Gender differences in grades versus grade penalties: Are grade anomalies more detrimental for female physics majors?

Alysa Malespina* and Chandralekha Singh†

Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA

(Received 19 January 2022; accepted 11 July 2022; published 21 October 2022)

Creating equitable learning environments has been an area of significant focus for physics education researchers in recent years. Here we introduce a framework that posits that grade penalty is a measure of academic self-concept and investigate if there are gender differences in grade penalties in physics courses for students majoring in physics. In order to quantify grade penalty, we define grade anomaly as the difference between a student’s grade in a course under consideration and their grade point average (GPA) in all other classes thus far. A grade anomaly lower than students expected grade based on their GPA is a grade penalty and higher than expected average grade is a grade bonus. Our framework posits that since women have traditionally been marginalized in physics, female physics majors are more likely to be negatively impacted by a grade penalty in their courses since their academic self-concept as a physics major hinges on them securing a certain grade. In the study presented here, we examine the average grade anomalies across a number of courses for female and male physics majors. We find that these students received grade penalties in almost all physics courses studied, though there were grade bonuses in a few laboratory courses. We also find that in physics courses, on average, women often had larger grade penalties than men, especially in introductory courses. We hypothesize that, because their grade penalties are often larger than men’s, women’s decisions to pursue a physics major and career may be particularly affected by grade penalties received in their various courses. Furthermore, the grade penalty measure can be easily computed by the physics programs concerned with equity.

DOI: 10.1103/PhysRevPhysEducRes.18.020127

I. INTRODUCTION

In recent years, physics education researchers have been particularly focused on creating equitable learning environments, which is particularly important for underrepresented students such as women [1–8], racial and ethnic minority students [7–10], and students with disabilities [11,12]. Here, we focus on women majoring in physics because they are more drastically underrepresented than women in many other science, technology, engineering, and mathematics (STEM) disciplines. Prior research has explored gender differences in performance and persistence in physics and other STEM fields [13,14]. This research has explored a range of potential factors that may lead to such differences. Some factors that researchers have investigated include societal biases regarding who can be successful in physics [15,16], lack of encouragement from families and instructors [17–19], and motivational characteristics and attitudes towards physics learning [3–5].

Societal stereotypes about physics are still prevalent, and both practitioners and students often believe that success in physics requires natural ability. For example, physics researchers are more likely than those in other STEM disciplines to endorse that their subject required natural ability for success [15], and young people often hold beliefs that physics has always been seen as really hard, and you know you have to be so clever to understand it [20]. This type of belief that physics requires a particular ability may combine with common societal stereotypes that men and boys are more likely than women and girls to be extremely intelligent [21–23], to discourage women from believing they have what it takes to pursue physics.

Lack of encouragement from families may also discourage early interest in science and physics. For example, parents tend to rate boys abilities in math higher than girls [17,18], and are less likely to explain science to girls than boys while using interactive science exhibits at museums [24]. Because of these lack of early experiences and encouragement, girls may be less likely to develop an early interest in science. Similar lack of encouragement from instructors later in life may discourage women from participating in physics as well. For example, in one study

*alm417@pitt.edu
Also at Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA.
†clsingh@pitt.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.
faculty members in physics rated men as more competent than women with an identical curriculum vitae. If women are not receiving the same amount of confidence or encouragement from their instructors, they may be less compelled to pursue physics.

Finally, some research on gender differences in physics performance and persistence has focused on motivational beliefs and attitudes towards physics learning [3–5]. One such attitudinal construct is academic self-concept, which describes a long-term expectation of success that students hold regarding their academic abilities and that depends on outside feedback, such as grades [25–27]. Low academic self-concept may lead to lower future achievement and persistence because it discourages student engagement in a domain [25]. When women leave STEM disciplines, particularly physics, they often do so with higher grades than the men who remain in the program [28–30]. For example, a recent investigation shows that among students with the same STEM GPAs, women were more likely to leave the major, while men were more likely to earn a degree [28].

