Editors' Suggestion

Demographics of physics education research

Stephen Kanim[*]
Department of Physics, New Mexico State University, Las Cruces, New Mexico 88005, USA

Ximena C. Cid[†]
Department of Physics, California State University Dominguez Hills, Carson, California 90747, USA

(Received 28 April 2020; accepted 1 May 2020; published 27 July 2020)

Is physics education research based on a representative sample of students? To answer this question we skimmed physics education research papers from three journals for the years 1970–2015 looking for the number of research subjects, the course the subjects were enrolled in, and the institution where the research was conducted. We combined this data with demographics data about these institutions to compile a profile of physics education research subjects, and compared the demographics of this population to those of all students taking physics in the United States. Our results suggest that physics education research subjects, as a whole, are better prepared mathematically and are from a narrow and unrepresentative subset of our intended target physics student populations. For this reason, findings from research may not be as generalizable to all student populations as we have previously assumed.

DOI: 10.1103/PhysRevPhysEducRes.16.020106

I. INTRODUCTION

In the past half-century, physics education researchers have probed student thinking, affect and identity, developed curricula and tools for measuring progress, and created theoretical models. The tremendous success of this collective endeavor, based on data collected from hundreds of thousands of students, has resulted in a greater expectation within and beyond the physics community that educational progress should be solidly grounded in evidence-based scientific investigation. As with any science, though, the applicability of the results depends on the degree to which the research data sample fairly represents the research target. The successes of physics education research (PER) have been achieved without an explicit accounting of the match between our research population and the population of students that we intend to benefit. The intent of this paper is to highlight, and to attempt to quantify, the disparities that exist between the level of preparation and the background of the general population of students taking introductory physics in the United States (or, for some comparisons, students taking physics at American universities and colleges) and the student population reported on in the physics education research literature.

The selection of physics education research subjects has primarily come about as a result of convenience: Because most PER researchers are at four-year colleges and universities, most research subjects are also at these institutions, leaving high school physics students and two-year college students relatively unstudied. The bulk of the research has been conducted at more selective universities, and within these universities in more selective courses (i.e., the calculus-based introductory physics sequence rather than from algebra-based, conceptual, or other physics courses), and as a result the level of preparation of most research subjects has been higher than the level of preparation of most students enrolled in physics courses. Effectively, the physics education research community has inadvertently cherry picked its data.

An additional consequence of this selection of convenient research subjects is that fewer underrepresented racially and/or ethnically diverse student populations are included in the research population than are present in the general population of introductory physics students. In addition, the PER research population comes from wealthier families than the general population. As a result, both the challenges and affordances of more diverse student populations are not well represented in the research.

In the following introductory sections, we offer a brief discussion of relevant past research and of the procedure we used to compare the PER research population to the population of students in all physics courses in the United States. Subsequently, we present the results of this comparison and highlight those characteristics of the research population that are different from those of the overall

[*]skanim@nmsu.edu
[†]xcid@csudh.edu

student population. Before discussing the implications of our findings, we describe the sources of error and uncertainty in this study that we think are most significant.

A. Previous results in PER and in demographic analysis of other fields

We are not aware of any previous studies that have looked in detail at the demographics of PER subjects. Several publications have surveyed the publication record of physics education research as we did, including (i) McDermott and Redish's *Resource Letter: PER-1: Physics Education Research* [1]; (ii) Docktor and Mestre's; *Synthesis of Discipline-based Education Research in Physics* [2]; and (iii) Meltzer and Otero's *A brief history of physics education research in the United States* [3]. While these surveys served various purposes, none of them were focused on demographics.

Several other papers have described the state of PER and have made recommendations for future directions, notably (i) Heron and Meltzer's *The future of physics education research: Intellectual challenges and practical concerns* [4]; and (ii) the National Research Council's *Adapting to a changing world: Challenges and opportunities in undergraduate physics education*. [5]. Neither of these studies provide analysis or recommendations with regard to research population demographics. In addition, several lectures associated with the AAPT's Millikan Award and Oersted Medal provide expert perspectives and recommendations for physics education researchers. While some of these lectures discuss the changes that have occurred in the level of student preparation that can be expected in physics courses, and others recommend changes to curricula and focus in order to attract and keep a more diverse student population in physics courses [6–9], they have not studied the relationship between physics course demographics and PER subject populations.

While we did not find previous research directly related to the demographics of research populations in PER, there are studies in other fields that compare research populations with the demographics of the overall population that field is purported to study. For example, in a paper titled *The weirdest people in the world?*, [10] Henrich, Heine, and Norenzayan look at the research basis for behavioral science. As a field, behavioral science depends almost exclusively on college student subjects from western, educated, industrialized, rich, and democratic (WEIRD) populations. (As university students, some of us have had the experience of serving as research subjects for psychology and behavioral science studies: Often, students in general education psychology classes are asked to sign up for these studies as part of their course credit. For researchers, these students form a convenient subject pool, because they are easy to recruit and are already on campus.) The authors point to various cross-cultural studies that show that the responses obtained from western college students are not predictive of the responses of humans in general, and in fact are often outliers. This does not mean that the results of studies depending on these students are incorrect or that they are not useful. It does however, cause concern when behavioral scientists use the over-generalized results of these studies to make claims, or to build theories, about the behavior of humans in general. While the behavioral science studies that are based on college students may be useful starting points for such claims and theories, and may be useful for establishing experimental protocols, validation requires cross-cultural replication. Similarly, in the field of genomics, Popejoy and Fullerton found in a 2016 study that 86% of all genome-wide association studies (GWAS) have samples taken from people of European descent [11]. While the studies that have been conducted have improved understanding of genetic influence on diseases, and have highlighted genetic risk factors for many of these diseases, the homogeneity of the research population limits the generalizability of findings.

These studies from other fields can serve as guides for what might be expected from studies of population variability in PER. First, as *the weirdest people in the world* [10] study suggests, it is possible that many reported PER results are not generalizable to the overall student population, and that some of our results may, in fact, be outliers. Second, PER is probably similar to behavioral science and genomics in that much of the disparity between research and overall populations is likely a result of reliance on readily available research subjects rather than careful selection of subjects to assure a good match with the overall population. Finally, like genomics, [11] we can expect that diversifying our research pool will be a significant challenge.

B. Description of this study

In order to evaluate which student populations the PER community has studied, we selected a sample of PER papers that we believed would provide a sufficiently broad but still manageable and representative sample of research descriptions. We chose to look at the research published from 1970 to 2015 inclusive from three journals: The American Journal of Physics (AJP), Physical Review: Physics Education Research (PRPER), and The Physics Teacher (TPT). Here we describe our procedure.

We scanned each physics education paper, looking for four elements: (i) The type of research study; (ii) the total number of students that data were obtained from; (iii) the institutions where the research was conducted; and (iv) the courses that the students were enrolled in. (The topic of study was also noted, but is not relevant to this population study.) We used a very broad definition of physics education research to select which papers to scan, and many of the scanned papers were subsequently eliminated as described below. Because this scanning was relatively

TABLE I. Number of papers included in this study and number of total students in various populations of courses from included papers.

	PRPER	AJP	TPT	Total
Papers: Total	342	372	317	1031
Papers: Included	179	159	79	417
Student population	*N* from PRPER	*N* from AJP	*N* from TPT	*N* total
K–9	10	193	0	203
High school	674	1590	19 275	21 539
Two-year college	380	321	0	701
Teacher prep	938	1598	260	2796
Disadvantaged	0	69	0	69
Conceptual	615	2597	469	3681
Honors	405	924	0	1329
Algebra-based	8341	23 718	4607	36 666
Calculus-based	56 991	109 218	13 487	179 696
Upper level	4600	3108	3	7711
Other	145	3061	60	3266
Total	73 099	146 397	38 161	257 657

quick, and in many cases we had to make guesses, results are necessarily rough.

From a spreadsheet compiled from these data, we then eliminated studies that did not include student data from physics courses (for example, purely theoretical papers, papers about textbook approaches, papers that described curricular or diagnostic modifications with only summaries of student success rates, or papers where all data were collected in mathematics or astronomy classes). Some papers included results from graduate students and faculty, and these populations were not included in our analysis. We also eliminated studies that were summaries of other research or were metastudies. In addition, we did not include studies or parts of studies that were conducted outside of the U.S.. Finally, in some cases we could not make reasonable guesses about the information we were looking for based on our scanning of the papers, and those studies were also eliminated.

To avoid eliminating a large number of the papers that remained, we developed some rules for dealing with uncertainties in the data. For cases where the paper we were scanning reported an aggregate number of students for different courses or for different schools, we simply split the number of students evenly among the courses or schools. When the name of the school was not given, we assumed that the research was carried out with students from the authors' home institutions. (If data were reported from multiple institutions without naming those institutions, we eliminated that paper from our dataset.) For some papers we estimated the number of students from whom data were collected based on the reported number of sections and average class size. For example, a paper might state that there were two sections of algebra-based introductory mechanics, which on average (for that institution) have about 150 students. We would assume there were 300 students total even though there were likely fewer or more students for that particular study. In some studies it was clear that students were enrolled in a post-secondary introductory course, but it was less obvious whether this course was algebra based or calculus based. For other studies, research subjects were chosen from both of these courses and the results were not separated. Rather than eliminate these studies, in both of these cases we arbitrarily chose to assign half of the students to each course.

