent missing
Physics coursework	20.0
Math coursework	20.2
All other coursework	20.2
Academic honors	22.1
Variety or duration of research	3.4
Quality of work	4.4
Technical skills	4.1
Research dispositions	4.7
Achievement orientation	4.4
Conscientiousness	4.4
Initiative	4.0
Perseverance	4.4
Alignment of research	7.2
Alignment with faculty	32.1
Community contributions	4.0
Diversity contributions	3.4
General GRE scores	2.2
Physics GRE score	2.5

To characterize an applicant’s undergraduate background, we used two binary measures. First, we used Barron’s value, which is a measure of an institution’s selectivity based on incoming students’ SAT scores, GPA and class rank, and overall acceptance rates. While not equivalent to prestigious, we treat selectivity as a proxy for prestige based on the assumption that more selective institutions are also prestigious institutions. For our analysis, we defined institutions with Barron’s values of “most competitive” or “highly competitive” as selective and all other institutions as not selective.

Second, we used the number of bachelor’s degrees awarded by the physics department at the applicant’s undergraduate institution to estimate the size and reputation of the department, with the assumption that a department that grants more degrees is more likely to be regarded as a “good” department and to be known by an admissions committee member. Because of variability in yearly degrees, we used the median number of degrees over the 2016–2017, 2017–2018, and 2018–2019 academic years as the number of bachelor’s degree awarded [81–83]. We then defined any program that was in the top quartile of physics bachelor’s degrees awarded during that period (more than 14 degrees per year during a typical year) as a large program and all other programs as smaller programs. For reference, the programs we classified as large produced nearly two-thirds of all physics bachelors degrees over the period.

To perform the comparisons in all cases, we used Fisher’s exact test to examine whether the rubric score was associated with any of the metrics of interest

(admission status, sex, institution selectivity, institution size). We used the standard choice of  $\alpha = 0.05$  to judge claims of statistical significance. Because we did 18 comparisons for each metric of interest, it is likely that there would be at least one false positive. Therefore, we used the Holm-Bonferroni procedure to correct the  $p$  values for multiple comparisons as it is less conservative than the traditional Bonferroni correction while maintaining statistical power [84].

For cases of missing data, we used pairwise deletion so that we could make the most use of the data we had. While Nissen *et al.* recommends using multiple imputations for missing data in physics education research studies [85], the goal of this paper is to understand what faculty did as opposed to estimate a larger trend or predict an outcome. Therefore, we do not believe that using multiple imputations is aligned with the goal of this paper. The percent of missing data per rubric metric is shown in Table I.

### IV. RESULTS

The results are largely unchanged based on whether we rounded up or rounded down when a faculty member gave a rating in-between levels of the rubric so we present only the rounded up results here.

When we examine the faculty’s rating of all applicants in Fig. 1, we notice two overarching trends. First, for traditional measures of academic success such as grades and test

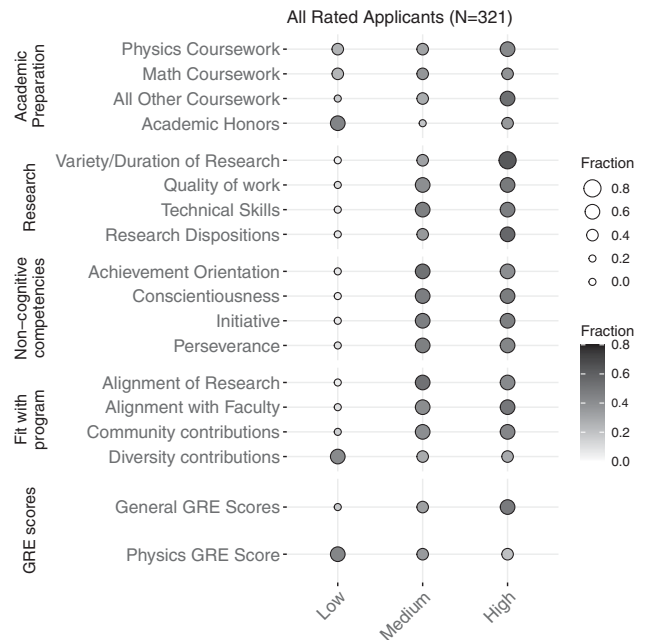


FIG. 1. Faculty ratings of domestic applicants on 18 constructs. In the plot, a larger, darker circle means that more applicants are in that bin. While many applicants are in each level of the academic preparation and test score constructs, few applicants are in the “low” bin of the research, noncognitive skills, and program fit constructs.

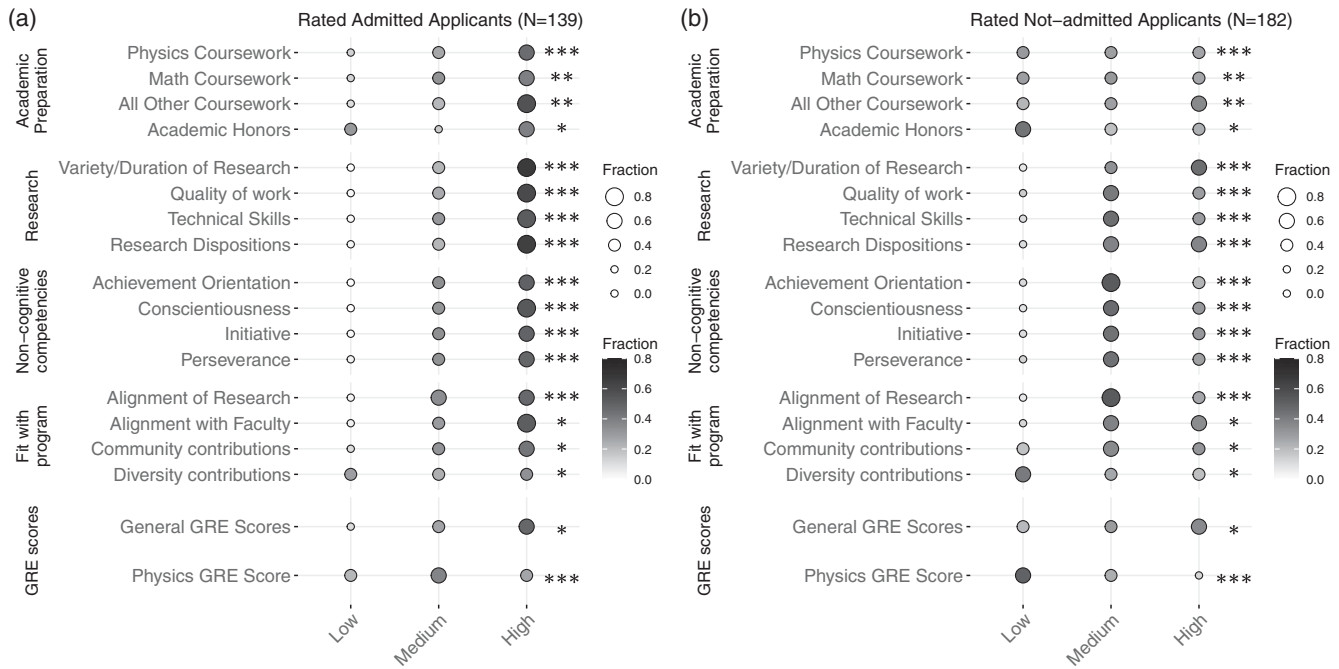


FIG. 2. Faculty ratings of domestic applicants on 18 constructs split by whether the applicant was admitted. Ratings that are statistically different between the two plots are marked on the right side of the plot, with a corrected  $p$  value  $<0.001$  represented by “\*\*\*,”  $<0.01$  by “\*\*,” and  $<0.05$  by “\*.” The distribution of ratings of all constructs is statistically different for admitted applicants compared to nonadmitted applicants. Overall, most admitted applicants were rated “high” while most nonadmitted applicants were rated “medium”.

scores, faculty tend to rate applicants using all three levels of the rubric. For the academic preparation constructs on the rubric, high is the most common rating given by faculty. However in terms of math and physics course grades, around 25% of applicants still scored in the low bin. Of the academic preparation constructs, academic honors follows a different structure than the others where faculty ratings are bi-modal, meaning that applicants either had no academic honors or had multiple academic honors.

