Data sharing model for physics education research using the 70 000 response Colorado Learning Attitudes about Science Survey for Experimental Physics dataset

John M. Aiken[1,2,3] and H. J. Lewandowski[3,4]

[1]The Njord Centre, University of Oslo, 0371 Oslo, Norway
[2]Centre for Computing in Science Education, University of Oslo, 0371 Oslo, Norway
[3]JILA, National Institute of Standards and Technology and the University of Colorado, Boulder, Colorado 80309, USA
[4]Department of Physics, University of Colorado Boulder, Boulder, Colorado 80309, USA

(Received 7 April 2021; accepted 18 October 2021; published 20 December 2021; corrected 19 January 2022)

We present a model for sharing quantitative data in the field of physics education research and use it to present a newly available dataset as an example. This model is in line with calls from across physics and science more generally to democratize data and results through open access. The model includes suggestions for data collection, creation of a data schema, and data sharing. It attends to the specific needs of the physics education research community, such as anonymization of human subjects data. As an example of this model, we use the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS) dataset, which includes over 70 000 responses to the E-CLASS survey. These data cover 133 institutions, 599 unique courses, and 204 instructors, and was collected between 2016 and 2019. These data are made available at the time of publication and can be used freely, without the need of any institutional review board approval.

DOI: 10.1103/PhysRevPhysEducRes.17.020144

I. INTRODUCTION

There are recent calls from the physics community for free access and open sharing of data (e.g., Refs. [1,2]). The United States Congress has taken up the issue [3] and the European Union has issued a strong mandate (known as Plan S [4]) towards publicly funded research being open access, both in the dissemination of results and in the raw data collected. The U.S. National Science Foundation's own rules require that the "primary data" gathered using NSF support must be freely shared "within a reasonable time" [5]. The American Institute of Physics (AIP) states "that all datasets underlying the conclusions of the paper should be available to readers" [6].

Within physics education research (PER), as with other fields of science, data sharing is becoming more common, but access to data could still be improved. PER, in particular, has had profound impacts on university physics learning through collecting and analyzing large datasets (e.g., Ref. [7]). However, sharing of these data have significant barriers due to the privacy rules surrounding student data, and, historically, raw data have not been broadly available to all PER researchers. We suggest that the practice of data sharing in PER could be expanded to help broaden the use of the data for research, and thus have greater impact on physics education.

Our goals for this paper include (1) introducing a framework for sharing data within PER and (2) using this framework to present a newly available dataset as an example. First, we present a framework for sharing large, quantitative datasets in PER. This model assumes that the data being shared are quantitative, such as tabulated pre- and postconcept inventory data, student information system data, or click data from learning management systems. It is not designed for qualitative data, such as video observations or interviews. The model suggests norms for data sharing, such as articulating the level of the data (e.g., data about an institution compared to data about an individual student), how to discuss the anonymization of the data, and also the organization, or schema, of the data. We suggest a productive model of data sharing supports open access to data, as it allows the reuse of previously collected data to answer new research questions, as well as studies to reproduce results [8]. Second, as an example of this framework, we present the 70 000 response Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS) dataset. This dataset is free to access for all researchers and contains pre- and postresponse data, demographics, course information, and other data for students attending 133 institutions of higher education. The courses represented are introductory

and upper-division laboratory courses. These data have been used to explore many research questions related to, for example, gender and labs [9] and the effect of instructional approaches on student views about experimentation [10,11]. The dataset has also been used to demonstrate the effectiveness of centrally administrated data collection in PER [12]. Making this dataset open and available to all researchers leverages the significant efforts required to collect such data to allow for studies of unexplored research questions by researchers not involved in the data collection. Our hope is that this will empower researchers with fewer resources and/or at smaller institutions to be able to pursue innovative research programs without the barriers of having to collect significant amounts of data themselves.

II. BACKGROUND

This paper aims to present a framework for data sharing in the field of physics education research (PER). In this section, we discuss the benefits of data sharing, a brief history of data sharing and datasets in PER, and documented issues in sharing social science data.