From a total of 1031 scanned papers, we included 417 papers in our dataset, as shown in the top part of Table I. By our count, the included papers describe research conducted on a population of 257 657 total students. There are likely cases of double counting included in this total, because many studies report results from multiple questions, and it is often impossible to tell whether the same student population was involved.

As shown in the lower part of Table I, we binned the students in each study into 11 categories based on their level and the courses in which they were enrolled: Kindergarten through ninth grade; high school; students at two-year colleges (TYC) enrolled in any physics course; students enrolled in pre- or in-service courses for teachers; students in special courses for disadvantaged or underprepared students; students in conceptual physics courses, students in introductory honors courses; students in introductory algebra-based courses; students in introductory calculus-based courses; students in upper-level physics courses; and students in other courses (e.g., lab courses). Students in all but the first two categories are in postsecondary courses.

We compare our results to data about the relative sizes of the student populations for the United States, obtained from the American Institute of Physics [12].

II. RESULTS

A. Overview

Six important results are suggested by our study and are described below: (1) PER in the U.S. pays scant attention to high school students; (2) PER in the U.S. almost completely neglects students in two-year colleges; (3) PER studies of introductory courses focus disproportionately on students in the calculus-based course; (4) PER studies are conducted primarily at institutions where the math preparation of incoming students is relatively strong; (5) PER studies are conducted at institutions that have wealthier students and have a smaller fraction of underrepresented racially and/or ethnically diverse students than the overall college-bound student population; and (6) Sampling in upper-division physics courses is also highly homogenized. These results are described individually in more detail below.

Result 1: PER in the U.S. pays scant attention to high school physics

During the 2012–2013 school year, 1.38 million students were enrolled in physics courses in the U.S. in both public and private high schools [12]. [This segment of the overall physics student population has also grown faster than introductory courses in colleges and universities, as shown in Fig. 1(a)] In comparison, there were about 0.5 million students enrolled in introductory physics courses in colleges and universities. From our study, only 76 of 417 total papers (24 PRPER, 21 AJP, and 31 TPT) reported data collected from a total of 21 539 students (674 PRPER, 1,590 AJP, 19,275 TPT) in high school classrooms. Whereas about 3 of every 4 American students are studying introductory physics in high school classrooms, fewer than 1 in 10 of our research subjects are high school students. Had we started our survey in 1972 rather than in 1970 this disparity would be even more pronounced: Over 10 000 of the high school students—about half—are reported on in a single survey published in TPT in 1971.

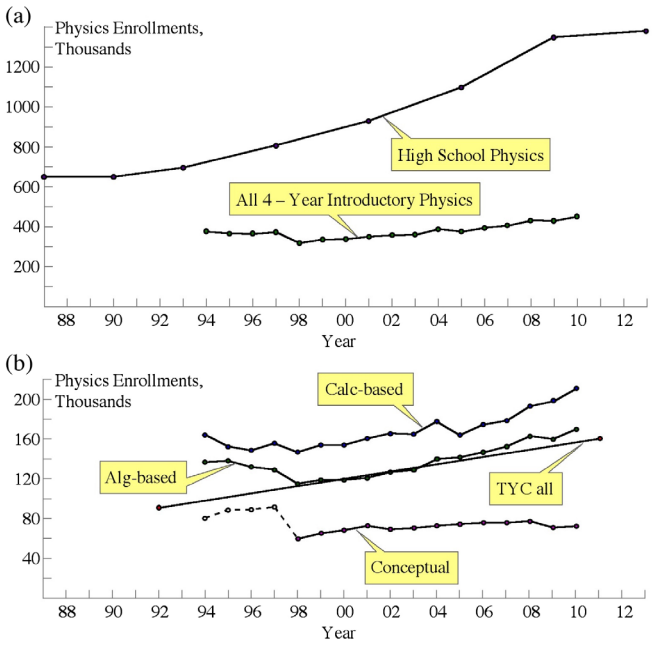

FIG. 1. (a) High school physics enrollment compared to physics enrollment at colleges and universities. (b) Enrollment in university courses broken down by type of course. Data for both graphs are from AIP statistics.

This scarcity of data from U.S. high school physics is distinctly different from PER conducted internationally, where it is much more common for research to be conducted in high schools. [For example, in the 2014 GIREP (Groupe International de Recherche sur l'Enseignement de la Physique) Conference Proceedings [13], 14 papers are studies of teaching and learning at the university level, with 7 of these studies from Europe. In contrast, 33 papers are included of secondary-level studies, with 27 of the studies from Europe.] It is also different from mathematics education research in the U.S., which places a higher emphasis on K–12 students. For example, in 2017 and 2018 the Journal for Research in Mathematics Education (JRME) [14], (which can crudely be considered as the equivalent in mathematics of PRPER in physics) had 11 papers including data from 24 433 K–12 students and 5 papers including data from 10 916 university and two-year college students (about 30% of total research subjects). These numbers reflect a recent *increase* in attention given to post-secondary students: For 1997 and 1998, JRME had 23 papers with data from 5935 K–12 students and only 2 papers with 830 students from universities (about 11% of all research subjects) [14].

It may be that one reason for the paucity of physics education research data from high schools is the difficulty of obtaining IRB approval for studies conducted in public schools compared to approval for research at universities. However, the apparent success of mathematics education researchers at obtaining approvals for K–12 research indicates that this is not an insurmountable obstacle.

Of the six results that we report on in this paper, our claim about high school physics is most dependent on our choice of journals. Had we chosen Journal of Research in Science Teaching (JRST) instead of TPT, for example, we might have seen better representation from high schools in the research population. We discuss this limitation in more detail in a subsequent section of this paper.

Result 2: In the U.S. we do almost no research on physics students in two-year colleges

About 44% of all undergraduates are enrolled in two-year colleges, and many of these students take physics. As shown in Fig. 1(b), in 2011 (the most recent year for which data is available for two-year colleges [15]) there were approximately 161 000 students taking physics at two-year colleges (calculus-based, algebra-based, and conceptual physics courses only), compared to about 450 000 students taking physics at four-year institutions [12]. In our study we

found only 6 papers (2 PRPER and 4 AJP) reporting on two-year college students in the United States, with a total student research population of 701 students (380 PRPER and 321 AJP). Although about one-quarter of all college physics students are enrolled in these institutions, only 0.3% of the total number of students studied by PER are from two-year colleges. Not only is data from two-year college students lacking in PER, but the American Institute of Physics rarely collects data about physics in two-year colleges: The report cited here was published in 2013 [15], and refers to surveys conducted in 2011, 2001, and 1995. This scarcity of information severely limits our understanding of physics learning in two-year colleges.

Result 3: In introductory physics, PER tends to focus on students in the calculus-based course

Our study also suggests a disproportionate reliance on data from the calculus-based course. As seen in Fig. 1(b), about 210 000 students were enrolled in calculus-based physics in 2010, accounting for about one-third of the students taking introductory physics of all kinds at colleges and universities. From our dataset, there were 179 696 students enrolled in calculus-based physics, accounting for about 82% of all research subjects in introductory courses. Our results for introductory physics at the university level are broken down by course in Fig. 2, and are compared to data about students enrolled in all courses.

We have included students at two-year colleges as a separate category in Fig. 2, because we wanted some visual representation that emphasized the degree to which these

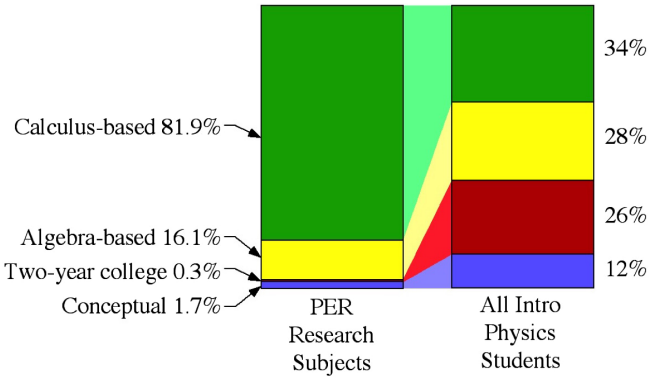

FIG. 2. Comparison of distribution of research subjects in all post-secondary introductory courses to distribution of all introductory students at colleges and universities in our study. Here we have included students in honors introductory physics courses with students in calculus-based physics. About one-third of all students taking post-secondary introductory physics in the U.S. take a calculus-based course at a four-year institution. In contrast, more than four-fifths of all students reported on in the research papers of our study were in the calculus-based course at a four-year institution. About one-quarter of all students taking introductory physics (of all varieties) do so at a two-year college; only 0.3% of the students in our study were from two-year colleges.

students are underrepresented in PER studies. Based on data from AIP in 2010, about 26% of these 161 000 two-year college physics students are enrolled in calculus-based physics, about 48% in algebra-based physics, and about 25% in conceptual physics. (Additionally, there are about 16 000 students taking some other physics course, and about 39 000 students taking physical science or technical physics.) If we apportion the two-year college students shown in red to the other categories, then about 41% of all post-secondary introductory physics students are in a calculus-based course, about 41% are in an algebra-based course, and about 19% in a conceptual physics course. Roughly, then, about twice as many calculus-based students are selected for PER studies as are represented in the overall target population.