There are many potential partial explanations that have been suggested regarding why women who are meeting or exceeding the requirements of their programs leave. These include lack of role models, societal stereotypes and biases about who can excel in these disciplines [15,17,21,31–33], gender discrimination in hiring [19], and differences in STEM motivational beliefs such as self-efficacy [2–4,34–45], recognition from instructors and peers [46,47], intelligence mindset [48,49], and sense of belonging [50,51]. One related reason for why this happens may be lower academic self-concept of female students in these courses compared to male students. Though none of these factors may provide complete explanations of gender differences in physics, aiming to address them simultaneously may create a better learning environment for women in these programs.

Here, we focus on physics majors and inquire about gender differences in grade penalties. In order to quantify grade penalty, we define grade anomaly as the difference between a student’s grade in a course of interest and their grade point average (GPA) in all other classes up to that point. The mean of this statistic for all students who took a course is the average grade anomaly (AGA). We divide average grade anomalies into bonuses and penalties. A course in which students on average earn a lower grade than usual has an AGA with grade penalty, while a course in which students on average earn a higher grade than usual has an AGA with grade bonus.

Within our framework, we posit that grade anomaly may allow us to track, through institutional grade data, an important measure of how courses may affect students academic self-concept. Our framework uses grade penalty as a central construct instead of grade because students academic self-concept is often based on comparisons, not absolute grades [52]. Students may compare their grades across courses to determine which disciplines they excel at or struggle with [52]. Additionally, students tend to have a fairly fixed view of what kind of student they are, e.g., students may endorse the idea that “If I get A’s, I must be an A kind of person. If I get a C, I am a C kind of person” [30]. Grade anomalies may challenge or reinforce students’ ideas about what kind of student they are, and if they are capable of succeeding in their chosen major. Many students who leave STEM majors explicitly cite lower grades than they are used to as a reason for doing so [29,30]. Grade penalties are more common and extreme in STEM disciplines than in humanities or social science departments [30,53–55], and women tend to have larger grade penalties than men in many subjects, including physics [55].

In this paper, we use situated expectancy value theory (SEVT), studies about why students leave STEM, and previous work on grade anomalies to explore whether the average grade anomalies for male and female physics majors are different, making grade anomalies an equity issue in physics. We also posit that grade anomaly may be a better measure of self-concept [52] than raw grades because it is a unique measure of “within-student” frame of reference (i.e., students are comparing their own grades across different courses as opposed to comparing their grades with others) [26].

A. Research questions

We aim to answer the following research questions regarding average grade anomalies:

RQ1. For which of their courses do students majoring in physics, on average, receive a “grade penalty” and for which courses do they receive a “grade bonus”?

RQ2. To what extent do men and women have different AGAs in their physics courses?

RQ3. To what extent do gender differences in AGAs follow the same trends as gender differences in average grades?

B. Theoretical framework

Expectancy value theory (EVT) [25] and SEVT [52] are frameworks to understand student achievement, persistence, and choice of tasks in a domain (e.g., physics or chemistry). EVT posits that performance and persistence is determined by someone’s expectation of success and the extent to which they value that task. If a student expects they will succeed in a task and believes that the task will be valuable to them (for personal interest, as a path to achieve another goal, etc.) they are more likely to pursue that task. If they do not expect to succeed and do not value a task, they are unlikely to attempt it. Here, we will focus primarily on student expectancies, though value is also important to understanding why some students may persist while others do not. Expectancies are a combination of academic self-concept, expectations for success, and perceptions of task
difficulty [25–27,52]. Academic self-concept is the most stable and the least task- and domain-dependent of the three, and it is based primarily on grades and outside (e.g., from parents, peers, and instructors) feedback [25–27,52,56]. Grades can inform academic self-concept as both an external (“How good at math am I compared to other students?”) and internal (“How good am I at math compared to English?”) frame of reference [25–27].

Expectancies of success are domain and task specific, and refer to a student’s belief in their ability to complete a specific task; which will include considerations such as knowledge and skill related to the subject, time allotted, and experience in a subject [25–27,52]. Expectancy for success closely relates to Bandura’s theory of self-efficacy [27,52,57]. A student may have a positive academic self-concept in math, but may have low expectancy for success if they take a math test on very new material they have not had adequate time to learn. The third issue related to expectancy, perceptions of task difficulty, is more straightforward; most students have less faith in their ability to do well on an exam if their peers have reported it to be particularly difficult [25].