Second, for the research, noncognitive, and fit constructs, faculty rarely used the “low” level of the rubric, with only three of the twelve constructs in those categories having more than 10% of applicants earning a low. For research, the most common rating was “high” while for the noncognitive traits, the most common rating varied between high and “medium.” In terms of the fit constructs, most applicants were rated as either medium or high for alignment of research, alignment with faculty, and community contributions. In contrast, for the diversity contributions construct, low was the most common rating, meaning that many applicants did not discuss how they promote or advocate for diversity in their applications.

When looking at how faculty rate applicants who would later be admitted compared to applicants who would not be admitted, we see statistically significant differences in the distribution of all ratings (Fig. 2). Overall, admitted applicants tended to be rated high on each construct while nonadmitted applicants tended to be rated medium on each

construct. There were a few exceptions to the general trend however. For academic honors, diversity contributions, and physics GRE scores, most admitted students were not rated as high and 25% of applicants received a low score while for all other course work, variety or duration of research, and general GRE scores, most nonadmitted applicants were rated as high.

When looking at the ratings broken down by sex independent of admission status (Fig. 3), we notice that the results tend to follow the overall patterns of all three ratings on academic success and test scores and mainly medium and high ratings on research, noncognitive skills, and fit with the program for both males and females. Comparing ratings between males and females, we find that only physics GRE score, community contributions, and diversity contributions showed statistically significant differences. While males tended to score higher on the physics GRE score, females tended to score higher on community contributions and diversity contributions. As we elaborate on in the discussion, differences in these three constructs do not necessarily mean that faculty are rating males and females differently but instead may be documenting inequities that already exist.

Likewise, when looking at the ratings broken down by the selectivity of the university where the applicant earned their bachelor’s degree independent of admission status (Fig. 4) or the size of the department where they earned their bachelor’s degree independent of admissions status



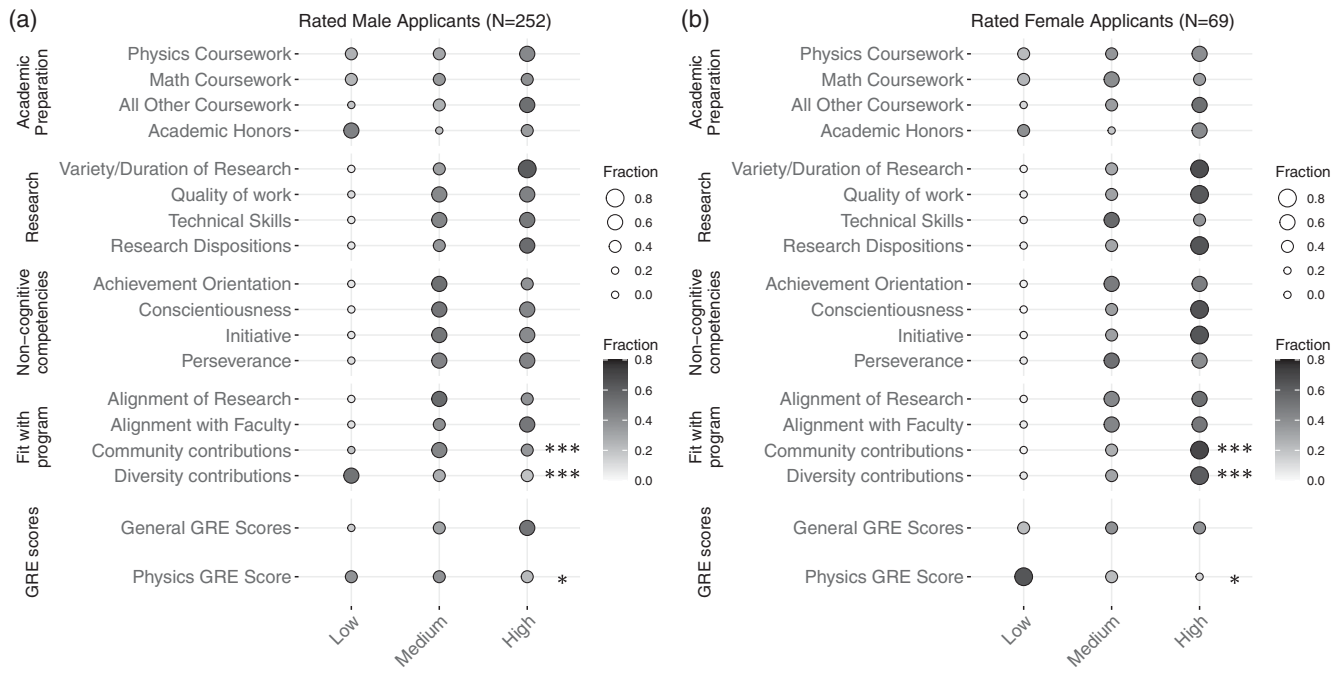


FIG. 3. Faculty ratings of domestic applicants (admitted and nonadmitted) on 18 constructs split by whether the applicant was male or female. Ratings that are statistically different between the two plots are marked on the right side of the plot, with a corrected  $p$  value  $<0.001$  represented by \*\*\*,  $<0.01$  by \*\*, and  $<0.05$  by \*. Only three of the constructs showed statistical differences between males and females: physics GRE score where males scored higher and community contributions and diversity contributions where females scored higher.