It is likely true that PER's oversampling of students from the calculus-based course has implications in terms of mathematical preparation, because students enrolled in this course are at least concurrently enrolled in calculus, whereas many students in other introductory physics courses will not have taken calculus. Beyond issues of mathematics preparation, however, there are other reasons why this nonrepresentative sampling should be of concern: At many universities, the choice of physics course is based on major, so life science and education majors are more likely to take algebra-based physics while physical science and engineering majors are more likely to take calculus-based physics. We expect that there are epistemological differences between these groups of students that are not adequately accounted for in PER studies, and it is possible that students in the calculus-based course are unrepresentative along a host of other dimensions as well.

Although it is probably no more difficult to obtain IRB approval for one post-secondary physics course compared to another, we can speculate about possible other reasons for the overselection of calculus-based physics students by PER researchers. First, it is often the case that these courses are larger, and that it is easier to collect larger datasets by choosing the calculus-based course for a study. Second, for PER researchers hoping to demonstrate the utility of PER to non-PER physics faculty members, it is useful to focus on the course that has most physics majors and potential physics majors, and that most physicists took when they were studying.

Result 4: In introductory post-secondary courses, PER tends to focus on students with stronger mathematics preparation

As result 3 shows, students in PER studies are likely to have stronger mathematical preparation than the overall population of students in introductory postsecondary physics courses based on the most advanced math course they must have taken in order to enroll in their selected physics course. In this section, we show that this likely difference in math preparation is compounded by PER's oversampling of research subjects from universities that are more selective

in terms of mathematical preparation. This argument requires several assumptions that we enumerate as we describe our procedure below.

Because it is easy to obtain and relatively ubiquitous, we have chosen math SAT scores [16] for entering freshmen for a university as a proxy for mathematics preparation. Discussion of the relationship between SAT scores and physics performance are included in Refs. [17–19] (For all SAT Math scores discussed in this paper we have used scores from the "old" SAT, administered before March 2016. For a few schools we were unable to obtain SAT scores and substituted converted ACT math scores.) We acknowledge that, in practice, SAT Math scores are at best a crude measure of the degree to which students are prepared to use math in physics and look forward to the development of more relevant measures (and to the development of measures that reduce or eliminate some of the biases against various populations that are inherent in standardized assessments including the SAT) [20,21].

We used data about SAT Math scores that were available for the incoming freshman class at each university. We expect that the SAT Math scores for students enrolled in introductory algebra- and calculus-based physics courses at any given university are higher than those scores for the freshman class as a whole (except institutions where all freshmen are required to take introductory physics), and we are assuming that the percent increase in scores obtained by considering this subset is about the same for all universities. We offer a crude test of this assumption in Sec. III.

To test result 4, from the 257 657 students included in this overall study we eliminated the K–12 students and the students in courses listed as Other in Table I, leaving 224 938 students in college courses of all types. We further eliminated students from unnamed institutions, leaving 210 784 students. (While entire studies without named universities had previously been eliminated from our study, these additional 14 154 students we eliminated were from studies where some of the data were from named institutions, and other data were not.) We then sorted the universities in our sample by the total number of students from each university that had been reported on. The 39 universities with the largest number of students account for 95% of all the U.S. university students reported on in the three journals, 200 079 of 210 784 students. For these 39 universities, we looked up the 25th and 75th percentiles of SAT Math scores for the entering freshman class for the year 2016 [16].

In Fig. 3, we illustrate disparities between the SAT scores of the student research population and the overall distribution of SAT scores for students in introductory physics. The light blue histogram shows the SAT Math scores for all students taking the exam in 2008 (the most recent year for which these data are available), binned in 50-point increments. The middle half is shaded a darker blue, with the 25th and 75th percentiles for all students taking the SAT indicated by the dashed vertical blue lines (430 and 590, respectively).

The red-shaded area represents the physics education research student population from our data. Each of the 39 universities (which comprise 95% of our data) in the research population is represented by a rectangle extending horizontally from 25th percentile to 75th percentile math SAT score for incoming freshmen at that university. The relative heights of the rectangles are determined by the fraction of the total research population represented in

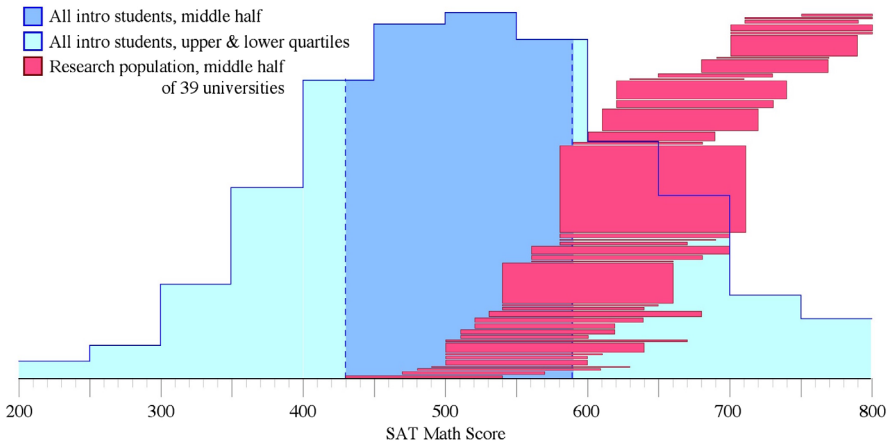

FIG. 3. Comparison of middle half SAT Math scores for all students with those of incoming freshmen at universities where most PER is conducted. The blue-shaded histogram gives the distribution of SAT Math scores for all students taking the SAT in 2008 [22], the last year that the College Board released these data in 50-point bins. The dashed vertical blue lines represent the 25th percentile score (430) and the 75th percentile (590) scores for all students; with the middle half shaded a darker blue than the lowest and highest quarters. The 39 stacked rectangles shaded in red are the 25th–75th percentile ranges for incoming freshmen at the universities whose students comprise 95% of all university students in our study. The horizontal span of each rectangle in the stack represents the middle half of incoming freshman students for a single university, with the height of that rectangle giving the proportional representation from that university of all students in our study.

studies from each university. For example, if the middle half of all math SAT scores from entering students at University X fell between 600 and 700, and if PER data from students enrolled at University X accounted for 10% of all research students from the 39 schools, then the rectangle representing University X would span horizontally from 600 to 700, and the height of this rectangle would be one-tenth of the overall height of the stacked red rectangles shown in Fig. 3. The rectangles are stacked in order of increasing 25th percentile scores for each school so that the diagram maximizes the overlap between the middle half of all students (the dark blue region) and the middle half of the students in the research population.

We recognize that not all high-school seniors who take the SAT will enroll in college, and that not all students who enroll in college will take physics courses. Nonetheless, as shown in Fig. 3, the 75th percentile for students at many universities is about where the 25th percentile lies for a large fraction of the students included in PER studies. For instructors at these universities—most physics instructors in the United States—that are trying to improve their physics instruction, it requires a substantial leap of faith to assume that typical PER research results have much relevance for students in their courses.

A second comparison can be made from the binned SAT Math scores for each school's incoming freshmen. For each of the 39 schools, we used College Board data [16] and the total number of students in our study from that school to calculate how many of the students were in each 100-point bin. For example, if we had 1000 students from a school, and 20% of the incoming freshmen from that school had SAT Math scores between 300 and 390, then we counted 200 students from that school with scores in this range. We then added up our binned data for all 39 schools to obtain the percentage of students in our study with scores in each bin. Since the College Board also reports binned data for all students taking the test [16], we can compare the results, as shown in Fig. 4. Again we see that the PER sample is not representative: For example, while 46% of the overall population has SAT Math scores below 500, only 7% percent of the students from these PER institutions do. At the other end of the scale, 25% of the overall population has SAT Math scores of 600 or more, but students with these scores account for 70% of students taking the math SAT at the 39 institutions representing the research population.

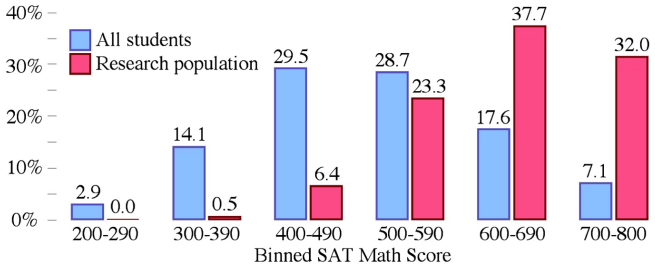

FIG. 4. Binned comparison of SAT Math scores for all college-bound students with students at universities where most PER is conducted. The blue bars are the SAT Math scores for all students who took the SAT in Spring 2016, collected into 100-point bins and given as a percent of all students taking the exam. The red bars are for students at the 39 universities where most PER is conducted, weighted by the contribution of each university to the total number of students in our sample.