A. Benefits of data sharing

Sharing of PER data does take effort and, depending on the dataset, it may require significant time and resources to be able to make the data widely available. Thus, we must understand the benefits that may be derived from data sharing. To discuss the benefits, we turn to ideas posited by PLOS ONE, an open access, broad topic journal. This journal requires that all researchers submit open and freely available data that are related to any study published in the journal and has clearly articulated some reasons why data sharing is beneficial to science. According to PLOS ONE (and replicated below), making data publicly available allows for the following [13]:

1. Validation, replication, reanalysis, new analysis, reinterpretation, or inclusion into meta-analyses;
2. Efforts to ensure data are archived, increasing the value of the investment made in funding scientific research;
3. Reduction of the burden on authors in preserving and finding old data, and managing data access requests;
4. Citation and linking of research data and their associated articles, enhancing visibility and ensuring recognition for authors, data producers, and curators.

Across social science fields there has been a replication crisis [14]. Many studies when revisited cannot produce the same result (e.g., within PER, Aiken *et al.* [15] demonstrated courses taken, not grades as previously demonstrated in Aiken and Caballero [16], is most predictive of whether or not a student remains in a physics degree program). In some cases, if data were freely available, studies at different institutions could be compared to see if the results could be replicated. Data sharing also encourages the long-term preservation of data, which maintains data integrity and can serve as training tools for future scientists [17]. Free and open data also encourage conversations around specific research questions. Results can be reanalyzed with different methods to further establish results.

Creating public datasets has had varied popularity across different fields focused on social science questions. In psychology, there is often poor availability of data for researchers outside of the original study [18]. A lack of public datasets is likely a strong contributing factor to the replication crisis [14]. It is important to avoid this outcome in PER. A replication crisis in PER would likely motivate departments to not adopt research-based pedagogy.

In the broad field of machine learning, it is common to have freely available datasets for a variety of tasks (e.g., Ref. [19]). These datasets are well known and, thus, are useful to test new machine learning models against data that is accessible to everyone. In PER, embracing an open data sharing policy can be used to leverage both research results and policy.

In addition to these points, we would also like to highlight that sharing data helps to democratize PER. Currently, PER is done primarily at large-enrollment research universities [20]. In some cases, this restricts the demographics of the students who are being studied. But it also has the effect of restricting the demographics of researchers by limiting access to data. By providing open and freely available data, we provide opportunities to participate in research for researchers who do not work at large-enrollment research universities. This point is particularly salient for quantitative data, where many analysis methods require a large amount of data.

B. Data sharing in PER

Historically, data from high-impact PER studies (e.g., Ref. [7]) have not been shared. In most cases, raw data from PER studies are not available. This leads to PER studies, which attempt to replicate previous work, relying on data reported in tables, since the raw data are unavailable (e.g., Ref. [21]). While laudable, secondary analysis can be problematic because it becomes impossible to account for correlations that may be found in the raw data unless they are directly reported. Instead, secondary analysis should ideally be built from the raw data of previous studies [22]. In the case that identifiable data are necessary to answer a research question, one must go through institutional review board approval in the U.S. or a similar process in other countries. However, many questions with regards to student learning, attitudes, and pathways through physics do not require identifiable data.

A common result in PER has been that interactive engagement increases conceptual learning in physics [7]. However, the exact effect sizes of factors that increase

engagement are not well understood. If there was a large, public dataset of concept inventory responses along with classroom description data, we could begin to build a testable and repeatable model of the alleged causal relationship between interactive engagement and physics content acquisition. These research results could then be leveraged towards influencing departmental policy. With a public dataset, the expected increase in learning at a single institution due to pedagogical changes could be compared directly to a national or international data metric. This would allow for direct examination of both the stability of learning increases due to pedagogical changes and an examination of when these pedagogical changes do not produce the desired effect. We believe there are currently two existing broad efforts that seek to accomplish this type of outcome.

Recently, there have been two large-scale initiatives to collect multi-instrument concept inventory and survey data in PER. The first large-scale data collection has been through the Learning About STEM Student Outcomes (LASSO) project [23]. The LASSO project implements a central storage and data collection framework similar to the ECLASS framework [12,24]. Instructors sign up to include a concept inventory in their course and provide a number of descriptors, such as estimated student enrollment, institutional descriptions, and a description of the course being taught. Students are then delivered the concept inventory online through the LASSO system. To access data via the LASSO system, researchers need to have local IRB approval to be able to purchase the data [25].

The second broad effort is the PhysPort system. PhysPort is primarily a website for "physics educators to learn to apply research-based teaching and assessment in their classrooms" [26]. PhysPort also includes a "Data Explorer" page, where faculty can upload survey data from their courses and compare their data to national responses. These data are then compiled into a dataset that is shareable. To access this dataset, researchers are required to apply to their local IRB first, then submit this IRB approval to the PhysPort administrators who will then review it and determine what data they are able to share [27].