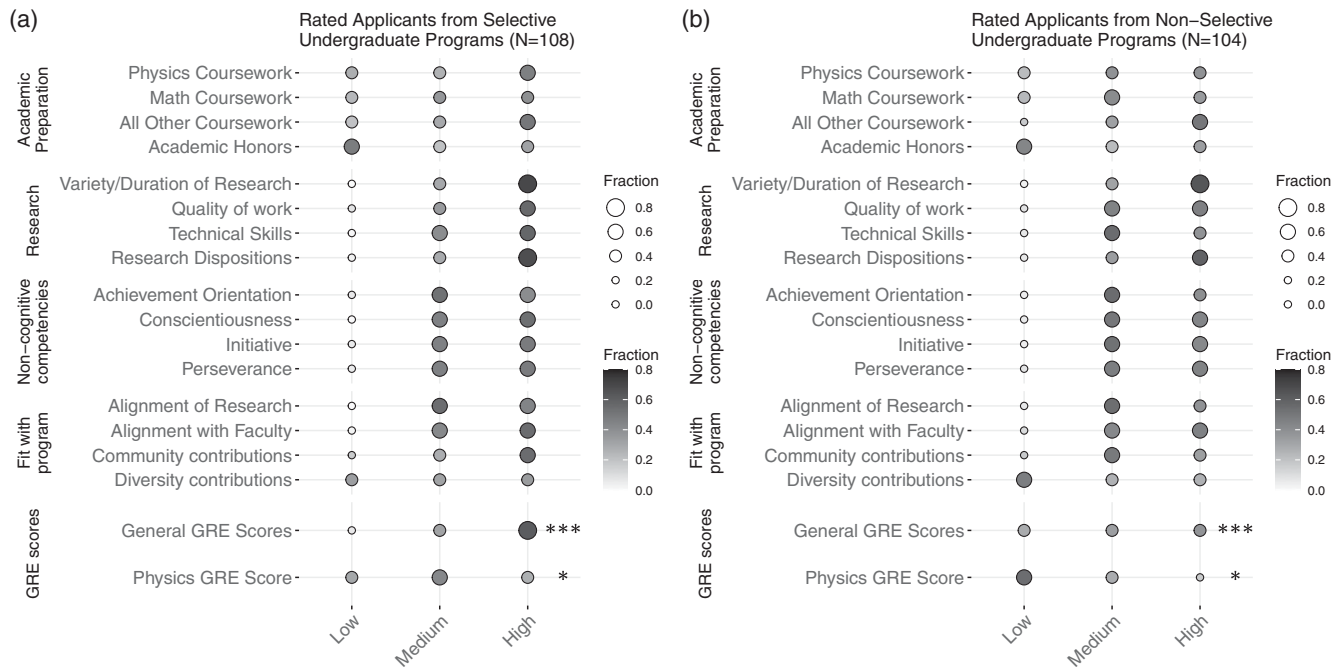


FIG. 4. Faculty ratings of domestic applicants (admitted and nonadmitted) on 18 constructs split by whether the applicant attended a more selective or less selective undergraduate university. Ratings that are statistically different between the two plots are marked on the right side of the plot, with a corrected  $p$  value  $<0.001$  represented by \*\*\*,  $<0.01$  by \*\*, and  $<0.05$  by \*. Only the general GRE and physics GRE scores showed differences.

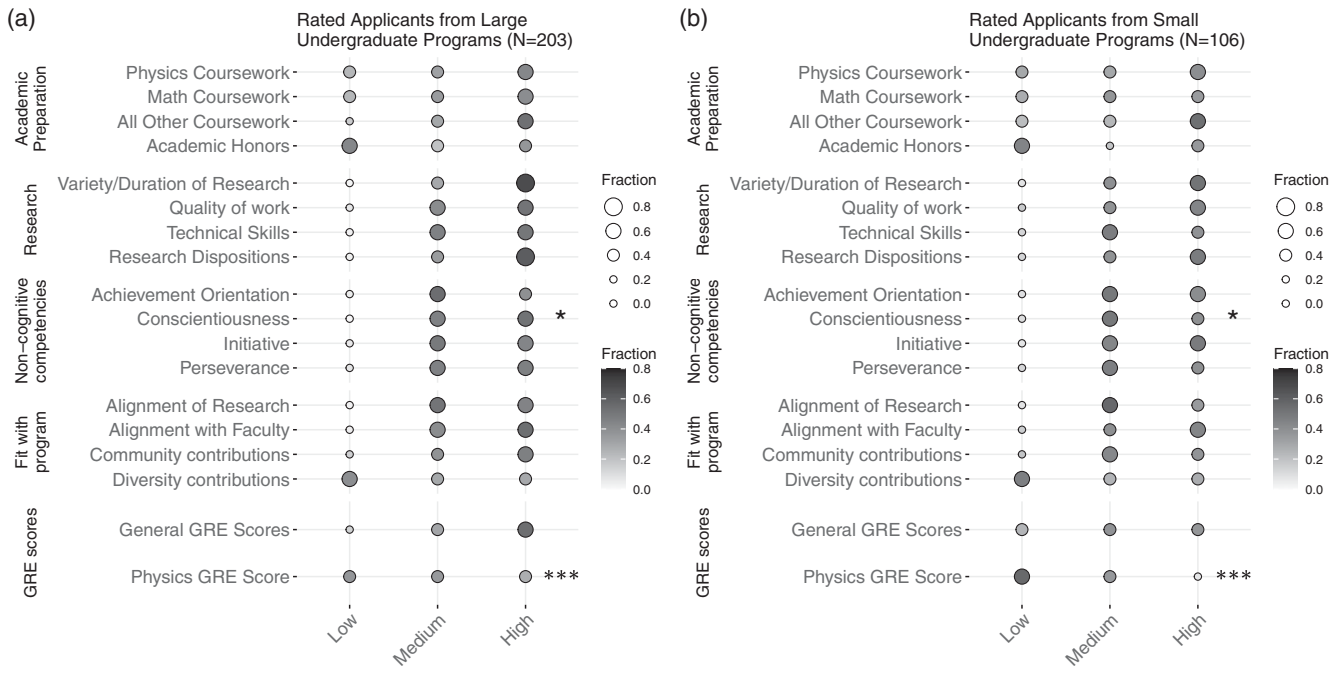


FIG. 5. Faculty ratings of domestic applicants (admitted and nonadmitted) on 18 constructs split by whether the applicant attended a university with a larger or smaller physics program. Ratings that are statistically different between the two plots are marked on the right side of the plot, with a corrected  $p$  value  $<0.001$  represented by \*\*\*,  $<0.01$  by \*\*, and  $<0.05$  by \*. Only the physics GRE score and conscientiousness showed differences between the groups of applicants, with the latter dependent on how larger physics program is defined.

(Fig. 5), we may also be observing existing inequities reflected in the faculty ratings. For example, applicants from more selective universities only had statistically higher ratings than applicants from less selective universities on the general GRE and physics GRE scores. Similarly applicants from larger programs had statistically higher ratings on the physics GRE score than applicants from smaller programs did. However, applicants from larger programs were also rated higher on conscientiousness than applicants from smaller programs, though this result is sensitive to how we define a large program.

While we could consider interactions between admission status and sex, institutional selectivity, or physics program size, we did not do so given the small sample sizes. For completeness, however, we include plots broken down by admissions status in the Supplemental Material [19].

### V. DISCUSSION

*How do faculty assign rubric scores to applicants and how do those differ between admitted and rejected applicants?* For academic achievement and test scores, faculty tended to use all three levels of the rubric when assigning scores to applicants. In contrast, faculty tended to use mainly medium and high when assigning scores to applicants in the research, noncognitive skills, and program fit categories. We argue that this result is more of a reflection of the rubric than it is of how faculty are using the rubric.

Given that grades and test scores are well defined via transcripts and test scores, rubric constructs measuring these tended to use quantitative measures to determine which score the applicant would receive. That is, a high score or high grades would correspond to a high rating while a low score or low grades would correspond to a low rating. Additionally, as the courses required for a physics degree tend to be similar regardless of the specific program, most applicants will have taken the courses mentioned in the rubric and hence, faculty can rank applicants based on those grades.