Result 5: PER in the U.S. oversamples from White and wealthy populations

PER researchers by and large conduct research on students at their own institutions, so the demographics of the research population reflect those of the institutions conducting research and are not necessarily similar to those of the overall population of physics students. Figure 5 was generated by looking at the racial demographics of the 39 institutions that make up 95% of PER student populations represented in the literature in our study. [While there are PER groups at Hispanic Serving Institutions (HSI) and Minority Serving Institutions (MSI), the data reported from these institutions is less than 5% of the overall data collected for physics education research as represented by our 39-school sample. There is no data in our sample from HBCUs (Historically Black Colleges and Universities) or from Tribal Colleges and Universities.] As with the SAT Math data, we scaled the demographics reported from each institution by the fractional contribution of that institution to the overall PER student population. We then compared the racial demographics of the overall research population to the racial demographics of all college-bound seniors for 2015, also obtained from the College Board [16]. (The categories in

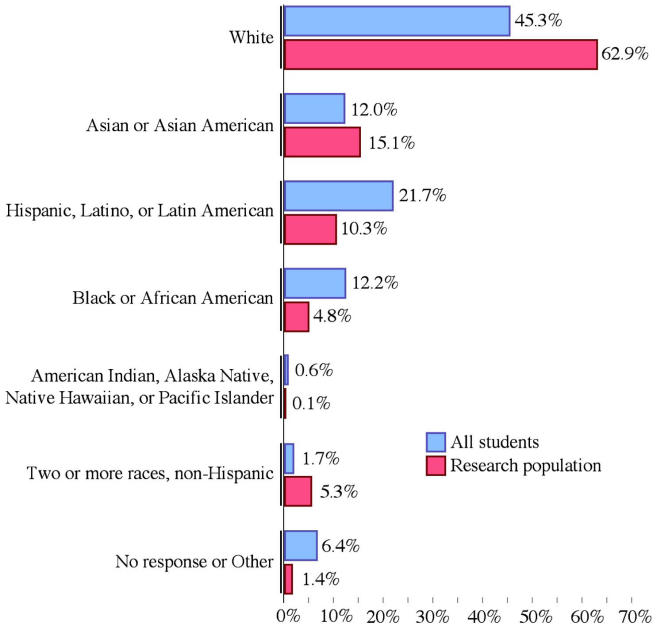

FIG. 5. Comparison of racial or ethnic distributions of all students taking the SAT in 2015 versus the PER research population from our study. Data and category titles are taken from the College Board website.

Fig. 5 were chosen to match the categories presented by the College Board website.)

The PER data are nonrepresentative of underrepresented racially and/or ethnically diverse populations: For example, while Latinx, Black, and Indigenous American students are 34.5% of the college-bound students taking the SAT, by our estimate only 15.2% of the research student population are from these groups. The only exception is the category for *two or more races, non-Hispanic*. We can only speculate about this unexpected result.

As with our discussion of math preparation, it is likely that the racial or ethnic makeup of all incoming freshmen are different from that of students in physics courses. Based on the demographics of STEM fields in general, it is probably true that there are fewer underrepresented students in physics courses than there are in the overall student population at any university. We don't know whether the difference is greater for the 39 schools in our sample than in all universities.

A similar analysis can be made comparing the research student population with the overall student population broken down by parents' income as shown in Figure 6. Data from the Equality of Opportunity Project [23] divides parents' income for students at each school into quintiles—for example, giving the percent of students at a school whose parents' incomes are in the bottom quintile of all parents' incomes in the country.

Data were not available for two of the 39 schools in our research population, representing a total of about 8% of the research student population. We took the data for the other 37 schools, representing 87% of the total research student population, weighted the percentages by the fraction of students from that school in the overall research population, and added the results to obtain a parent income distribution for our research population.

From the list of all schools included in the Equality of Opportunity Project study, we eliminated categories of schools that we thought were unlikely to offer physics—schools categorized as two-year for profit, and schools that

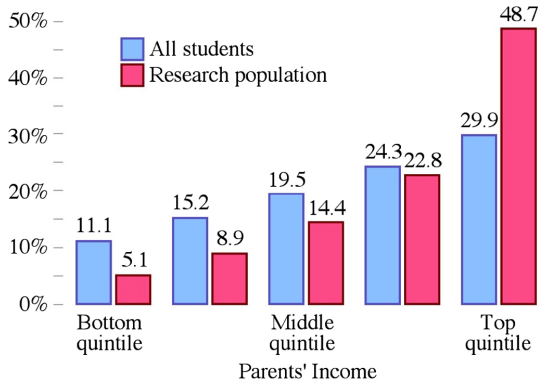

FIG. 6. Comparison of parents' income for research student population (from 37 universities representing 87% of the PER research population) to overall student population.

offered programs of study lasting less than two years. We compared the results from our research population to the percentages in each quintile of the remaining schools included in the study weighted by student population. (Repeating the comparison while including all schools made almost no difference in our results, with the largest changes to any quintile of about 0.3%.)

For eight of the 37 institutions on our list, representing about 23% of the total research population, data were given for a university system rather than for a specific university (e.g., cumulative data are presented for all branches of a state university system rather than for individual universities within that system). In each of these eight cases, the PER research data were collected at the flagship university of the university system, and it is likely that research students' parents are wealthier than the data suggest.

Almost half of the students in our research population have parents with incomes in the top quintile of all parent incomes, compared to about 30% of the overall population. At the lower end of the income scale, 14% of the students in the research population have parents with incomes in the bottom 40% (bottom 2 quintiles), whereas about 26% of the students in the overall population fall into these categories. Overall, students in our research population come from wealthier families, and probably were educated in wealthier K–12 school districts.

Result 6: Sampling in upper-division physics courses is highly homogenized

Table I shows that we counted 7711 upper-division students as part of our PER research subjects. As with our analysis for result 4, we needed to remove an additional 989 students from studies where some of the data were from unnamed institutions, leaving 6722 students in upper-level physics courses Only 22 different universities are represented, and over 95% of these students attend 9 universities. Since the two universities reporting the most students account for 80% of the upper-level students included in this study, the demographics of these two universities strongly determines the demographics of physics education research conducted in upper-level courses, and by extension for upper-level topics such as quantum mechanics and statistical mechanics.

Using a proportional distribution of expected SAT Math scores as we did in our description of result 3, we find that about 58% of the research sample of upper-level physics students can be expected to have scores above 600, compared to about 25% expected of students at all schools. This result is particularly salient for upper-level students in light of a study [18] that claims that a practical minimum SAT Math score of 600 is necessary for students to succeed as physics majors.

Based on our analysis, we expect that about 60% of our sample population of upper-level students will be White (compared to 45% of all students), about 11% Black or Latinx (compared to 34%), and a reasonable number of

Native Americans to expect in our upper-level research sample is zero.

III. LIMITATIONS OF THIS STUDY

In this section we describe what we believe are the three most important sources of uncertainty and error in our study, and what we did to gauge the impact of these uncertainties.

Limitation 1: Errors and uncertainties introduced by our data collection methods

This study was initiated by S. K., who scanned through all papers from PRPER, AJP, and TPT, with results shown in Table I. As described earlier, only enough time was spent on each paper to find (a) the type of research study, (b) the total number of students involved in the study, (c) the institutions where the research was conducted, and (d) the course that the students were enrolled in. How dependent are the reported results on the interpretations of the papers made by an individual researcher? How different might the results be if more time had been spent on each paper, or if a different researcher had scanned the data?

In our scanning of the papers, it was unusual to spend more than 5 min on any given paper. We recognized at the outset that, because we were not reading the papers in detail, we could expect to make errors in our data tabulation. Often the information we sought was not explicitly provided as part of the paper, and we had to make judgments or guesses. For these reasons, we expect that any attempt to replicate our results would create a different dataset. Our hope is that the conclusions that we draw are sufficiently robust that a replication would not substantially change the conclusions drawn.

To attempt to quantify the effect of quick scanning and of data variations that result from individual researcher choices on our results, the authors chose a random subset of 300 papers (100 from PRPER, 100 from AJP, and 100 from TPT), and X. C. C. independently scanned these papers so that we could compare results.

Often, the two authors disagreed because we looked in different places for the information we sought. For example, data on the number of students in a study appear in a variety of places: abstracts (sometimes an approximation), introductions, descriptions of methodology, tables, table captions, figures, figure captions, and the body of the text (as numbers, words, or both). It is very common for a single paper to include multiple studies and multiple reports of data, and perhaps the most common source of disagreement and ambiguity was a determination of whether data from students in one study were the same students presented in another study reported in the same paper. Additionally, it was also often unclear if pre- and postdata were from the same sample population, as pre- and postdata often report different numbers of students, and for some studies they appear to be pretreatment and posttreatment values for completely different populations altogether.

Almost all of the differences in our tabulations were in the number of students, rather than in the course category or in the institution represented. Often, we made different decisions about whether multiple sets of reported data represented the same group or different groups of students. For each paper where we obtained different results, both authors reread the paper more carefully, discussing where we obtained the results we used and then decided which result was more consistent with our established rules for data selection. Based on this discussion we arrived at a consensus dataset for the 300 papers we both reviewed. These papers represent a sampling of about 30% of the 1031 total papers that form our dataset.