In addition to these projects that collect data from multiple survey instruments, there are two well-established single-instrument systems, the E-CLASS and the Physics Lab Inventory of Critical thinking (PLIC) [28]. Both of these data collection efforts are built on similar core processes and code. We will describe the E-CLASS system in detail below.

C. Challenges for data sharing

Data sharing cannot be done freely without first attending to some challenges. These challenges fall into two groups: (1) administrative challenges and (2) privacy challenges. Administrative challenges include finding appropriate technological solutions to share data affordably and reliably. Recent efforts across science have been made to create free data repositories for researchers to store data in, such as open-source repositories like Harvard Dataverse [29], Data World Bank [30], the U.S. Government's data.gov [31], CERN's Open Data warehouse for particle physics data [32], or the Open Science Framework [33]. Resources such as the Open Data Handbook describe different data sharing procedures for open data [34]. These procedures include what license should be used for sharing the data [for example, the ECLASS dataset shared in this paper uses the Open Data Commons Open Database License (ODbL) v1.0 [35]]. These efforts have eliminated the need for funding locally hosted computing resources to store and share data. In addition to finding a way to share data in a reliable way, one needs to be make sure the data shared openly are no longer considered human subjects data.

Data privacy is a complex issue. Typically, data such as names and telephone numbers are never released because these create obvious identifiers within data. Groups of specific data types (e.g., zipcode, date of birth, gender) can be used in some cases to identify a person when other public data (such as census data) are available [36]. In highly sensitive contexts, such as medical research, data can be released using sophisticated statistical information methods [36]. In the case of PER, it is unlikely to be the case that these sophisticated methods are necessary because students rarely divulge private information that would compromise their lives or livelihoods. However, it is important that data are anonymized or deidentified at the appropriate level to protect both the students who are learning and the pedagogical practices of the instructors. This process should be done in consultation with an IRB.

D. E-CLASS

One common goal for instructors of physics lab classes is for their students to develop scientific habits of mind and expertlike epistemology around experimental physics. To assess the impact of lab instruction on these views, the E-CLASS was developed [37] and validated for all levels of college physics [38]. The E-CLASS measures student epistemologies and expectations for experimental physics in lab classes. The E-CLASS is administered to students at the beginning and end of an academic term to measure the impact of instruction. The survey asks students to respond with their level of agreement (from strongly agree to strongly disagree) to two questions for 30 statements. The questions ask "What do you think when doing experiments for class?" and "What would experimental physicists say about their research?" The 30 statements include a variety of ideas surrounding experimental physics and were chosen to align with a large range of learning goals from the community of lab instructors. An example statement is "Calculating uncertainties usually helps me understand my results better." In addition to the 30 statements asked both on the pre- and postsurvey, there are

23 related statements about what is important for earning a good grade in the class, which are asked on only the postsurvey. An example question is "How important for earning a good grade was calculating uncertainties to better understand my results?" Questions about demographics, interest in physics, and career plans are also asked on the postsurvey.

There have been many research studies done with student responses to the E-CLASS [9–12,37–52]. A summary of many of these results can be found in Ref. [53]. These studies cover a wide range of research questions, including exploring the impact on E-CLASS scores from open-ended vs guided labs [11], gender [9], and skills vs concept focused labs [39]. Additional work looked at how student responses evolved as they progressed in their physics courses [44], and how course grading practices [41] and student course grades [42] correlated with E-CLASS responses. Finally, there have been several studies examining the effectiveness of lab transformations using E-CLASS for a single course [46–48,51,52]. Even with all of these previous research questions being probed, there are considerably more still to be answered. In particular, questions that require more data than was available in 2016–17, when most of the previous research was done.

III. DATA SHARING MODEL

The data-sharing model presented in this paper is based on a collection of sharing recommendations across many fields including physics, psychology, and physics education research. It is presented here as a four step process: (1) data collection, (2) data schema, (3) data anonymization, and (4) data sharing (see Fig. 1). Each process is discussed below.

A data-sharing model explicitly articulates how the data were collected, the original purpose for the collection, how the data are organized for sharing, and how to access said data. It provides better transparency for research that was done with the data collected [54]. It clearly articulates the anonymization process [55]. It provides a description of the data that are being shared, called a data *schema*, by establishing the relationships between the different data presented in the dataset [56]. Additionally, a data-sharing model provides a framework for potential research questions to be investigated by other researchers. In this section, we will discuss each of the components of a data-sharing model.