In contrast, the research, noncognitive skills, and fit with department are less well defined and instead depend on what applicants write in their statements and what information letters of recommendation contain. While applicants are asked to respond to specific points in their two statements that align with constructs on the rubric (prompts in the Supplemental Material [19]), it is up to the student to do so and to provide enough detail for an admissions committee member to rate the applicant. This means that not all constructs on the rubric may necessarily be addressed. For example, if an applicant takes quantum mechanics it will certainly appear on their transcript but if that applicant was also active in departmental service activities, it may not be reflected in any parts of the application. As a result, the rubric needs to take into account that applicants may not display a trait either because they do not exhibit it or because they did not

mention it (the instructions for the applicant's statements ask them to address multiple topics that map onto various rubric constructs). Any display of the trait could then not fall into the low level of the rubric, which would then explain why faculty tended to use only the medium and high ratings.

A reasonable follow up is then whether combining "no evidence" with "evidence not presented" as a single level on the rubric represents an issue with the rubric. We argue that it does not, as it provides the best option given the data faculty have available. Applicants are asked to discuss certain topics in their statements that map broadly onto the rubric constructs but that does not necessarily mean they will. While interviews could be useful in separating no evidence cases from evidence not presented cases, we worry these would increase admissions committee members' work load.

In terms of comparing admitted and nonadmitted applicants, all 18 rubric constructs showed statistically significant differences. Given the goal of the rubric is to aid faculty in determining who to admit, we would expect the rubric to show such differences. That all rubric constructs show differences suggests all parts of the rubric are useful for determining who to admit.

*How do the scores assigned by faculty differ by applicant's sex?* We found only three constructs on the rubric that showed sex differences: physics GRE score, community contributions, and diversity contributions. Given known scoring gaps on the physics GRE [26], it is not surprising that males are rated more highly than females on the physics GRE score. Given that females perform larger amounts of service work in academia [86] and are more likely to volunteer in general [87], it is also not unexpected that constructs measuring these would show a difference between sexes. Because the constructs that show sex differences are related to effects documented in the literature, we believe that the rubric is reflecting inequities that already exist rather than creating additional ones. Therefore, we conclude that the rubric is not providing an advantage to male or female applicants and thus is equitable in terms of sex.

Additionally, the constructs of the rubric that do not show differences between sexes also align with what we would expect based on the literature. The result that physics and math GPA did not differ by sex aligns with the findings of [32] and the result that noncognitive skills did not differ by sex aligns with the general finding that noncognitive skills do not appear to depend on demographics [64,67].

*How do the scores assigned by faculty differ by the type of institution the applicant attended?* When we compared applicants based on whether their undergraduate institution was a more or less selective institution, we found that the only constructs that showed differences were the general GRE and physics GRE scores. This result aligns with the results of our previous work investigating the physics GRE

scores by undergraduate institution type [27,88]. We note that if we instead define more-selective universities to include large state universities, such as Michigan State University, University of Colorado, Boulder, and University of Washington, our results are unchanged. This redefinition is equivalent to considering Barron's values of 1–3 as more selective and everything else as less selective compared to the definition of more selective as Barron's values of 1 and 2 in Secs. III and IV.

The interpretation of the results when comparing applicants from larger or smaller physics departments is less straightforward because the results do depend on how we define "larger" and "smaller" departments. When we define larger programs as those that ranked in the top quartile of physics bachelor's degrees granted as measured by the median number of degrees awarded over the last three years of available data and rounded up in-between ratings, we find that the physics GRE score and conscientiousness showed differences between applicants from larger and smaller programs. However, if we rounded down on in-between ratings instead, only physics GRE score showed a difference between applicants from larger and smaller programs.

Furthermore, alternative definitions of "larger programs" also produced varying results. One could also have reasonably defined "larger" to mean (1) in the top half of physics bachelor's degrees granted as measured by the median number of degrees awarded over the last three years, (2) in the top quartile of physics bachelor's degrees granted as measured by the total number of degrees awarded over the last three years, and (3) in the top half of physics bachelor's degrees granted as measured by the total number of degrees awarded over the last three years. When we also consider rounding up or rounding down in-between ratings, we could make various combinations of physics GRE score, general GRE score, physics coursework, and conscientiousness show a statistically significant difference. The only rubric construct that always showed a statistically significant difference regardless of how we defined "larger programs" was the physics GRE score. Therefore, the results suggest that applicants from larger physics programs score higher on the physics GRE than applicants from smaller program do, but the results are inconclusive as to whether other areas of the rubric might show differences based on the size of the physics program the applicant attended.

Because we found only one consistent construct that varied based on the institution attended and that difference is expected based on previous research, the results suggest that the rubric is supporting equitable admissions when it comes to undergraduate institutions. However, other constructs can show differences based on undergraduate institution and how we binarize it, and we should therefore be cautious about claiming the rubric is equitable when it comes to undergraduate institutions.

One area that unexpectedly did not show differences regardless of how we defined larger program was the research section. It is often assumed that students at larger programs have more opportunities to engage in research than students at smaller programs. Yet, even if that is true, it does not appear to be reflected in the rubric scores.

## VI. LIMITATIONS

Our study has four main limitations. First, our study does not include many disadvantaged groups in higher education who might not have the same opportunities as their more privileged peers and hence, may score lower on the rubric. While gender and race are the most obvious due to the way our university records applicant data and interprets proposal 2, our study does not include a comparison of low-income applicants to higher-income applicants or first generation applicants to continuing generation applicants.

Additionally, the size of our study does not allow us to explore intersections and where possible inequities may lie. As Rudolph *et al.* noted, using small sample sizes with subgroups has insufficient statistical power and could lead to invalid inferences [42]. Hence, we refrained from performing such analyses in this paper.

Second, our data only contained ratings from the initial reviewer and none of the ratings of later reviews. As a result, we were unable to look into differences in how individual faculty members use the rubric. For example, it is possible that one faculty member might systematically rank applicants lower on the rubric constructs than a different faculty member might. With only a single rating per application, we are unable to disentangle differences in faculty ratings and differences in the applicants the faculty members reviewed.

Third, this study included only a single program. Under a more traditional graduate admissions system, physics has been called a “high consensus” discipline [23], meaning that physics faculty tend to agree on what a “quality” applicant is and therefore, a single department’s admissions process would be more or less representative of graduate admissions processes in physics. When switching to rubric-based admissions, we cannot necessarily make that same claim. As our rubric was created based on what faculty value, it is not unreasonable to assume that the results would generalize to other departments that also use rubric-based admissions. However, until such processes are evaluated at other departments, we cannot make such a claim.

Fourth, as a result of using only one program, the applicants are likely not representative of the larger population. The data in this study comes from (i) people who applied to our program and (ii) applicants who had a nearly complete application. Thus, if we consider those with an interest in attending physics graduate school as our population, we first selected on those who applied to graduate school, then selected on those who applied to

our program, and finally selected on those who provided enough information in their applications for faculty to evaluate. At each step, we are excluding some of the larger population and thus our claims cannot necessarily be expected to hold for the larger population of potential applicants. For example, anecdotal evidence suggests minoritized applicants are more likely to not complete their applications than majoritized applicants are.

## VII. FUTURE WORK

As noted in the limitations, our study compared rubric scores of males and females and applicants from larger or more selective programs with applicants from smaller or less selective programs. Future work could then explore how rubric-based admissions may impact other historically and currently underrepresented groups in physics such as Black, Latinx, or Indigenous applicants. Racism, and specifically anti-Black racism, is still prevalent in physics [89–93] and therefore might be reflected in rubric-based admissions.