Table II represents the consensus total for the number of students in each category after more detailed reading of those papers where the authors disagreed. Cells that were unchanged from the values given in Table I, or that had changes of less than 1%, are reported here without

TABLE II. Data corrections based on review of one-third of the papers. Changes to values given in Table I are shown in parentheses: Entries without percentages given remained the same after review, or were changed by less than 1%.

Student population	N from PRPER	N from AJP	N from TPT	N total
K–9	10	193	0	203
High school	674	1590	19 292	21 556
Two-year college	380	1400 (+336%)	0	1780 (+154%)
Teacher prep	1188 (+27%)	1623 (+2%)	260	3071 (+10%)
Disadvantaged	0	69	0	69
Conceptual	803 (+31%)	2692 (+4%)	519 (+11%)	4014 (+9%)
Honors	523 (+29%)	816 (−12%)	0	1339
Algebra-based	8505 (+2%)	24 139 (+2%)	4609	37 253 (+2%)
Calculus-based	65 308 (+15%)	108 907	13 426	187 641 (+4%)
Upper level	5594 (+22%)	3489 (+12%)	3	9086 (+18%)
Other	191 (+32%)	3356 (+10%)	82 (+37%)	3629 (+11%)
Total	83 176 (+14%)	148 274 (+1%)	38 191	269 641 (+5%)

percentages; the percent change from the values given in Table I are shown for all other entries.

By far the most significant change in a single entry was in the data from AJP for two-year colleges, which became 1400 students for our consensus dataset, roughly 4 times as large as our initial count of 321 students. This change was a result of a reassessment of a single paper that included seven universities and a single two-year college. Strict application of our rules assumes an even number of students are from each of the eight institutions, resulting in the larger number, even though all reported results on individual questions are from Ph.D. granting institutions only and there are no published results from the two-year college. Our revised number for two-year colleges, 1780 students for all three journals, still means that fewer than 1% of PER subjects are from two-year colleges even though about one-quarter of all introductory students are enrolled at these institutions. Despite the steep increase in the number of students, our conclusion that two-year colleges are all but ignored is unchanged.

Some of the other changes for a population category for a single journal are also quite substantial—in the worst case, a 37% increase (though this is for a small-N value for the "other" category, which was not used for any of our results). On the other hand, the total number of students for any given student population for all journals combined was 18% or less in all cases except for two-year colleges. Our sense is that the conclusions we have drawn about population disparities are not that sensitive to the variations in categorization and counting due to differing interpretations of paper descriptions, or due to overly hasty scanning of papers.

Limitation 2: Effect of limiting our study to three journals

The claim we make in this paper is that the research basis for PER in the U.S. is not based on a representative sampling of the students enrolled in physics. However, this study looks at PER only in the American Journal of Physics, Physical Review Physics Education Research, and The Physics Teacher. How well do these three journals represent the publication record of all of PER? Is it possible that a wider sampling of journals would change our claim? To attempt to answer these questions, we looked at two summaries of PER with extensive publication listings, a paper by Docktor and Mestre [2] and an earlier resource letter by McDermott and Redish [1].

Docktor and Mestre synthesis: In a 2014 paper, Docktor and Mestre [2] use the results of a paper commissioned by the National Research Council to summarize and categorize physics education research. The paper has 539 references, of which 84 are books, 40 are conference proceedings, 26 are websites, 10 are dissertations, and 2 are videos. The remaining 378 references are journal articles from a total of 41 journals. Of these articles, 250, or about two-thirds, are included in our three-journal sampling. Twenty of the remaining 38 journals have only one or two papers listed from those journals. Four journals are reasonably well represented: the Journal of Research in Science Teaching, JRST, (20 papers), Cognition and Instruction (12 papers), the Journal of the Learning Sciences (12 papers, one of which is in reference to an entire issue), and Science Education (11 papers), collectively accounting for 55 of the 128 papers in other journals. For comparison, the Docktor and Mestre summary cites 133 papers from the American Journal of Physics, 100 papers from Physical Review Special Topics Physics Education Research, and 17 papers from The Physics Teacher.

McDermott and Redish resource letter: In 1999, McDermott and Redish published a physics education resource letter [1] containing 243 references. (Some footnote numbers reference more than one paper.) Of these, 52 are books, 10 are conference proceedings, and 3 are websites, leaving 178 references to journal publications distributed over 25 journals. Not quite half (86 or 48%) of these publications are included in our three-journal study, 64 papers in the American Journal of Physics and 22 papers in The Physics Teacher. (Because this resource letter was published in 1999, there are no reference papers for PRPER, which began publication in 2005.) Fifteen of the remaining 22 journals have only one or two referenced papers. The four most commonly cited of the remaining journals are the International Journal of Science Education (32 papers), the Journal of Research in Science Teaching (16 papers), Physics Education (9 papers), and Physics Today (7 papers).

Based on these two PER summaries, our three-journal choice is reasonably effective at representing PER publications during the years 1970–2015. It might have been better to have also included the Journal of Research in Science Teaching, as papers there were cited 38 times in the two summaries, about the same as the 39 citations for papers from The Physics Teacher. We can get some sense of the effect of this inclusion by looking at these 38 citations from JRST. Only two papers were cited in both summaries, so there were 36 distinct papers included. Of these, 5 papers were summaries or theory papers that included no student data, 11 papers had no data from students in the United States, one paper included data only from students not taking physics, and one paper was based on data from graduate students. The 18 remaining papers were based on 4186 research subjects.

Fifty-eight percent (2437 subjects) were elementary- and middle-school students, reported on in two studies about scientific reasoning in the late 1970s. In contrast, there were only 203 K–9 students among the subjects of the papers in the 3 journals we chose. Almost 20% (815 subjects) were high school students. This more than doubles the percentage of high school students who were research subjects in our study of AJP, TPT, and PRPER papers. It is safe to say, then, that inclusion of JRST would have weakened our

claim that PER in the U.S. pays scant attention to high school students. If we assume (based on the two summaries described in this section) that the inclusion of JRST would add about the same number of papers and therefore about the same number of research subjects to the pool as TPT does, and that the number of research subjects from high schools in JRST remains 20%, then we can expect that the total fraction of high school students in an expanded study would be about 10%, rather than 8%. Keeping in mind that about three-fourths of all introductory physics students are in high school, our conclusion (result 1) that PER is dramatically undersampling high school students is likely unchanged by inclusion of JRST.

Only 899 of the JRST subjects, or 21%, were university students. All but 57 of these students were from the 39 schools described previously that make up the bulk of PER research subjects. There were no studies that included upper-level students. We do not believe that including JRST would have resulted in any changes to our five other research results, because the students by and large come from the same schools and therefore have similar demographics as the students that we are basing our conclusions on.

With more time, a valuable follow-on study would be to look at the effects of including JRST and perhaps other journals in more detail. Based on our analysis above of a small sample of these additional studies, though, we do not expect that expanding the study in this way would have much effect on the results we describe in this paper. An interesting additional study would be of the degree to which the PER research community overlaps the high school science learning research community, perhaps as measured by intra- and intercitation instances for these two groups.

We note in closing that, while it might seem beneficial to add journals to our study, there is also an inherent confounding factor: Many of the papers that were included in the two syntheses but published in other journals were based on the same research studies as those already published in the three-journal analysis that forms the core of our discussion, with the authors providing an alternate analysis, or perhaps one better suited to a different audience. As a result, a number of the students included by adding journals would be double counted.

Limitation 3: Use of university-wide SAT Math data instead of physics class data

In this paper we have used SAT Math scores obtained from the College Board as a measure of the mathematics preparation of students in physics courses at the 39 institutions where physics education is primarily conducted. These data are for all incoming freshman at each institution, and we assume that at almost all institutions the SAT Math scores are higher for students enrolled in physics courses than those scores for all freshmen. Result 4 (research students are better prepared than the overall population of physics students) and to a lesser extent result 6 (research on upper-level physics students is highly homogenized) depend on an assumption that SAT Math scores of physics students are monotonically related to those scores for all incoming freshmen. With this assumption, (since the upward shift in scores happens for all schools) a comparison of incoming freshmen scores will inform us about the relative scores of students in physics courses.

This assumption may not be valid for schools with extremely high SAT Math admission scores. For a hypothetical school with a 25th percentile of 700 and a 75th percentile of 790, for example, we would not expect the scores in physics courses to be much different. (Some STEM-oriented schools require that all students take physics; for these schools the incoming freshmen would have the same score distribution as students in physics.)