A. Data collection

In the data collection step, it is important to articulate what data were collected, how they were collected, and results that have been produced using these data so far. Data in education research have a number of constraints that should be articulated. First is what theoretical assumptions went into the data collection [57]. Second is what level of data are being collected [58]. Third is characterizing the data types and encoding of the data that are collected [59].

Ding [57] identifies that quantitative PER historically has had three genres of studies that are performed:

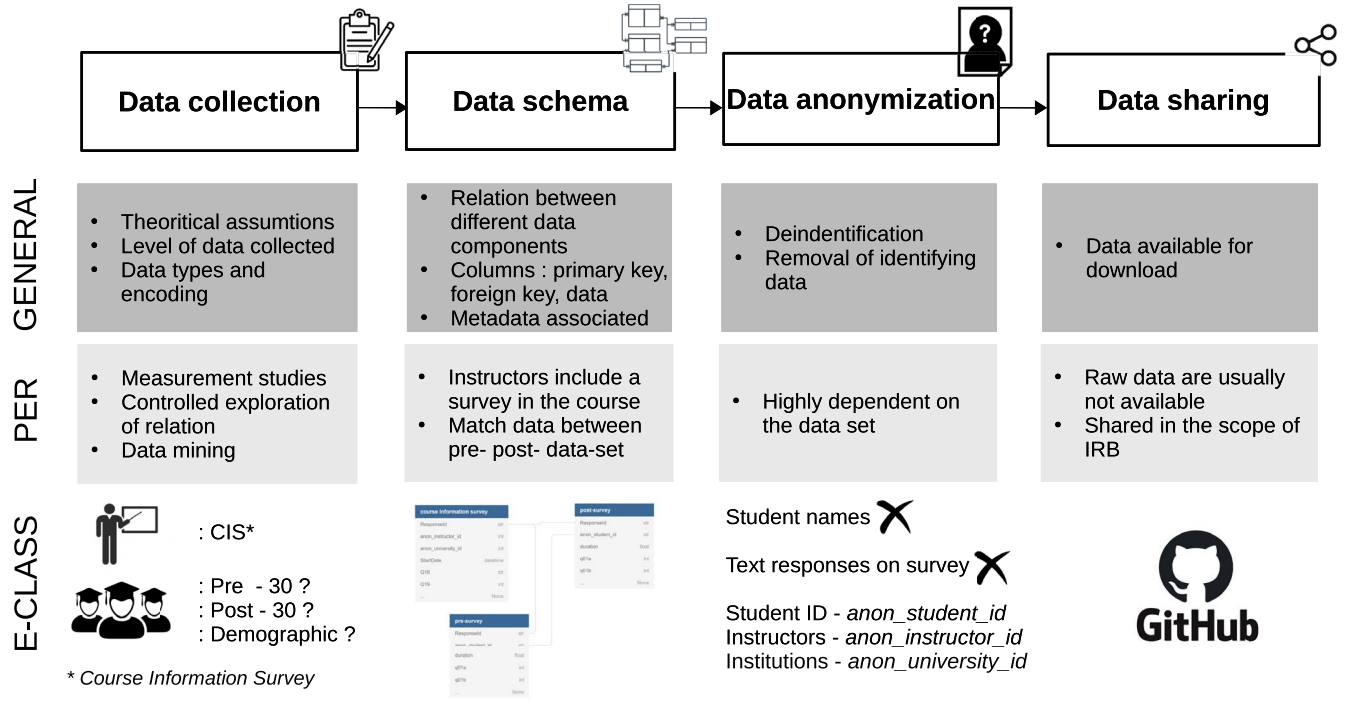

FIG. 1. A data-sharing model for PER. While data sharing occurs across all of science, PER has specific needs to make data sharing generally available, as does the E-CLASS itself.

(1) Measurement studies, such as building an understanding of student knowledge of a physics concept, (2) controlled explorations of relations where relationships between confounding factors are explored (such as conceptual understanding and social interactions in class), and (3) data mining, where researchers use data they lack control over, such as student information systems. The theoretical assumptions of what can be said and done within each kind of study is directly relevant to sharing data. For example, data gathered to examine causal relationships of student conceptual understanding and self efficacy may be data mined later and even enhanced with data from other sources. These data cannot be used to produce new causal relationships. By articulating the theoretical assumptions that went into gathering the data in the first place, researchers that are using a newly shared dataset can better understand the boundary conditions governing research questions they can ask.