In EVT, the three expectancy concepts were collapsed into one factor. However, the updated framework, SEVT, has called for a separation of these three concepts [52]. According to Eccles and Wigfield [52], combining academic self-concept, expectancies for success, and perceived task difficulty has led to a lack of understanding of the unique developmental mechanisms of each and how the three concepts relate. We posit that grade anomaly may be a better measure of how students’ self-concept evolves [52] due to feedback about performance than raw grades. This is because students often judge their ability by comparing their grades across courses rather than comparing their grades to other students’ (in EVT or SVT, this is called the within-student frame of reference). Poor performance from a within-student frame of reference may cause students to question if they should continue in a discipline [52].

Grade anomalies allow us to measure how courses can affect students’ academic self-concept [27,52,53] using institutional grade data [55], which may be more accessible to instructors and researchers than surveys or interview data. While students’ raw grades are a useful measure, e.g., because they allow for direct comparison between students and because they are used by institutions to award scholarships and track student academic standing, we propose that using grade penalty in addition to raw grades gives researchers and instructors more insight into student self-concept (that is, a students’ view of what “kind of student they are”).

A student who receives lower grades in their science courses than their humanities courses may take this as a sign that they are not capable of excelling in the sciences, even if the grades they earn are high enough for them to continue in their major [29,30]. This experience may be common, because grade penalties tend to be more extreme and widespread in STEM disciplines than in other subjects [30,54,58,59]. Women may also have fewer resources available to cope with these grade penalties than men, in part due to lack of role models and societal stereotypes about who belongs in these disciplines and can excel in them [16,58,60]. When students leave STEM fields, they often list lower grades than expected as a reason [29,30], and women tend to leave these fields with higher grades than men [28]. Larger grade penalties may be one potential reason for this discrepancy.

Several studies [54,55,61] have utilized “grade anomaly” or grade penalty, the difference between a student’s GPA excluding a course of interest and their grades in all courses thus far. Koe eter al. [54] conducted the first study we know of that focuses on average grade anomaly. They used AGA because it was perceived to be a better measure of how students view their comparative performance than their raw grades across different courses. They found that, at their institution, grade penalties were greater for STEM than non-STEM courses. Further, within STEM courses, grade penalties were smaller for men than women. In particular, they found that physics courses had the largest grade penalty and largest gender difference in AGA. The researchers theorized that large grade penalties and gender differences may be partially attributed to high-stakes assessments [62–66] and stereotype threat [67]. The Matz et al. [55] study had similar findings but with a larger student sample across multiple institutions. Across five universities, STEM courses had larger grade penalties and larger gender differences in AGA that usually favored men. Their study also raised concerns over high-stakes assessments. They emphasized that large grade anomalies often reflect grading decisions made by instructors (for example, choosing high-stake assessments may increase gendered grade differences [62,64,66,68]), rather than being an accurate measure of student learning.

Additionally, Witteveen and Attewell [61] found that having lower STEM GPAs than overall GPAs during the first two semesters of university were negatively correlated with completing a STEM degree, even when controlling for gender, race, high school preparation, and college performance. This was not the case for courses taken later in students’ college career, which speaks to the importance of introductory courses in student retention. Past research has found that during times of transition, the usually stable academic self-concept becomes more dependent on grade feedback and less dependent on outsider (e.g., parental) feedback [26]. These findings hint at the importance of monitoring and reducing grade penalties in students’ first few semesters.

Thus, past work provides evidence for the existence of average grade penalties in many STEM courses, and
the existence of gender differences in these anomalies. Here, we present an investigation that focuses on average grade anomaly in various courses for physics majors. We analyze data to study if these trends hold in a more homogeneous population of students in the same major at a large university in the U.S., rather than combining students across institutions and many majors. We hypothesize that grade anomalies, e.g., grade penalty, discussed here can negatively influence students internal frame of reference.