For the university with the lowest 25th and 75th percentiles of the 39 schools that comprise our PER research sample (the bottom red rectangle in Fig. 3) we would expect no such ceiling effect. This university uses ACT scores rather than SAT scores, and we have converted them to SAT equivalent scores in Fig. 3 and below. We obtained data from two fall semesters that included four sections of the algebra-based introductory physics course ($N = 438$) and four sections of the calculus-based introductory physics course ($N = 507$). We compared SAT Math scores for these classes to the SAT Math scores for incoming freshmen at that university during the same two semesters. Both the 25th and 75th percentile scores were about 30 points higher in the algebra-based course than for all incoming freshmen. In the calculus-based course the 25th percentile was about 90 points higher than the 25th percentile for all freshmen; the 75th percentile was about 70 points higher. Roughly, then, compared to all incoming freshmen, the algebra-based middle half had SAT Math scores about 30 points higher and the calculus-based course had scores about 80 points higher. The median scores also increased by about the same amount.

We also had access to data from a calculus-based course ($N = 277$) at a university whose SAT Math scores placed about halfway up our list, the 22nd of 39 schools ranked by increasing 25th percentile scores. The 25th percentile for the SAT Math was 650 in the calculus-based course, or 70 points higher than for all freshmen at this university. The 75th percentile was 720, only about 20 points higher.

Using the data points for the calculus-based courses (about 82% of the research population), and assuming that the school with the highest SAT Math scores will have no appreciable difference between the scores of students in their physics courses and the scores of all students, we can make some very crude guesses about how Fig. 3 might change if we had data for physics classes rather than for incoming freshmen. With three 25th percentile score conversions (430 for all freshmen to 520 for physics students; 580 to 650; 750 to 750) we can fit a quadratic

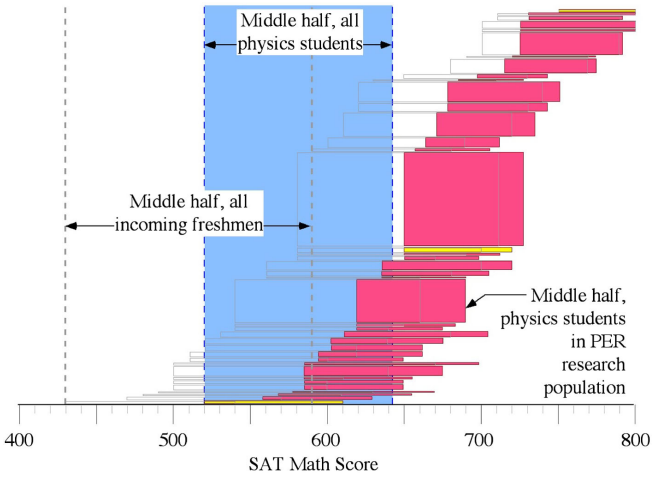

FIG. 7. Our best guess about the overlap between research population and overall physics student population. Based on our knowledge of how SAT Math scores change from incoming freshmen to students in calculus-based physics courses for 2 institutions (shown in yellow), and on the assumption that the scores do not change for the school with the highest scores (also in yellow), the 25th–75th percentile ranges shown in Fig. 3 have been mapped to predicted ranges for students in calculus-based physics courses. Score mapping is done by quadratic fit. The data for incoming freshmen from Fig. 3 is shown outlined in light gray for comparison, as are the limits for the middle half range.

to our data, and use this quadratic to map the 25th percentile scores for all schools. We can do the same for the 75th percentile (540 to 610; 700 to 720; 800 to 800). We use the same mapping for the 25th and 75th percentile range for all physics students. The results of this mapping are shown in Fig. 7.

The most noticeable result of our mapping is a narrowing of the score range for the middle half of students (both for the research population schools and to a lesser extent for the overall student population, since the 75th percentile increases less than the 25th percentile. There is relatively little change in the overlap between the research population and the overall student population, and there is still a clear mismatch in levels of preparation as measured by SAT Math scores. We do not believe that our conclusions about this mismatch would change if we had access to scores for physics students rather than for incoming freshmen. The demographics of physics classrooms is currently under study by Amy Robertson and Raphael Mondesir, who are looking at data taken from 9 universities. We anticipate that their results will help to determine what can and cannot be said about the demographics of physics classrooms based on university-wide demographics [24].

As a final note about uncertainties in SAT Math scores, our use of incoming freshmen scores overlooks the scores of students who transfer to four-year institutions from two-year colleges. About one-quarter of the students who enter a two-year college transfer to a four-year institution within five years. Some of these students will have already taken physics when they transfer, and some will take physics after transferring. Depending on institutional admissions policies, these students may not be using SAT scores (even if they did take the SAT in high school) for their admission on transfer. Without more information, it is hard to compare the level of preparation of these students to non-transfer students. However, we believe that the transfer rates are small enough that any effects do not alter the claims we have made.

IV. DISCUSSION

The results reported in this paper suggest that American physics education research generally relies on research subjects who in many respects are not representative of the overall population of physics students. As a result, the PER community needs to be very careful about the assumptions that we make when interpreting the data that we have. While we are not aware of researchers explicitly claiming that their results are generalizable to the overall population of American physics students, there is usually an implicit expectation that the results obtained by a researcher in one physics classroom should be replicable in another physics classroom: The idea that "my students = your students" might be considered the zeroth law of PER, because it is an unspoken assumption that underlies much of our discourse. Based on our results, however, we as a community do not have enough data to confidently extend our research results to a majority of physics students, including high school students, students in two-year colleges, and students at less selective four-year institutions.

The effect of a research focus on a non-representative research population is hard to assess. It might be that for some studies there is no effect at all, and repeating the same research on a more representative population would yield the same results. For example, we suspect that student responses to conceptual questions about Newton's 3rd law or about electric current in circuits would be similar across introductory student populations. On the other hand, questions that depend on an understanding of the relationship between a quantity and its rate of change might be strongly dependent on mathematical proficiency, and we might expect to see significant variation in responses from one student population to another. For questions about expectations, epistemology, or affect we might expect to see variations for other reasons. But these are just guesses: If the *intent* of research-based curriculum development or the creation of theoretical models is to benefit physics students *in general*, then the PER community needs to be more attentive to demographics. On the other hand, if the intent is to describe, modify, and model the behavior of a subset of this population, then our community needs to establish norms for describing these subsets, and to be more specific about limitations on the applicability of our findings.

The disparities described in this paper between the PER research population and the general population of physics students are not a result of individual ill intent, and our goal in highlighting them is not to assign blame or to criticize past researchers for their choices of research subjects. As with other systemic biases, however, the absence of individual ill intent does not mean that real harm is not done, that existing inequities are not exacerbated, or that we are continuing to reinforce implied notions of what "normal" physics students look like. Nor does this absence absolve us of responsibility for fixing the problem.

In this discussion section, we first outline the benefits that have accrued to PER as a result of working with a relatively homogenous research population, and outline the consequences associated with continuing to conduct research on a nonrepresentative student population. We then suggest some steps that can be taken to improve our demographics.

A. Benefits of working with a nonrepresentative sample population

The perspective we have adopted is that in many ways the homogeneity of our research population has been to the benefit of PER, by generating less ambiguous research results than would otherwise have been possible. In turn, the curricular materials developed on the basis of this research have been demonstrably successful at improving instruction at many institutions. The rapid growth of PER over the years of our study, and the widespread adoption of research-based materials is likely in large part a consequence of limiting the variability in our student research populations.

Physics education research has, up to now, been conducted primarily in what we consider to be a simplified context that includes calculus-based courses and well-prepared students with relatively homogenous and privileged backgrounds. We do not argue that the results obtained and the theoretical models proposed by physics education researchers lack value because they have failed to take population differences into account and because they have tended to focus on a well-prepared homogeneous population. A useful analogy can be made to the sequence of research in a given field, where initial studies are often restricted to situations where phenomena are simplified and many variables are eliminated. For example, early climate change studies ignored effects of interactions between oceans and the atmosphere, and did not include effects of changes to the planet's biomass [25]. Only after the models developed with these simplifying assumptions were better understood did researchers start to include more complicated and realistic scenarios. Similar strategies are employed when we teach physics: When Newton's laws are taught, for example, first we teach students in idealized situations without resistive effects. Once students have learned how to apply Newton's laws in simple situations, we add a few more selected variables such as friction and air resistance, and then we might include multiple dimensions or different coordinate systems.

PER is now reasonably well established as a field, however, and it is our sense that the time is ripe to start to explore the effects of variations in student population. It is possible that our results remain unchanged when many of the studies that we have already conducted are repeated with different student populations. On the other hand, we may see dramatically different results, which will allow us to focus on the causes of these differences. We expect that additional influences to overall findings will be discovered, much in the same way that adding friction to a mechanics problem changes final velocities without eliminating the known effects of mass and gravitation.

Perhaps the most important reason for doing so is to improve instruction for *all* physics students by better understanding the impediments and affordances that students of diverse backgrounds bring to physics learning. A deeper understanding of variations in population and preparation will allow the development of curricular materials that are better matched to specific populations and that are more likely to be successful for the general population of physics students. Our understanding of the process of adoption of materials will benefit from paying attention to the effects of population differences. In addition, theoretical models of student thinking will need to become more nuanced in order to account for population variations, and the quality of these models will improve as they are tested against more general student populations. Exploring variation in population is a tremendous opportunity for the PER community, and will yield enormous benefits for our students and for our understanding of physics learning and teaching.