Data can be collected in a variety of ways. Traditionally, concept inventory data in PER have been collected by instructors individually (often on paper forms) and then compiled into larger datasets by researchers (e.g., Ref. [7]). This model of data collection is similar to the PhysPort Data Explorer model except that the PhysPort Data Explorer has automated the dataset compilation process. Data can also be collected through centralized administration and storage procedures. This is how data for E-CLASS [12], PLIC [28], and LASSO [23] projects collect data. Beyond concept inventory data, datasets can also be compiled from institutional data (e.g., Ref. [60]). This is typically done first by accessing local institutional data, then transforming it into an analytics database that can be shared.

One way of articulating the data collection process is by describing the level at which the data are gathered. The level of data is important because it implies specific constraints about what can and cannot be said with the data being shared. Quantitative educational data can be defined as being at three levels [58,61,62]:

- **Macrolevel data**: Data concerning institution-level importance such as grades, demographics, etc. These data are frequently collected on a timescale of semesters or years.
- **Mesolevel data**: Response data from concept inventories, course observations, submissions of course assignments, etc. This level of data allows the capture of concepts and affect that students are learning and experiencing in a classroom or online environment.
- **Microlevel data**: Data from interacting with learning systems, such as video lecture clickstreams or learning management software. These data are often captured on a per second level as students click on a web platform, search through a video lecture, or interact with an intelligent tutor system.

These data-level descriptions are not necessarily completely distinct from one another and can often complement each other. For example, the way a student uses a web platform as they complete a homework assignment can contain data that are both the clicks they make and the responses to questions they provide.

Macrolevel data have been used in PER to investigate success in learning [63] or staying in the major [15]. The ECLASS dataset presented in this paper is an example of mesolevel data. Microlevel data have been used in PER to investigate how physics students learn from video lectures [64,65].

Once a good understanding of how and what data are collected, the organization, or schema, can be articulated.

B. Data schema

A schema is a description of how different components of data relate to one another. By having a schema, one can construct a dataset that is designed to easily assemble data components that answer a specific research question. Quantitative educational data are, by nature, relational [58]. Relational data are data that are formed into a collection of tables based on relationships within columns from the tables [66]. These tables can be organized such that they contain nonredundant data, which can support efficiency in database querying [67]. These relationships can be leveraged via different schemas (the structure of the tables and their relationships) and can be used to increase productivity in analytics [68]. The type of data in each column should also be clearly articulated. Different data types have different assumptions about what can and cannot be done with them [59]. There is no one correct schema to organize data. Schemas should be designed with the broadest number of research questions possible in mind.

A schema provides definitions for different types of columns in a multitable dataset. The three types of columns are (i) primary key, (ii) foreign key, and (iii) data. The primary key column represents a unique identifier for that row of data. The importance of this column is that it provides a unique identifier for the unique data in that row that connects to other data. For example, a table of student data could have the columns "ID" and "name." The ID column would be the primary key, since students can share a name. A foreign key is a unique identifier in a separate table that points to a primary key in another table. A foreign key cannot exist without a primary key. For example, if we create a second table that contains courses students take, it would have an ID column and a "course ID" column. The ID column points to the previously described student table column ID. This serves two primary functions: (1) There is a reduced amount of data needed to be stored, since student names do not need to appear in both the student table and the courses table, and (2) this process promotes data integrity. This data integrity comes from the fact that data cannot be stored in the courses table for student IDs that do not exist in the student table. Thus, there can be no

orphaned data rows. The columns "name" in the student table and course ID in the courses table are data columns.

Data are columns that have no relationship outside of the table they reside in (e.g., grades for courses in a courses table). A simple data schema for PER data would be the relationships between a pretest and post-test dataset for a single course taught by a single instructor.

Closely related to the data schema is the metadata associated with the dataset. The metadata is associated with each column in the dataset. These data provide descriptions that can help researchers know how to use each data column. Typically, the metadata is stored in a separate description file, such as an excel spreadsheet. For example, in the ECLASS dataset, columns have short ID names such as the course information survey question "Q5." Metadata found in the "question_lookup" spreadsheet gives a short description of Q5, which is whether the course is offered at institution that is on the semester or quarter system. The metadata also has the original question text that is presented to instructors.