II. METHODOLOGY

A. Participants and procedures

Participants in this study were students who declared a physics major at any time during the thirteen-year duration of the study. All participants attended the University of Pittsburgh, which is a large, public, urban institution. Thus, our sample included both students who graduated with a physics major and students who declared a physics major but later switched to another field. We excluded summer introductory courses they are not a typical representation of courses at our institution. For example, many summer students do not primarily attend our institution, but are local students visiting home for the summer. In addition, summer courses are typically taught by graduate students, the class sizes are an order of magnitude smaller than those in the fall and spring semesters, and many students work full time while taking an introductory course which is not as common during the fall and spring semesters. We had a total of 671 students who took 23,154 classes (including 5,713 physics courses). The sample consisted of 23.2% women and 76.8% men. All students in the sample selected one of these binary gender options. Students identified with the following races or ethnicities: 79% White, 9% Asian, 3% Hispanic or Latinx, 4% multiracial, 2% African American or Black, and 3% unknown or unspecified. This research was carried out in accordance with the principles outlined in the University of Pittsburgh Institutional Review Board (IRB) ethical policy, and deidentified demographic data were provided through university records.

We chose the twenty most common physics courses taken by students in our sample, many of which are mandatory for physics majors. All the courses we studied are listed in Table I, along with information about the year in which the students typically take the course. The total number of students who took each course, as well as the gender distribution for each course can be found in Table II. Because of changing course requirements over time (for example, Quantum Mechanics 1 was optional for the first three years of the study) and the tendency of students to leave the physics major over time [29,30,69,70], more students took introductory than advanced courses.

<table>
<thead>
<tr>
<th>Course</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>≥ 5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics 1</td>
<td>74</td>
<td>19</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Physics 2</td>
<td>54</td>
<td>34</td>
<td>9</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Physics 1 Honors</td>
<td>87</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Physics 2 Honors</td>
<td>83</td>
<td>9</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Introductory Lab</td>
<td>17</td>
<td>58</td>
<td>15</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Intro to Astronomy</td>
<td>35</td>
<td>41</td>
<td>15</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Hon Introductory Lab</td>
<td>4</td>
<td>75</td>
<td>13</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Modern Physics 1</td>
<td>6</td>
<td>49</td>
<td>27</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Modern Physics 2</td>
<td>1</td>
<td>44</td>
<td>23</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>Electronics Lab</td>
<td>2</td>
<td>43</td>
<td>32</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>Modern Physics Lab</td>
<td>0</td>
<td>32</td>
<td>44</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Thermo and Stat Mech</td>
<td>2</td>
<td>10</td>
<td>42</td>
<td>32</td>
<td>14</td>
</tr>
<tr>
<td>Mechanics</td>
<td>4</td>
<td>30</td>
<td>41</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>E&amp;M 1</td>
<td>2</td>
<td>14</td>
<td>53</td>
<td>22</td>
<td>9</td>
</tr>
<tr>
<td>E&amp;M 2</td>
<td>0</td>
<td>7</td>
<td>42</td>
<td>34</td>
<td>17</td>
</tr>
<tr>
<td>Comp. Methods</td>
<td>2</td>
<td>13</td>
<td>44</td>
<td>25</td>
<td>16</td>
</tr>
<tr>
<td>Optics</td>
<td>0</td>
<td>4</td>
<td>29</td>
<td>47</td>
<td>20</td>
</tr>
<tr>
<td>Optics Writing</td>
<td>1</td>
<td>12</td>
<td>28</td>
<td>42</td>
<td>17</td>
</tr>
<tr>
<td>Quantum Mechanics 1</td>
<td>0</td>
<td>4</td>
<td>25</td>
<td>50</td>
<td>21</td>
</tr>
<tr>
<td>Quantum Mechanics 2</td>
<td>0</td>
<td>4</td>
<td>25</td>
<td>49</td>
<td>22</td>
</tr>
</tbody>
</table>

B. Measures

1. Course grade

Course grades were based on the 0–4 scale used at our university, with A = 4, B = 3, C = 2, D = 1, F = 0, or W (late withdrawal), where the suffixes + and −, respectively, add or subtract 0.25 grade points (e.g., B+ = 2.75 and B− = 3.25), except for the A+, which is reported as 4. We are unable to report detailed grading schemes of each physics instructor, type of course, or any other detailed course-level information but a majority of courses are traditionally taught primarily using lectures.