In addition, efforts to systematically compare results across populations will allow refinement of our understanding of what matters in terms of preparation for physics instruction, and of which descriptors of student background are useful and which are not. From our perspective, the physics education research community is now at a point in the overall research enterprise where we would greatly benefit by explicitly including population variability in our studies.

B. Recommendations for improving PER demographics

The results of studies in other fields suggest that improving the diversity of community-wide research samples is difficult, and it requires sustained effort. In genomics, for example, Bustamente *et al.* [26] found in 2009 that only 4% of all subjects included in genome-wide association studies were of non-European descent. In a follow-up study in 2016, Popejoy *et al.* [11] found that although there was a 20-fold increase in the number of genetic samples included in these studies, the fraction of

subjects of African descent had increased by only 2.5%, and of Latin-American subjects by only 0.5%. The fraction of samples from Native American, Aboriginal Australian, and Pacific Islanders actually decreased. (The percentage of non-European subjects increased from 4% to 19%, largely due to an increase in the number of studies from Asian countries. [26].

We can expect that improving the demographics of the PER student population will be difficult as well, and that improvement will only come with a sustained and community-wide effort. In this section we suggest actions that individuals and groups can take that will contribute to this improvement. Many of these suggestions are based on recommendations made in the references cited previously describing demographic research in behavioral science and genomics [10,11,26].

Recommendation 1: Describe student populations as part of our research communications

How should the physics education research community move toward accounting for population differences? A helpful first step would be improving our research community's characterization of our research subjects: We should be more explicit about who our populations are, including details about them that we think influence the results we are seeing. This will help the PER community, and others, to gauge the generalizability of our research findings. Currently, the description of institution and population sample is often vague, for example "Our study was conducted at a large Midwestern university," and sometimes we do not even learn what courses the students were enrolled in. If we hope to gain a better understanding of the variation in student responses from one population to another, it is important to undertake a characterization of the research population that provides a basis for comparison. This involves some guesswork about the factors that may be relevant: for example, it might be that for some studies race, gender, and sexual orientation are very important, while for other studies they are not. We expect that the physics education research community will become more adept at selecting relevant characteristics if we increase our focus on variation across populations, and if more research is conducted with a less homogeneous student population.

We suspect that part of the reluctance to offer more details about our student populations and their preparation stems from a recognition that as researchers we have an ethical obligation to protect the privacy of our research subjects, and possibly also from a desire to shield our institutions from potential embarrassment if student responses are not what we might hope for. However, the tendency to be overly vague is not without cost. A potential adopter of materials needs to be able to understand what the sample populations are in order to evaluate how the results might apply to their population of students. Without sufficient information about the research population, how can we understand the results and then assess if we would or would not expect similar results with *our* populations of students?

Recommendation 2: Develop better measures of student preparation for learning physics

Because the prevalence of PER-based materials have been developed on the basis of research conducted on better-prepared students, it is difficult to determine whether PER-based materials are going to be as effective with students who are less prepared. In addition, because PER has not looked in detail at variations in student populations, it is not really clear what *better preparation* for physics courses really means. Is it simply more effective traditional mathematics preparation? Are there correlations with scientific reasoning skills? Are there correlations with other cognitive skills such as spatial reasoning? Does better preparation primarily mean a difference in epistemological development and stance? Does it depend on past experience with more challenging problems? Are there strong cultural effects? These are all open, interesting, and probably difficult questions, and an improved understanding of the differences in student responses from one population to another will help to answer them.

Recommendation 3: Increase the number of replication studies

PER needs replications of well-known studies that include diverse sample populations. Though scientists in general recognize the importance of replicability as a basic tenet of scientific practice, this recognition often does not translate to useful incentives for actually conducting replication studies [27]. Tenure committees, funding agencies, grant proposal evaluators, and journal reviewers and editors all tend to reward innovation, and the culture of academia is such that across science, very few replication studies actually take place [28]. However, a better understanding of the applicability of PER's published research results will require replication studies done with more representative student populations. Increasing the number of replication studies will require changing the expectations of both tenured and untenured faculty, of journal editors and reviewers, and of funding agencies. At the request of congress, a committee of the National Academies of Science, Engineering, and Medicine has been formed to explore issues of replication and repeatability in research. Their report, *Reproducibility and Replicability in Science* was published in May 2019 [29].

The NSF has published *Companion Guidelines on Replication and Reproducibility in Education Research* as a supplement to their more general guidelines for education research [30]. The guidelines note the special challenges of replication in an education research due to the variability of research contexts. In addition, they offer guidelines for effective replication studies, and for designing nonreplication studies so that replication is easier. For example, they recommend that original data and sufficient

details of the analysis conducted should be made available. They also note the importance of documenting the features of the student population, consistent with recommendation 1.

Some funding agencies, recognizing that replication studies are difficult to fund through traditional programs, are beginning to provide research support through programs targeted at replication. For example, the Dutch Research Council (NWO) has a Replication Studies Program that funds "cornerstone research" [31]. While we are not aware of equivalent earmarking of funding opportunities for replication at NSF, the Directorate for Social, Behavioral, and Economic Sciences (SBE) specifically encourages replication studies for submission [32].

The role of science journals in increasing the prevalence of replication studies cannot be overstated. It may be useful for editorial boards to set specific goals for content dedicated to replication studies. In addition, by rewriting guidelines for reviewers, journals can explicitly request that these studies be evaluated on the basis of their scientific value rather than using originality as an overarching criterion for publishing decisions.

Recommendation 4: Provide incentive for including diverse sample populations

In the long run, it is important for our research community to make sure that we are not ignoring entire student groups. We should ensure that we have chosen research subjects in a manner that ensures that our results apply to all students who might potentially benefit. The Belmont report [33] that summarizes the basis for ethical research involving human subjects includes a "principle of justice" that makes this obligation clear:

> *The choice of participants in research needs to be considered carefully to ensure that groups (e.g., welfare patients, particular racial and ethnic minorities, or persons confined to institutions) are not selected for inclusion mainly because of easy availability, compromised position, or manipulability.*
> *In order to achieve an equitable distribution of the risks and potential benefits of the research, investigators must determine the distribution of different groups (men and women, racial or ethnic groups, adults and children, age, etc.) in the populations that … are anticipated to benefit from the knowledge gained through the research.*

While individual Institutional Review Boards probably work to ensure that equitable choices of research subjects are made *within* each institution, the fact that PER is typically conducted at more selective and homogenous institutions means that as a nationwide systemic issue, there is little that is done to ensure equitable distribution of the benefits of what is usually publicly funded research.

Recommendations for diversifying research populations in other fields emphasize the role of funding agencies, and many of these recommendations apply to PER as well. Bustamente *et al*. [26] recommend that peer reviewers and granting bodies stress the importance of racial and ethnic diversity in medical genetics studies. They note that even though the National Institutes of Health mandated the inclusion of diverse subjects in 1985, only 7% of genome-wide association studies have included minority subjects, which suggests that the mandate has been largely neglected in the awarding of funding. Popejoy [11] recommends a funding prioritization on behalf of granting agencies that rewards proposals that intend to study underrepresented populations. In the WEIRD people paper, Heinrich [10] makes a similar suggestion, noting in addition that there are typically also added costs associated with population diversification that should be taken into account.

Recommendation 5: Reframe discussions about race and privilege

PER's focus on students who are wealthier, Whiter, and better mathematically prepared than the overall student population reinforces the notion that the students that we study are normal physics students and other students are deviations from this norm. (This mindset was exemplified by Supreme Court Justice Roberts' question in *Fisher vs University of Texas* about what diverse students bring to a physics classroom—a question that contains an implicit assertion that a default physics classroom contains no racially and/or ethnically diverse students [34].) This framing goes hand in hand with a "deficit model" of racially and/or ethnically diverse student performance in physics classes, where analysis of racially and/or ethnically underrepresented students' participation is viewed in terms of how they compare to White students. With this perspective, racially and/or ethnically diverse students are likely to be framed as lacking in some physics trait when compared to the implied norm' and intervention involves changes to instruction to normalize these students. It is more useful to approach differences in results, as a function of population, as simply that: differences. We should not be implicitly trying to understand how to make our population more like some assumed "norm" when differences do emerge. Rather we should be allowing ourselves to value differences and to recognize that the populations themselves are adding valuable knowledge and contributions to our overall understanding just as they are.

The relative racial, gender, and socioeconomic homogeneity of the overall physics community has increasingly become a focus of the physics education research community. The causes of this lack of equity are systemic and numerous, and obviously go beyond considerations of participation in research [35]. However, we believe that working towards equity in our research would be a great contribution to promoting inclusiveness in physics and in the disciplines that require physics, as this equity will provide the baseline information that we need to increase the likelihood that *all* students are more successful in their

physics courses and are thus more likely to pursue physics-related careers.

Recommendation 6: Support faculty professional development on equitable practices that support diverse populations

Student demographics are changing in physics classes at all levels. In high schools, two-year colleges, and universities, more women are enrolling in physics, as are more students from traditionally underrepresented student populations. This is a trend that we would like to encourage and if possible to accelerate, and the physics community would be well served by better understanding how to optimally serve (and in some cases not actively hinder) a student population that is more varied in terms of background and expectations. This will require professional development that is specifically geared toward fostering equitable practices that support diverse populations. For researchers in particular, additional professional development on implicit bias and its possible effects on the generation of research questions and on the interpretation of results would be helpful.