While physics faculty tend to think of diversity mainly in terms of race [23], we acknowledge that diversity is broader than race and studies of equity around the rubric should also consider first generation applicants, low-income applicants, disabled applicants, and veterans. Studies of undergraduate admissions suggest that when extracurriculars and subjective assessments of character and talent gleaned from essays and recommendations are added to the admissions process, existing inequalities may increase [94] and these applicants may become further disadvantaged in the admissions process. Therefore, future work should ensure that rubric-based admissions do increase equity rather than just use a new tool to perpetuate existing inequities.

Second, future work should examine how the use of rubrics may affect what parts of an application drive the admissions process. In our prior work, we found that the physics GRE and grade point average were the main drivers of the admissions process [24]. Given the rubric is designed to emphasize more than just grades and test scores, we would hope to see these factors deemphasized under the rubric system. Such a result would suggest that the rubric is fundamentally changing how faculty are reviewing applicants.

Third, future work could examine the impact of rubric-based admissions in terms of admitted applicants and student outcomes. In terms of who is admitted, our initial, but limited departmental data on admitted applicants suggests that women and applicants from underrepresented racial and ethnic groups make up a larger portion of admitted applicants under the rubric-based admissions model. In 2017, the last year before the implementation of the rubric, 13% of admitted applicants were women and 9% were applicants from underrepresented racial and ethnic groups. During the three years of this study, those numbers were 27%, 29%, and 31% for women



and 6%, 10%, and 12% for applicants from underrepresented racial and ethnic groups. While the percent of admitted applicants from a minoritized group in physics seems to be increasing under our rubric, that does not necessarily mean that the percent of minoritized applicants admitted is increasing. Future work should also investigate the latter.

In terms of student outcomes, future work can investigate how students admitted under the rubric compare to students admitted under the previous approach in terms of graduation rate, time to completion, and passing comprehensive exams. Faculty skeptical of holistic admissions may worry that by deemphasizing grades and test scores, their program is admitting less academically prepared students and therefore, students may take longer to achieve program milestones. Future work can explore if these fears have any merit. Research at the undergraduate level on holistic admissions has found that adding noncognitive traits increased graduation rates, especially among those from disadvantaged backgrounds [95]. At the graduate level, a study of a materials science and engineering program found that after changing their admissions to include noncognitive skills, their incoming students won more university fellowships, though the authors cautioned they could not attribute the increase in fellowships solely to their changes in admissions [96]. In addition, another study of rubric-based holistic admissions in engineering found that most faculty did not believe the quality of admitted students decreased after switching to more holistic admissions [97]. Thus, evidence from outside of physics suggests that these fears may be unfounded, but we will not know for sure until physics specific studies are conducted.

Additionally, future work can examine noncognitive skills in physics more broadly. Physics has been characterized as a brilliance-dominated field [98] and hence, it is not surprising that most studies of success in physics have also focused on cognitive measures such as grades, exam scores, and standardized test scores. While such studies could be useful at all levels of physics, studies at the graduate level are especially important given the limited number of studies exploring their usefulness for predicting success in graduate school. [42].

Finally, future work around equity in graduate admissions should investigate who is invited to apply to graduate school in the first place, what barriers those who do not apply but wish to do so encounter, and how those barriers may be removed. In previous work, Cochran *et al.* investigated what barriers applicants to physics graduate school, via the APS Bridge Program, perceived, finding that GRE scores, lack of research experience, low GPA, program deadlines, and application costs were common concerns [28]. Unless we also work to make the application process more equitable, making the evaluation process more equitable will not result in large-scale changes in equity at the graduate level.

Shifting from a researcher lens to a practitioner lens, future work can also examine how graduate programs as a whole can become more equitable and how graduate programs can ensure their program goals align with their admissions goals. Milestones in a typical physics graduate program include passing a series of courses and exams followed by completing independent research for a dissertation. While common, such assessment practices should be evaluated as to whether they are aligned with the goals for admission. For example, after implementing rubric-based admissions, our department no longer requires incoming students to take and pass a qualifying exam.

## VIII. RECOMMENDATIONS FOR DEPARTMENTS

The results of this study suggest a general recommendation to implement rubrics in physics graduate school admissions. Rubrics can aid reviewing applications by standardizing the process and limiting bias and using rubrics does not appear to increase the time to review applications.

Of course, simply using a rubric will not result in changes unless it is implemented well. We therefore propose three more specific recommendations.

First, we recommend that admissions committees have multiple members review each application. For a well-constructed rubric, there should be limited uncertainty as to what rating an applicant will receive. However, for constructs that are more subjective in nature, faculty may have differing opinions about what counts as achieving each level. For example, for the quality of work construct on our rubric, what counts as “making significant contributions to the project” might vary based on the reviewer. Therefore, having multiple reviewers can reduce potential bias when reviewing applications.

Second, following the call of others [16,99], we recommend that members of the admissions committee should be of diverse backgrounds and representative of the applicant pool. To accomplish that, departments might also consider adding non-tenure stream faculty, post-docs, and current graduate students to their admissions committees, providing appropriate recognition and compensation as necessary. Prior work has shown that faculty may prefer to admit applicants like themselves [23] and therefore, a representative admissions committee is needed to ensure that minoritized applicants are given equal consideration.

Finally, we recommend that departments conduct regular self-studies of their graduate admissions processes and share the results. While Rudolph *et al.* have previously called for departments to conduct self-studies of their admissions process [42], we believe it is equally important to share the results of those self-studies so that the physics community can know what is and what is not working. This collective knowledge of what is working and what is not

working can then be used by all to improve graduate admissions in physics for everyone.

For the sharing of results to be impactful however, the results must be easy to access and easy to understand. While individual departments could post their results on their websites, we believe doing so adds an extra layer of complexity and makes the results harder to access. Instead, we advocate for a centralized system to be created so that departments can easily report their data in a standardized way and practitioners can easily see and compare results across programs. Such a system could be maintained by professional societies such the American Physical Society or the American Institute of Physics, or other organizations. A system like this has been designed for research-based assessments [100], but to our knowledge, there exists no such system for graduate admissions.

However, when conducting such self-study of what is working well and what is not working well, it is important to consider the question of “working well for whom?” As Razack *et al.* note, “working well” depends on one’s social positioning [77] and therefore, a change that works well for applicants of one background may not be working for applicants of a different background. By considering the “for whom?” the physics community can ensure that changes made are for the benefit of all rather than as new methods to continue the existing exclusionary practices in graduate admissions.

## IX. CONCLUSION

In this paper, we demonstrated that rubric-based admissions are a promising avenue for increasing equity in graduate admissions. We showed that faculty ratings of applicant’s grades, research experiences, and noncognitive abilities do not differ based on the applicant’s sex or undergraduate background. The differences we did observe in faculty ratings could be explained as observing known systematic issues in physics regarding test scores and service work expectations.

Based on the results of this study, we recommend that departments use rubric-based holistic review for their graduate admissions process. Multiple people should review each application and those people should be representative of the applicant pool to limit any bias in the review process. Finally, departments should engage in self-study to see how their graduate admissions process is working and share those results so that the physics community can collectively learn what is working and what is not working in making graduate admissions more equitable.

## ACKNOWLEDGMENTS

We would like to thank Nicole Verboncoeur and Tabitha Hudson for compiling the data in this project. This project was supported by the Michigan State University College of Natural Sciences and the Lappan-Phillips Foundation.