2. Grade anomaly

Grade anomaly (GA) was found by first finding each students grade point average excluding the course of interest (GPAexc), including all courses taken prior and simultaneously with the course of interest. This was done by using the equation

\[ \text{GPA}_{\text{exc}} = \frac{(\text{GPA}_c \times \text{Units}_c) - (\text{Grade} \times \text{Units})}{\text{Units}_c - \text{Units}}, \]

where GPAc is the student’s cumulative GPA, Unitsc is the cumulative number of units the student has taken, Grade is...
TABLE II. Mean and standard deviation (SD) of average grade anomalies and grades, number of students (N) for each course of interest, and the percentage of women that took the course. Courses are also marked as required or optional, which is accurate as of the final semester of data collection. Optional courses are typically chosen from a group of electives to fulfill degree requirements, though no specific course is mandatory. We used the following abbreviations: Laboratory (Lab) and Statistical Mechanics (Stat Mech).

<table>
<thead>
<tr>
<th>Course</th>
<th>Course type</th>
<th>% Women (%)</th>
<th>N</th>
<th>AGA Mean</th>
<th>AGA SD</th>
<th>Grade Mean</th>
<th>Grade SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics 1</td>
<td>Required b</td>
<td>23</td>
<td>335</td>
<td>−0.36</td>
<td>0.92</td>
<td>2.89</td>
<td>0.92</td>
</tr>
<tr>
<td>Physics 2</td>
<td>Required b</td>
<td>26</td>
<td>389</td>
<td>−0.39</td>
<td>0.86</td>
<td>2.87</td>
<td>0.94</td>
</tr>
<tr>
<td>Physics 1 Honors</td>
<td>Required b</td>
<td>25</td>
<td>198</td>
<td>−0.18</td>
<td>1.06</td>
<td>3.31</td>
<td>0.77</td>
</tr>
<tr>
<td>Physics 2 Honors</td>
<td>Required b</td>
<td>21</td>
<td>183</td>
<td>−0.11</td>
<td>0.54</td>
<td>3.43</td>
<td>0.70</td>
</tr>
<tr>
<td>Introductory Lab</td>
<td>Required b</td>
<td>23</td>
<td>371</td>
<td>0.27</td>
<td>0.71</td>
<td>3.49</td>
<td>0.82</td>
</tr>
<tr>
<td>Introduction to Astronomy</td>
<td>Optional</td>
<td>23</td>
<td>151</td>
<td>−0.04</td>
<td>0.88</td>
<td>3.10</td>
<td>0.99</td>
</tr>
<tr>
<td>Honors Introductory Lab</td>
<td>Required b</td>
<td>15</td>
<td>116</td>
<td>−0.02</td>
<td>0.58</td>
<td>3.53</td>
<td>0.69</td>
</tr>
<tr>
<td>Modern Physics 1</td>
<td>Required a</td>
<td>19</td>
<td>315</td>
<td>−0.50</td>
<td>1.03</td>
<td>2.84</td>
<td>1.08</td>
</tr>
<tr>
<td>Modern Physics 2</td>
<td>Optional</td>
<td>19</td>
<td>286</td>
<td>−0.28</td>
<td>0.71</td>
<td>3.04</td>
<td>0.93</td>
</tr>
<tr>
<td>Electronics Lab</td>
<td>Optional</td>
<td>22</td>
<td>241</td>
<td>−0.29</td>
<td>0.66</td>
<td>3.00</td>
<td>0.89</td>
</tr>
<tr>
<td>Modern Physics Lab</td>
<td>Optional</td>
<td>19</td>
<td>125</td>
<td>0.13</td>
<td>0.29</td>
<td>3.46</td>
<td>0.63</td>
</tr>
<tr>
<td>Thermodynamics Stat Mech</td>
<td>Required</td>
<td>21</td>
<td>219</td>
<td>−0.52</td>
<td>−0.78</td>
<td>2.88</td>
<td>1.05</td>
</tr>
<tr>
<td>Mechanics</td>
<td>Required</td>
<td>20</td>
<td>291</td>
<td>−0.53</td>
<td>0.70</td>
<td>2.83</td>
<td>0.93</td>
</tr>
<tr>
<td>Electricity and Magnetism 1</td>
<td>Required</td>
<td>20</td>
<td>318</td>
<td>−0.64</td>
<td>0.83</td>
<td>2.68</td>
<td>1.08</td>
</tr>
<tr>
<td>Electricity and Magnetism 2</td>
<td>Optional</td>
<td>23</td>
<td>100</td>
<td>−0.25</td>
<td>0.62</td>
<td>3.33</td>
<td>0.80</td>
</tr>
<tr>
<td>Computational Methods</td>
<td>Required</td>
<td>19</td>
<td>272</td>
<td>−0.34</td>
<td>1.00</td>
<td>2.95</td>
<td>1.24</td>
</tr>
<tr>
<td>Optics</td>
<td>Optional</td>
<td>20</td>
<td>196</td>
<td>−0.41</td>
<td>0.81</td>
<td>2.90</td>
<td>1.06</td>
</tr>
<tr>
<td>Optics Writing</td>
<td>Optional</td>
<td>25</td>
<td>166</td>
<td>−0.04</td>
<td>0.78</td>
<td>3.30</td>
<td>0.98</td>
</tr>
<tr>
<td>Quantum Mechanics 1</td>
<td>Required</td>
<td>19</td>
<td>204</td>
<td>−0.17</td>
<td>0.81</td>
<td>3.27</td>
<td>0.99</td>
</tr>
<tr>
<td>Quantum Mechanics 2</td>
<td>Optional</td>
<td>23</td>
<td>103</td>
<td>−0.15</td>
<td>0.59</td>
<td>3.41</td>
<td>0.74</td>
</tr>
</tbody>
</table>