Some recent conferences have focused on issues related to creating more welcoming physics classrooms and communities. On January 5th, 2018, the National Society of Hispanic Physicist (NSHP) and the American Association of Physics Teachers (AAPT) hosted the Conference on Enhancing Physics Programs at Hispanic-Serving Institution (CEPP-HSIs). Though this conference was specifically focused on Latinx populations, some of the recommendations that were generated and outlined in the conference report [36] are directly aligned with the need of broadening participation.

Unfortunately, it is often the case that faculty members who are least aware of the challenges facing nontraditional student populations in physics are also the least likely to actively seek out professional development that supports equitable practices. For organizations hoping to improve the academic climate (including but not limited to physics departments, schools, professional societies, and funding agencies) this means that beyond providing effective professional development, there is a need to provide some incentives for participation. (A similar dissemination issue exists with instructional materials in physics—faculty members whose instruction could benefit from research-based approaches may not be aware that these materials exist. Since 1996 AAPT has been running NSF-supported New Faculty Workshops for physics faculty to attempt to improve adoption rates [37]. It may be that a similar workshop-based approach would be effective at improving the atmosphere for underrepresented students in physics.) We expect that the selection of strategies for increasing participation in and effectiveness of equity-oriented professional development will become better understood as more and more institutions focus on these systemic issues.

Recommendation 7: Garner support from professional societies and funding agencies

Funding agencies and professional societies have the ability to influence institutions of higher education as well as individual researchers. It is vital for these entities to be intentional when it comes to equitable practices: Professional societies should actively advocate for improved and increased collaborations between institutional types. For example, the report for the CEPP-HSI [36] has suggested professional societies help departments make connections to build bridges to other institutions and other disciplines. This could mean that subcommittees help develop guidelines for creating partnerships between research institutions and teaching institutions and/or two-year colleges. These guidelines could include topics that address shared goals and visions as opposed to one institution dominating the goals of a grant. These guidelines could also include guidelines designed to prevent one institution from abusing another institution that might have more diverse populations that are only interested in partnerships to fulfill broader impact clauses from funding agencies.

V. CONCLUSION

In many ways, results of physics education research have shaped the way education has evolved over the past few decades—from content, to delivery, to classroom layouts, etc. However, as a research community, we have not been sufficiently attentive to whether these robust, impactful results have applied to *all* students. We have implicitly assumed that the populations that we have researched are representative. But the preparation, motivations, and goals of students at a nonresearch university in a predominantly rural state are likely quite different from those of students growing up in cities and attending highly competitive research institutions. Moreover, the resources available to instructors at these institutions are likely to be different as well. Published studies that gloss over the disparities that exist between these groups of students might compare the performance of these groups of students without any discussion of issues related to population. This leads to results that are likely not reproducible and that do not form a solid basis for future research.

It is our hope that the PER community increasingly approaches studies with the question "What differences in results are due to variations in student population?"

We recognize that many of the mismatches that we describe in this paper between PER subjects and physics students overall might not be surprising to many education researchers or more broadly to physicists and educators. Our hope is that describing the mismatches that we are aware of in some detail and quantifying them to some extent will promote discussion of these systemic issues and will prompt both individual and institutional action to steer PER toward more representative selection of research subjects. The PER community is at a point where we can acknowledge that there is a problem when the demographics of a study are not taken into account.

ACKNOWLEDGMENTS

The authors would like to thank Susan White and Patrick Mulvey and the American Institute of Physics Statistical Research Center for their help with compiling statistics for overall physics enrollments. We would also like to thank Ramon Barthelemy for his help in identifying relevant previous studies and for his careful reading of our manuscript.

[1] L. C. McDermott and E. F. Redish, Resource letter: PER-1: Physics education research, Am. J. Phys. **67**, 755 (1999).
[2] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).
[3] D. E. Meltzer and V. K. Otero, A brief history of physics education in the United States, Am. J. Phys. **83**, 447 (2015).
[4] P. R. L. Heron and D. E. Meltzer The future of physics education research: Intellectual challenges and practical concerns, Am. J. Phys. **73**, 390 (2005).
[5] National Research Council, *Adapting to a changing world: Challenges and opportunities in undergraduate physics education* (National Academies Press, Washington, DC, 2013).
[6] P. G. Hewitt, Millikan Lecture 1982: The missing essential—a conceptual understanding of physics, Am. J. Phys. **51**, 305 (1983).
[7] A. Eisenkraft, Millikan Lecture 2009: Physics for all: From special needs to Olympiads, Am. J. Phys. **78**, 328 (2010).
[8] A. Hobson, Millikan Award Lecture, 2006: Physics For all, Am. J. Phys. **74**, 1048 (2006).
[9] J. Tobochnik, 2017 Oersted Medal Presentation: The changing face of physics, and the students who take physics, Am. J. Phys. **85**, 409 (2017).
[10] J. Henrich, S. J. Heine, and A. Norenzayan, The weirdest people in the world?, Behav. Brain Sci. **33**, 61 (2010).
[11] A. B. Popejoy and S. M. Fullerton, Genomics is failing on diversity, Nature (London) **538**, 161 (2016).
[12] Data from the American Institute of Physics https://www.aip.org/statistics.
[13] http://www1.unipa.it/girep2014/proceedings/GIREP-MPTL%202014%20Conference%20Proceedings.pdf.
[14] J. Res. Math. Educ. **49** (2018); **48** (2017); **29** (1998); **28** (1997).
[15] https://www.aip.org/sites/default/files/statistics/undergrad/tyc-enrollments-p-11.pdf.
[16] Information from the College board about individual colleges and universities including SAT Math scores can be found at https://bigfuture.collegeboard.org/college-search.
[17] V. Coletta, J. Phillips, and J. Steinert, Interpreting force concept inventory scores: Normalized gain and SAT scores, Phys. Rev. ST Phys. Educ. Res. **3**, 010106 (2007).
[18] S. Hsu and J. Schombert, Nonlinear psychometric thresholds for physics and mathematics, arXiv:1011.0663v1.
[19] D. E. Meltzer, The relationship between mathematics preparation and conceptual learning gains in physics: A possible "hidden variable" in diagnostic pretest scores, Am. J. Phys. **70**, 1259 (2016).
[20] R. Zwick, *Fair Game?: The Use of Standardized Admissions Tests in Higher Education* (Psychology Press, London, 2002).
[21] C. Miller and K. Stassun, A test that fails, Nature (London) **510**, 7504 (2014).
[22] College Board SAT Scores from http://media.collegeboard.com/digitalServices/pdf/research/Total_Group_Report_CBS_08.pdf.
[23] Equality of Opportunity Project data at: http://www.equality-of-opportunity.org/data/.
[24] A. Robertson (personal communication).
[25] S. R. Weart, *The Discovery of Global Warming* (Harvard University Press, Cambridge, MA, 2008).
[26] C. D. Bustamante, E. G. Burchard, and F. M. De La Vega, Genomics for the world, Nature **475**, 163 (2011).
[27] J. C. Travers, B. G. Cook, W. J. Therrien, and M. D. Coyne, Replication research and special education, Remed. Spec. Educ. **37**, 195 (2016).
[28] J. J. Van Bavel, P. Mende-Siedlecki, W. J. Brady, and D. A. Reinero, Contextual sensitivity in scientific reproducibility, Proc. Natl. Acad. Sci. U.S.A. **113**, 6454 (2016).
[29] National Academies of Sciences, Engineering, and Medicine, *Reproducibility, and Replicability in Science* (The National Academies Press Washington, DC, 2019).
[30] https://www.nsf.gov/pubs/2019/nsf19022/nsf19022.pdf.
[31] https://www.nwo.nl/en/news-and-events/news/2017/social-sciences/repeating-important-research-thanks-to-replication-studies.html.
[32] https://www.nsf.gov/pubs/2018/nsf18053/nsf18053.jsp.
[33] K. J. Ryan, J. Brady, R. Cooke, D. Height, A. Jonsen, P. King, K. Lebacqz, D. W. Louisell, D. W. Seldin, E. Stellar, and R. H. Turtle, *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Washington, DC, 1979).
[34] D. MacIsaac, AAPT committee on diversity issues statement on Fisher v. University of Texas at Austin, Phys. Teach. **54**, 379 (2016).
[35] D. F. Carter, J. E. R. Dueñas, and R. Mendoza, Critical Examination of the Role of STEM in Propagating and Maintaining Race and Gender Disparities, in *Higher Education: Handbook of Theory and Research* (Springer, Cham, 2019), pp. 39–97.
[36] https://www.aapt.org/Resources/upload/AAPT-NSHP-HSI-Report-110125.pdf.
[37] S. V. Chasteen, R. Chattergoon, E. E. Prather, and R. Hilborn, Evaluation methodology and results for the new faculty workshops, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. L. Jones, L. Ding, and A. Traxler (AIP, New York, 2016).