Finally, it is important to characterize the dataset as a whole in some way. This can include summary tables and figures, description of expected clusters within the data, time lines of data changes, or any other visualizations that can help researchers understand the scope of the dataset. It is important that the goal of these plots and tables is to characterize the dataset, and not answer a particular research question.

C. Data anonymization

We do not provide detailed suggestions here, as each dataset is unique and poses different challenges to deindentification. Regardless of the process, an IRB should be consulted during the deindentification process and before the data are shared openly. Below, we describe data deindentification and provide an argument for why this process is necessary and facilitates free and open sharing of data collected in PER.

Data deindentification is the removal of identifying data for students, instructors, institutions, etc. In many cases, data may be deidentified simply by removing names and ID numbers. However, it is likely that in complex datasets deidentification will be a more complex process. Deidentification of data is a subset of the field of data anonymization. Data anonymization is not considered as a binary state and is instead on a spectrum [55]. Strongly anonymized data may take the form of aggregate tables, where students or groups with small numbers are removed (e.g., IPEDS data [69]). Data can also be anonmyized algorithmically [70]. In general, we note that with access to auxiliary databases and modern machine learning techniques some deidentified data might be partially reidentified, and so one should consider this when going through this process to limit this possibility.

By deidentifying data, it also allows for it to be shared outside the scope of the IRB (with regards to U.S. data sharing, these rules are different elsewhere). The data that were gathered can be requested to be made public in the initial IRB submission. Alternatively, it can be requested as an amendment depending on the particular situation. Deidentifying data allows for data to be shared. Sharing data reduces barriers for researchers from all institutions especially those that do not commonly work with human subjects research and may not have an IRB.

D. Data sharing

Data sharing means that the data are made available, electronically, for download for researchers to use. This could be as simple as providing a link to a github repository, as we do with the E-CLASS data [71]. For more complex datasets stored in databases, it might be useful to establish an "application programming interface" (API) that scientists can make requests against (e.g., the census data APIs [72]).

Data should be shared with the minimum number of barriers to access possible. This includes that the data should be open access and therefore free for all researchers to download and use. Additionally, the data, having been deidentified, should not require IRB review to use to answer new research questions. Other barriers could include continued funding for private storage on locally hosted servers. Thus, we recommend using free resources available on the web when possible. In the case of E-CLASS, we have used github, since there is an analysis library associated with the dataset. However, researchers should consider other data repositories (e.g., Harvard Dataverse [29]).

IV. EXAMPLE IMPLEMENTATION: E-CLASS

In addition to describing a model for sharing data for PER scientists, this paper also presents, as an example, the E-CLASS dataset. The E-CLASS dataset contains 39 505 responses to the presurvey and 31 093 responses to the postsurvey. Students can, in some cases, respond to the survey more than once. Thus, there are a total of 35 380 unique responses to the presurvey and 28 282 unique responses to the postsurvey. In this case, "unique" is defined as the first response to the pre or postsurvey. The dataset represents 133 unique universities, 204 unique instructors, and 599 unique courses (Tables I, II, and III). The dataset contains data for both students in introductory and "Beyond the First Year" courses (BFY). The total data collected per semester have increased over the course of the data collection period as shown in Fig. 2.

This dataset is freely available to download [71]. The data are stored in comma-separated files and the repository includes python-based analysis related to this paper in jupyter notebooks [73]. The repository includes additional helper jupyter notebooks, which demonstrate how to