*Students may take either Physics 1 and 2 or Physics 1 and 2 Honors.
*Students may take either Introductory Lab or Honors Introductory Lab.

III. RESULTS

A. For which of their courses do physics students receive a grade penalty and for which courses do they receive a grade bonus?

To answer RQ1, we calculated average grade anomaly for our courses of interest. We show the descriptive statistics for both grades and AGA in Table II. On average, students received grade penalties in most of their courses. However, two courses tended to give students grade bonuses: Introductory Physics Lab and Modern Physics Lab. Three courses gave neither a grade penalty nor a grade bonus (i.e., the AGA of the course was within 1 standard error of 0): Optics Writing, Honors Introductory Lab, and Introduction to Astronomy. The courses that gave students grade bonuses or no grade anomalies were all lab courses except Astronomy and the Optics wiring practicum.

Four courses gave students particularly large grade penalties (≤ −0.50): Modern Physics 1, Thermodynamics and Statistical Mechanics, Mechanics, and Electricity and Magnetism 1. This means that a student taking one of these courses can expect to receive half a letter grade lower in these courses than they do on average. Notably, the courses with the largest grade penalties are mandatory to complete (see Table II).

The courses in Table II do not average to an AGA of zero because student sample changes over time due to students who leave the major and because not all courses students take are included. For example, students may receive grade bonuses in general education courses which students can select from several hundred courses.

B. To what extent do men and women have different AGAs in their physics courses?

To investigate if there are differences in grade anomalies for men and women, we performed four multivariate
analyses of variance (MANOVA) with courses grouped to reduce listwise deletion into courses typically taken in students’ first, second, third, and fourth years in the physics program, which can be seen in Table III.

Women and men had statistically significantly different outcomes in their first-year courses, as shown by the MANOVA analysis in Table III. Figure 2 displays the mean and standard error of both men’s and women’s AGAs in first-year courses. From Fig. 1, one can see that there is a distinguishable difference between men’s and women’s AGAs for each of the first-year courses, and that men tended to have smaller grade penalties than women. The largest gender differences (shown in Fig. 1 and Table III) tend to be in the courses that students take earliest in their time as physics students: Physics 1 and Physics 1 Honors. Thus, it is not surprising that first-year courses show a statistically significant gender difference in Table II.