- 
- [1] Patrick J. Mulvey, Starr Nicholson, and Jack Pold, *Trends in Physics PhDs*, Tech. Rep. (American Physical Society, 2021).
  - [2] Theodore Hodapp and Erika Brown, Making physics more inclusive, *Nature (London)* **557**, 629 (2018).
  - [3] The AIP National Task Force to Elevate African American Representation in Undergraduate Physics and Astronomy, *The Time is Now: Findings from TEAM-UP Report to Increase the Number of African Americans with Bachelor’s Degree in Physics and Astronomy*, Tech. Rep. (American Institute of Physics, New York, 2020).
  - [4] Brian J. Rybarczyk, Leslie Lerea, Dawayne Whittington, and Linda Dykstra, Analysis of postdoctoral training Outcomes that broaden participation in science careers, *CBE Life Sci. Educ.* **15**, ar33 (2016).
  - [5] Özlem Sensoy and Robin DiAngelo, “We are all for diversity, but ...”: How faculty hiring committees reproduce whiteness and Practical suggestions for how they can change, *Harv. Educ. Rev.* **87**, 557 (2017).
  - [6] Arri Eisen and Douglas C. Eaton, A model for postdoctoral education that promotes minority and majority success in the biomedical sciences, *CBE Life Sci. Educ.* **16**, ar65, 1 (2017).
  - [7] Needhi Bhalla, Strategies to improve equity in faculty hiring, *Molecular Biol. Cell* **30**, 2744 (2019).
  - [8] Michelle I. Cardel, Emily Dhurandhar, Ceren Yayar-Fisher, Monica Foster, Bertha Hidalgo, Leslie A. McClure, Sherry Pagoto, Nathaniel Brown, Dori Pekmezi, Noha Sharafeldin, Amanda L. Willig, and Christine Angelini, Turning chutes into ladders for women faculty: A review and roadmap for equity in academia, *J. Women’s Health* **29**, 721 (2020).
  - [9] Julie R. Posselt, *Equity in Science* (Stanford University Press, Palo Alto, CA, 2020).
  - [10] KerryAnn O’Meara, Dawn Culpepper, and Lindsey L. Templeton, Nudging toward diversity: Applying behavioral design to faculty hiring, *Rev. Educ. Res.* **90**, 311 (2020).
  - [11] Julie R. Posselt, Toward inclusive excellence in graduate education: Constructing merit and diversity in Ph.D. admissions, *Am. J. Educ.* **120**, 481 (2014).
  - [12] Julie Posselt, Theresa Hernandez, Geraldine Cochran, and Casey Miller, Metrics first, diversity later? Making the shortlist and getting admitted to physics Ph.D. programs, *J. Women Minorities Sci. Engin.* **25**, 283 (2019).
  - [13] Julia D. Kent and Maureen Terese McCarthy, *Holistic Review in Graduate Admissions: A Report from the Council of Graduate Schools* (Council of Graduate Students, Washington DC, 2016).

- [14] <http://igenetwork.org/>.
- [15] Casey Miller and Julie Posselt, Equitable admissions in the time of COVID-19, *Physics* **13**, 199 (2020).
- [16] Sonia F. Roberts, Elana Pyfrom, Jacob A. Hoffman, Christopher Pai, Erin K. Reagan, and Alysson E. Light, Review of racially equitable admissions practices in STEM doctoral programs, *Educ. Sci.* **11**, 270 (2021).
- [17] Poorna Talkad Sukumar and Ronald Metoyer, A visualization approach to addressing reviewer bias in holistic college admissions, in *Cognitive Biases in Visualizations*, edited by Geoffrey Ellis (Springer International Publishing, Cham, 2018), pp. 161–175.
- [18] Rachel E. Scherr, Monica Plisch, Kara E. Gray, Geoff Potvin, and Theodore Hodapp, Fixed and growth mindsets in physics graduate admissions, *Phys. Rev. Phys. Educ. Res.* **13**, 020133 (2017).
- [19] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.18.020140> for figures showing the rubric scores split by admission status and gender, undergraduate institution selectivity, and undergraduate physics program size, our admissions rubric, and our personal and academic statement prompts.
- [20] Jacqueline Doyle and Geoff Potvin, In search of distinct graduate admission strategies in physics: An exploratory study using topological data analysis, in *Proceedings of PERC Conf. 2015, College Park, MD* (2015), pp. 107–110, [10.1119/perc.2015.pr.022](https://doi.org/10.1119/perc.2015.pr.022).
- [21] Geoff Potvin, Deepa Chari, and Theodore Hodapp, Investigating approaches to diversity in a national survey of physics doctoral degree programs: The graduate admissions landscape, *Phys. Rev. Phys. Educ. Res.* **13**, 020142 (2017).
- [22] Deepa Chari and Geoff Potvin, Understanding the importance of graduate admissions criteria according to prospective graduate students, *Phys. Rev. Phys. Educ. Res.* **15**, 023101 (2019).
- [23] Julie R. Posselt, *Inside Graduate Admissions* (Harvard University Press, Cambridge, MA, 2016).
- [24] Nicholas T. Young and Marcos D. Caballero, Using machine learning to understand physics graduate school admissions, in *Proceedings of PER Conf. 2019, Provo, UT*, [10.1119/perc.2019.pr.Young](https://doi.org/10.1119/perc.2019.pr.Young).
- [25] Casey Miller and Keivan Stassun, A test that fails, *Nature (London)* **510**, 303 (2014).
- [26] Casey W. Miller, Benjamin M. Zwickl, Julie R. Posselt, Rachel T. Silvestrini, and Theodore Hodapp, Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion, *Sci. Adv.* **5**, eaat7550 (2019).
- [27] Nils J. Mikkelsen, Nicholas T. Young, and Marcos D. Caballero, Investigating institutional influence on graduate program admissions by modeling physics Graduate Record Examination cutoff scores, *Phys. Rev. Phys. Educ. Res.* **17**, 010109 (2021).
- [28] Geraldine L. Cochran, Theodore Hodapp, and Erika E. Alexander Brown, Identifying barriers to ethnic/racial minority students' participation in graduate physics, in *Proceedings of PER Conf. 2017, Cincinnati, OH*, [10.1119/perc.2017.pr.018](https://doi.org/10.1119/perc.2017.pr.018).
- [29] <https://www.ets.org/gre/test-takers/general-test/register/fees.html>.
- [30] <https://www.ets.org/gre/test-takers/subject-tests/register/fees.html>.
- [31] Lindsay M. Owens, Benjamin M. Zwickl, Scott V. Franklin, and Casey W. Miller, Physics GRE requirements create uneven playing field for graduate applicants, in *Proceedings of PER Conf. 2020, virtual conference*, [10.1119/perc.2020.pr.Owens](https://doi.org/10.1119/perc.2020.pr.Owens).
- [32] Kyle M. Whitcomb, Sonja Cwik, and Chandrekha Singh, Not all disadvantages are equal: Racial/ethnic minority students have largest disadvantage among demographic groups in both STEM and non-STEM GPA, *AERA Open* **7**, 1 (2021).
- [33] Kyle M. Whitcomb and Chandrekha Singh, Underrepresented minority students receive lower grades and have higher rates of attrition across STEM disciplines: A sign of inequity?, *Int. J. Sci. Educ.* **43**, 1054 (2021).
- [34] Stuart Rojstaczer and Christopher Healy, Where A is ordinary: The evolution of American college and university grading, 1940-2009, *Teachers College record* **114**, 1 (2012).
- [35] Mike Verostek, Casey W. Miller, and Benjamin Zwickl, Analyzing admissions metrics as predictors of graduate GPA and whether graduate GPA mediates Ph.D. completion, *Phys. Rev. Phys. Educ. Res.* **17**, 020115 (2021).
- [36] Catherine D'ignazio and Lauren F Klein, *Data Feminism* (MIT Press, Cambridge, MA, 2020).
- [37] <https://www.ets.org/gre/institutions/admissions/planning/benefits/>.
- [38] Sang Eun Woo, James M. LeBreton, Melissa G. Keith, and Louis Tay, Bias, fairness, and validity in graduate-school admissions: A psychometric perspective, *Perspect. Psychol. Sci.*, [174569162110553](https://doi.org/10.1177/174569162110553) (2022).
- [39] Sandra L. Petersen, Evelyn S. Erenrich, Dovev L. Levine, Jim Vigoreaux, and Krista Gile, Multi-institutional study of GRE scores as predictors of STEM Ph.D. degree completion: GRE gets a low mark, *PLoS One* **13**, e0206570 (2018).
- [40] Joshua D. Hall, Anna B. O'Connell, and Jeanette G. Cook, Predictors of student productivity in biomedical graduate school applications, *PLoS One* **12**, e0169121 (2017).
- [41] Linda Sealy, Christina Saunders, Jeffrey Blume, and Roger Chalkley, The GRE over the entire range of scores lacks predictive ability for PhD outcomes in the biomedical sciences, *PLoS One* **14**, e0201634 (2019).
- [42] Alexander Rudolph, Gibor Basri, Marcel Agüeros, Ed Bertschinger, Kim Coble, Meghan Donahue, Jackie Monkiewicz, Angela Speck, Keivan Stassun, Rachel Ivie, Christine Pfund, and Julie Posselt, Final Report of the 2018 AAS task force on diversity and inclusion in astronomy graduate education, *Bull. AAS* **51**, 1 (2020), <https://baas.aas.org/pub/2019i0101>.
- [43] Carrie Hawkins, The impact of a holistic admissions review process in a doctor of physical therapy program, Graduate Theses, Dissertations, and Capstones, Bellarmine University, 2020, <https://scholarworks.bellarmino.edu/tdc/85/>.