However, Table II also shows that there are no statistically significant gender differences for courses taken primarily by second-, third-, and fourth-year students. This finding is supported by Fig. 1, which shows that women had indistinguishable AGA outcomes to men in most individual courses, including Honors Introductory Physics Lab, Electronics Lab, Introduction to Astronomy, Modern Physics 1, Modern Physics 2, Modern Physics Lab, Thermodynamics and Statistical Mechanics, Optics, Optics Writing, Electricity and Magnetism 1, Electricity and Magnetism 2, Computational Methods, and Quantum Mechanics 1.

Though these courses are not included in any course groups that show significant AGA gender differences, we note that women had favorable AGA outcomes (e.g., smaller grade penalties or larger grade bonuses) compared to men in two courses: Introductory Physics Lab and Quantum Mechanics 2.

### C. To what extent do gender differences in AGAs follow the same trends as gender differences in average grades?

Next, we explore if grade anomaly and raw grades can provide different information. That is, does calculating AGA reveal additional trends beyond what raw grades can provide? Table III shows that the only group of courses that has a statistically significant gendered AGA difference is courses taken primarily by first-year students. Similarly, Table III shows that the only group of courses that has a statistically significant gendered grade difference is also courses taken primarily by first-year students. However,
it is important to note that the gender difference is larger for AGA than raw grades.

This trend is further supported by Figs. 1 and 2. These figures show that the gendered grade differences appear to be larger (i.e., their standard errors are further from overlapping) for AGAs than for grades, especially for first-year courses such as Physics 1 and Physics 1 Honors. This trend is also shown by the between-gender effect sizes listed in Table III: Cohen’s $d$ [71] is larger ($\Delta d = 0.18$) for AGA than grade for Physics 1, which is most students’ introduction to the major.

Beyond first-year courses, there are a few differences in effect size between some courses, which are shown in Table III. For example, both Honors Introductory Lab ($\Delta d = 0.20$) and Modern Physics Lab ($\Delta d = 0.15$) appear to strongly favor women in terms of raw grades, but this effect is smaller for AGAs. Similarly, Modern Physics Lab appears to have favorable outcomes for women in raw grades (seen in Fig. 2), but not AGAs (seen in Fig. 1).

Importantly, there are some additional trends revealed in AGA that cannot be seen with grade data alone.

IV. DISCUSSION

Our results show that there are grade penalties in the majority of courses studied. First, we discuss why grade penalties can potentially be harmful. It is important to note that students at the University of Pittsburgh do not declare their major until the end of their second year, so we are unable to track students who decided to leave the physics discipline before their third year. However, lower than expected grades, even in one course, can be a catalyst for students to leave STEM majors [29,30]. This does not just include D and F grades or withdrawal from the course, but grades that were high enough to continue in the program that did not meet a student’s personal expectations [29,30]. This can especially be an important issue among high-achieving students, who are more likely to endorse...
perfectionism and feeling that their identity as “good STEM students” is threatened by B’s and C’s, or even a low grade on a single exam [30]. Thus, we hypothesize that the courses with largest grade penalties, in this case Modern Physics 1, Thermodynamics and Statistical Mechanics, Mechanics, and Electricity and Magnetism 1, are the courses that are more likely to discourage most students from continuing in physics.

In addition to seeing evidence of grade penalties in some physics courses, we also see evidence of gender differences in average grade anomalies in over half of the courses studied, particularly Physics 1, Physics 2, and Physics 1 Honors. We find these introductory courses to be particularly concerning. Our research shows that women have statistically significantly lower grades and AGAs than men in first-year courses (see Table III), which has the potential to affect women’s academic self-concept more than other courses and they are taken during students’ first year at the university when their academic self-concept is in major flux [26]. Additionally, prior research suggests that low STEM GPAs during a student’s first year are correlated with lower degree completion [61]. Because women leave majors with higher grades than men who remain both in [28,30] and outside [58] of STEM, this raises serious equity concerns. Past work also suggests that women tend to have lower self-efficacy and sense of belonging that relate to academic self-concept [3,5,37,62,72,73]. Women report feeling more demoralized than men when they receive lower grades than expected, and cite more worry about not understanding the material even if they receive A’s, B’s, or C’s (all of which are grades that allow students to continue in most programs) [30,74]. This trend can be particularly strong among high-achieving women [30].