- [44] Annie M. Francis, L. B. Klein, Sharon Holmes Thomas, Kirsten Kainz, and Amy Blank Wilson, Holistic admissions and racial/ethnic diversity: A systematic review and implications for social work doctoral education, *J. Social Work Educ.* **58**, 227 (2021).
- [45] Marena A. Wilson, Max A. Odem, Taylor Walters, Anthony L. DePass, and Andrew J. Bean, A model for holistic review in graduate admissions that decouples the GRE from race, ethnicity, and gender, *CBE Life Sci. Educ.* **18**, ar7, 1 (2019).
- [46] Wendy I. Pacheco, Richard J. Noel, James T. Porter, Caroline B. Appleyard, and Hannah Sevian, Beyond the GRE: Using a composite score to predict the success of puerto rican students in a biomedical PhD program, *CBE Life Sci. Educ.* **14**, ar13 (2015).
- [47] Blaire Lauren Moody Rideout, A Study of the Inter-Rater Reliability of University Application Readers in a Holistic Admissions Review Process, Ph.D. thesis, Bowling Green State University (2017).
- [48] Tim Kautz, James J Heckman, Ron Diris, Bas Ter Weel, and Lex Borghans, *Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success* (National Bureau of Economic Research, Cambridge, MA, 2014).
- [49] Brent W. Roberts, Back to the future: Personality and assessment and personality development, *Journal Res. Pers.* **43**, 137 (2009).
- [50] Mathilde Almlund, Angela Lee Duckworth, James Heckman, and Tim Kautz, Personality psychology and economics, in *Handbook of the Economics of Education* (Elsevier, New York, 2011), Vol. 4, pp. 1–181.
- [51] W. E. Sedlacek, Noncognitive Measures for Higher Education Admissions, in *International Encyclopedia of Education*, 3rd ed., edited by Penelope Peterson, Eva Baker, and Barry McGaw (Elsevier, Oxford, 2010), pp. 845–849.
- [52] Lisa G. Smithers, Alyssa C.P. Sawyer, Catherine R. Chittleborough, Neil M. Davies, George Davey Smith, and John W. Lynch, A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes, *Nature Hum. Behav.* **2**, 867 (2018).
- [53] Pamela Scorza, Ricardo Araya, Alice J. Wuermli, and Theresa S. Betancourt, Towards clarity in research on “non-cognitive” skills: Linking executive functions, self-regulation, and economic development to advance life outcomes for children, Adolescents and youth globally, *Hum. Dev.* **58**, 313 (2016).
- [54] Terence J. Tracey and William E. Sedlacek, Noncognitive variables in predicting academic success by race, *Meas. Eval. Guid.* **16**, 171 (1984).
- [55] Terence J. Tracey and William E. Sedlacek, Prediction of college graduation using noncognitive variables by race, *Meas. Eval. Counseling Develop.* **19**, 177 (1987).
- [56] Niki Medrinos, Beyond the SAT/ACT: An Examination of Non-Cognitive Factors That Contribute to Students’ College Success, Ph.D. thesis, Temple University (2014).
- [57] Stephen Carp, Kyle Fry, Brittany Gumerman, Kevin Pressley, and Alyssa Whitman, Relationship between grit scale score and academic performance in a doctor of physical therapy program: A case study, *J. Allied Health* **49**, 29 (2020).
- [58] Scott K. Stolte, Stephanie B. Scheer, and Evan T. Robinson, The reliability of non-cognitive admissions measures in predicting non-traditional doctor of pharmacy student performance outcomes, *Am. J. Pharm. Educ.* **67**, 18 (2003).
- [59] Kristin Zakariassen Victoroff and Richard E. Boyatzis, What is the relationship between emotional intelligence and dental student clinical Performance?, *J. Dental Educ.* **77**, 416 (2013).
- [60] Christopher Peskun, Allan Detsky, and Maureen Shandling, Effectiveness of medical school admissions criteria in predicting residency ranking four years later, *Med. Educ.* **41**, 57 (2007).
- [61] Jay Burmeister, Erin McSpadden, Joseph Rakowski, Adrian Nalichowski, Mark Yudelev, and Michael Snyder, Correlation of admissions statistics to graduate student success in medical physics, *J. Appl. Clinical Med. Phys.* **15**, 375 (2014).
- [62] Chan Kulatunga Moruzi and Geoffrey R. Norman, Validity of admissions measures in predicting performance outcomes: The contribution of cognitive and non-cognitive dimensions, *Teach. Learn. Med.* **14**, 34 (2002).
- [63] Melissa C. O’Connor and Sampo V. Paunonen, Big five personality predictors of post-secondary academic performance, *Personality Indiv. Diff.* **43**, 971 (2007).
- [64] Patrick Kyllonen, Alyssa M. Walters, and James C. Kaufman, Noncognitive constructs and their assessment in graduate education: A review, *Educ. Assess.* **10**, 153 (2005).
- [65] Robert E. Ployhart and Brian C. Holtz, The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection, *Personnel psychology* **61**, 153 (2008).
- [66] William E. Sedlacek, Why we should use noncognitive variables with graduate and professional students, 2001.
- [67] Casey W. Miller, Using non-cognitive assessments in graduate admissions to select better students and increase diversity, *Status* (American Astronomical Society, 2015), [https://aas.org/sites/default/files/2019-09/Status2015\\_Jan\\_s.pdf](https://aas.org/sites/default/files/2019-09/Status2015_Jan_s.pdf).
- [68] Lindsay Owens, Benjamin M. Zwickl, Scott V. Franklin, and Casey W. Miller, Identifying qualities of physics graduate students valued by faculty, in *Proceedings of PER Conf. 2019, Provo, UT*, 10.1119/perc.2019.pr.Owens.
- [69] Yuanyuan Chen, Shuaizhang Feng, James J. Heckman, and Tim Kautz, Sensitivity of self-reported noncognitive skills to survey administration conditions, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 931 (2020).
- [70] Equity in Graduate Education—Non-Cognitive Assessment (2021), <https://equitygraded.org/non-cognitive-assessment/>.
- [71] Penny Salvatori, Reliability and validity of admissions tools used to select students for the health professions, *Adv. Health Sci. Educ.* **6**, 159 (2001).
- [72] Jacqueline M. Zeeman, Jacqueline E. McLaughlin, and Wendy C. Cox, Validity and reliability of an application