We hypothesize that women may be more likely to have a low academic-self-concept than men at similar performance levels for several reasons. First, prior research suggests that women are less likely to separate their academic self-concept from their grades which is one of
the clearest types of recognition in a domain [29,30,58]. In particular, grades are the resource that women have the most access to. Academic self-concept is formed through grades and feedback from outsiders. Because women are less likely to receive recognition as someone with potential in STEM from their parents [17,18,24], society at large [16,21], and their instructors [19,31,75], they are more likely to rely on grade information to develop their academic self-concept. Also, women often tend to earn higher grades than men with the same standardized test scores [30,76]. Because women are often more accustomed to higher grades, they may have more concern about grades that are lower than what they are accustomed to, or they may compare their relatively low STEM grades and leave for another subject that gives them the recognition for their work that they are accustomed to [29,30].

We find that average grade anomalies and raw average grade data do not always reveal the same trends. Some courses have larger gender differences in AGA than in grades, such as Physics 1 and 2, Physics 1 Honors, Modern Physics Lab, and Optics Writing. This further speaks to the usefulness of tracking both AGA and grades of the students. An instructor may see a small or negligible grade difference by gender and assume that there is gender equity in their classroom based upon grade outcomes by gender, but without knowing the gender differences in AGA, the instructor will not understand how those grades are perceived by female and male students. Understanding both grades and AGA differences may allow instructors to understand both classroom-level inequities and the extent to which their course may be pushing out students, particularly those from underrepresented groups, such as women, out of STEM fields.

From our results, we make several recommendations to instructors and departments. First, measuring grade anomaly in addition to grades may be a useful way to find inequities in the learning environment. Measuring grades and gendered grade differences is both valuable for and accessible to individual instructors, but grade anomalies may be useful to departments concerned about student retention over longer periods and finding which courses may be discouraging students, particularly those from underrepresented groups, to leave a major.

V. CONCLUSION AND FUTURE RESEARCH
In this investigation, we found that grade anomalies exist for all studied physics courses at our institution for those who had declared a physics major. Further, several physics courses had an average grade anomaly that favored men over women. These findings raise particular concern about the need for an equitable learning environment and outcome for these students. These results are very important because they provide some evidence that courses in physics departments tend to have grade penalties. They support the grade penalties found in introductory courses from prior work on grade anomalies. The relatively new measure of AGA may also act as one measure of academic self-concept that is easy for institutions to access and evaluate over time. This can also be useful to researchers as they develop separate measurements for academic self-concept and expectancies for success. Although we have strong evidence of grade penalties in the studied physics courses for those who had declared a physics major as well as gendered AGA differences, we did not have access to syllabi or other information about individual courses offered over the thirteen-year period of data collection. Therefore, we are not able to pinpoint specific practices that may lead to grade penalties, grade bonuses, or gender inequities at our institution. However, we know that out of the courses currently offered, most of these courses focus on teaching in a traditional, lecture-based, and exam-reliant format.

Finally, this research is based at a primarily white, large, public university. While our results may generalize to similar institutions [54,55], we do not know what patterns of grade anomalies exist at smaller liberal arts colleges, minority-serving institutions, or community colleges. Conducting research at a diverse range of institutions in different countries, as well as a focus on how grade anomaly affects students from a variety of underrepresented groups, will help us more fully understand how grade anomalies differ for a range of students.

ACKNOWLEDGMENTS
This work was funded by National Science Foundation Awards No. DUE-1524575 and No. PHY-1806691. Additionally, APC charges for this article were fully paid by the University Library System, University of Pittsburgh.

[3] E. Marshman, Z.Y. Kalender, T. Nokes-Malach, C. Schunn, and C. Singh, Female students with A’s have similar physics self-efficacy as male students with C’s in
MALESPINA and SINGH

PHYS. REV. PHYS. EDUC. RES. 18, 020127 (2022)


gender.


[56] A. Wigfield and J. Cambria, Students achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes, Development Pol. Rev. 30, 1 (2010).


[70] B. King, Does postsecondary persistence in STEM vary by gender?, AERA Open 2, 1 (2016).