- review process using dedicated reviewers in one stage of a multi-stage admissions model, *Curr. Phar.Teach.Learn.* **9**, 972 (2017).
- [73] Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman, Science faculty's subtle gender biases favor male students, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16474 (2012).
- [74] Asia A. Eaton, Jessica F. Saunders, Ryan K. Jacobson, and Keon West, How gender and race stereotypes impact the advancement of scholars in STEM: Professors' biased evaluations of physics and biology post-doctoral candidates, *Sex Roles* **82**, 127 (2020).
- [75] Nicolás E. Barceló, Sonya Shadravan, Christine R. Wells, Nichole Goodsmith, Brittany Tarrant, Trevor Shaddox, Yvonne Yang, Eraka Bath, and Katrina DeBonis, Reimagining merit and representation: Promoting equity and reducing bias in GME through holistic review, *Academic Psych.* **45**, 34 (2021).
- [76] David M. Quinn, Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales, *Educ. Eval. Policy Anal.* **42**, 375 (2020).
- [77] Saleem Razack, Torsten Risør, Brian Hodges, and Yvonne Steinert, Beyond the cultural myth of medical meritocracy, *Med. Educ.* **54**, 46 (2020).
- [78] Keivan G. Stassun, Susan Sturm, Kelly Holley-Bockelmann, Arnold Burger, David J. Ernst, and Donna Webb, The Fisk-Vanderbilt Master's-to-Ph.D. Bridge Program: Recognizing, enlisting, and cultivating unrealized or unrecognized potential in underrepresented minority students, *Am. J. Phys.* **79**, 374 (2011).
- [79] This category was not used in 2021 and will not be used in 2022 due to the impacts of COVID-19 on students' ability to take these tests.
- [80] Julie R. Posselt, Trust networks: A new perspective on pedigree and the ambiguities of admissions, *Rev. High. Educ.* **41**, 497 (2018).
- [81] Starr Nicholson and Patrick J. Mulvey, *Roster of Physics Departments with Enrollment and Degree Data, 2017*, Tech. Rep. (American Institute of Physics, New York, 2018).
- [82] Starr Nicholson and Patrick J. Mulvey, *Roster of Physics Departments with Enrollment and Degree Data, 2018*, Tech. Rep. (American Institute of Physics, New York, 2019).
- [83] Starr Nicholson and Patrick J. Mulvey, *Roster of Physics Departments with Enrollment and Degree Data, 2019*, Tech. Rep. (American Institute of Physics, New York, 2020).
- [84] Sture Holm, A simple sequentially rejective multiple test procedure, *Scand. J. Stat. Theory Appl.* **6** (1979), <https://www.jstor.org/stable/4615733>.
- [85] Jayson Nissen, Robin Donatello, and Ben Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, *Phys. Rev. Phys. Educ. Res.* **15**, 020106 (2019).
- [86] Cassandra M. Guarino and Victor M. H. Borden, Faculty service loads and gender: Are women taking care of the academic family?, *Res. High. Educ.* **58**, 672 (2017).
- [87] Bureau of Labor Statistics, Volunteering in the United States, 2015, Tech. Rep. USDL 16-0363 (2016).
- [88] Nicholas T. Young and Marcos D. Caballero, Physics graduate record exam does not help applicants "stand out", *Phys. Rev. Phys. Educ. Res.* **17**, 010144 (2021).
- [89] Charles D Brown III, Commentary: Disentangling anti-Blackness from physics, *Phys. Today* (2020).
- [90] Katemari Rosa and Felicia Moore Mensah, Educational pathways of Black women physicists: Stories of experiencing and overcoming obstacles in life, *Phys. Rev. Phys. Educ. Res.* **12**, 020113 (2016).
- [91] Paul H. Barber, Tyrone B. Hayes, Tracy L. Johnson, and Leticia Márquez-Magaña, Systemic racism in higher education, *Science* **369**, 1440 (2020).
- [92] Danielle Dickens, Maria Jones, and Naomi Hall, Being a token black female faculty member in physics: Exploring research on gendered racism, identity shifting as a coping strategy, and inclusivity in physics, *Phys. Teach.* **58**, 335 (2020).
- [93] Chanda Prescod-Weinstein, Making black women scientists under white empiricism: The racialization of epistemology in physics, *Signs: J. Women Culture Soc.* **45**, 421 (2020).
- [94] Kelly Ochs Rosinger, Karly Sarita Ford, and Junghee Choi, The role of selective college admissions criteria in interrupting or reproducing racial and economic inequities, *J. Higher Educ.* **92**, 31 (2020).
- [95] David Kalsbeek, Michele Sandlin, and William Sedlacek, Employing noncognitive variables to improve admissions, and increase student diversity and retention, *Strategic Enrollment Management Quart.* **1**, 132 (2013).
- [96] La'Tonia Stiner-Jones and Wolfgang Windl, Work in Progress: Aligning What We Want With What We Seek: Increasing Comprehensive Review in the Graduate Admissions Process, in Proceedings of the 2019 ASEE Annual Conference & Exposition, Tampa, Florida (2020), 10.18260/1-2-33592.
- [97] Shannon Barker and Amy Clobes, Work in Progress: A Holistic PhD Admissions Rubric-Design & Implementation, in *Proceedings of the 2021 ASEE Virtual Annual Conference Content Access* (ASEE Conferences, 2021).
- [98] Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland, Expectations of brilliance underlie gender distributions across academic disciplines, *Science* **347**, 262 (2015).
- [99] Quinn Capers, Leon McDougale, and Daniel M. Clinchot, Strategies for achieving diversity through medical school admissions, *J. Health Care Poor Underser.* **29**, 9 (2018).
- [100] Jayson M. Nissen, Manher Jariwala, Eleanor W. Close, and Ben Van Dusen, Participation and performance on paper- and computer-based low-stakes assessments, *Int. J. STEM Educ.* **5**, 21 (2018).