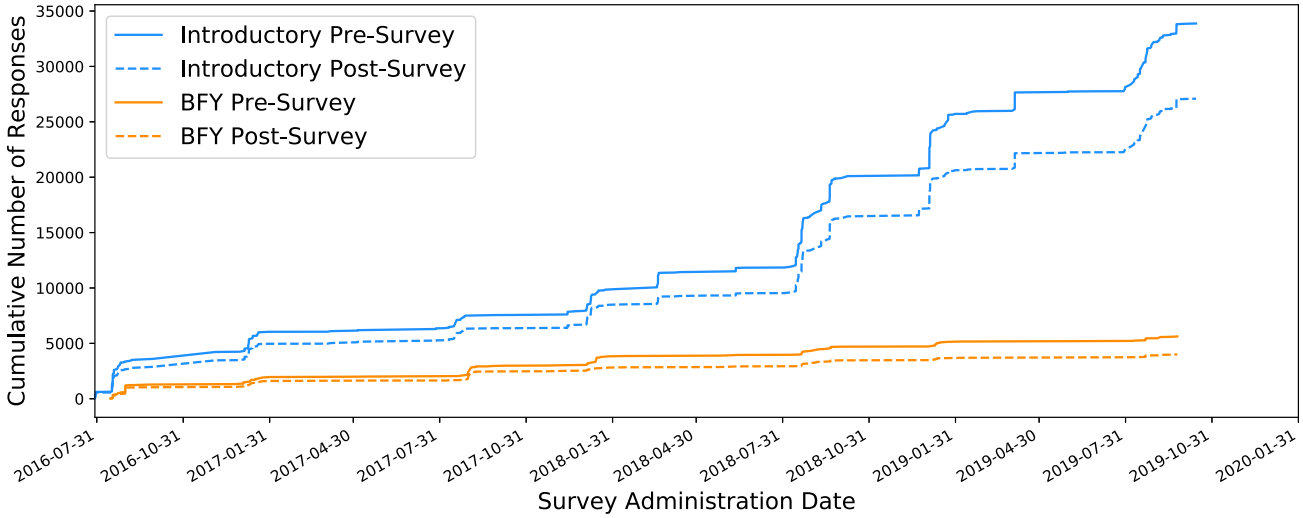

FIG. 2. Cumulative number of student responses to the E-CLASS since automated online administration began in 2016. More student responses are collected from introductory courses than BFY labs, as those classes tend to have larger enrollment. Additionally, in general, more students respond to the presurvey than the postsurvey, which accounts for the different between the two lines for each lab level. Currently, there are 70 598 total responses in the dataset.

organize and interact with the data. Sharing jupyter notebooks is in line with calls within the physics and broader scientific communities for sharing code and data [74,75].

Currently, E-CLASS data collection is ongoing. However, due to the COVID-19 pandemic and active data handling procedures, the current dataset is offered as a snapshot for data gathered between 2016 and 2019. Data from 2020 onward are not currently available. Future work may expand the dataset to include additional semesters. Overall, the schema for the E-CLASS will not change substantially. The structure of each table will remain the same. However, it can be the case that extra questions can be added to the survey. For example, in 2020, the postsurvey had several questions added related to laboratory instruction during the COVID-19 pandemic. If these data were to be included in the shared dataset, the schema diagram (Fig. 4) would have additional columns associated with the postsurvey table.

A. E-CLASS: Data collection

The E-CLASS was developed to help instructors and PER researchers measure the impacts of different lab course implementations and interventions. It was developed to address a large variety of learning goals that can be roughly categorized as exploring students' epistemology and expectations of experimental physics. To be able to address such a large range of goals, the survey was not designed to measure just one or a few latent factors. Additionally, the survey was designed to measure students' progression of ideas as they move from introductory courses to more advanced-level courses. To achieve this, many questions are directed at either the introductory or advanced level. Thus, we stress that although one can consider an "overall" E-CLASS score, the real power of the assessment comes from examining responses to individual questions, and in particular, ones that align with a particular course's learning goals [39].

Based on how the survey instrument was designed, there are research questions that can and cannot be answered with the data. For example, while the E-CLASS was designed to assess student attitudes towards laboratory physics, it was not designed to measure latent factors such as student affect (see Appendix). Thus, while exploratory factor analysis can be performed on the E-CLASS dataset, there should be no assumption that any latent factors will be found because the survey was not designed to assess latent factors. This is one particular limitation of the data collected.

E-CLASS data are centrally collected using the Qualtrics survey application and a custom automation system hosted at University of Colorado, Boulder [12]. Central collection of data provides both researchers and instructors with a higher quality of data and supports pedagogical changes [24]. This is due to the connection between local changes and national datasets [12]. Centrally collected data can be better standardized since the data collection process is the same for all data collected. Centrally collected survey repositories have recently become more popular in PER (e.g., PhysPort [26], LASSO [23], PLIC [28]).

There is a standardized process that all instructors follow who wish to use E-CLASS in their courses. Instructors who want to use the survey answer a course information survey (CIS), which automatically generates a pre- and postsurvey for their course. Instructors need to answer the CIS each semester and for each course they plan to use the survey. Answering this CIS automatically generates links for the surveys the instructors can then give to their students for responding to the E-CLASS itself. It also automatically

populates the internal E-CLASS database with the information from the CIS.

The CIS asks the instructors about survey administration logistics, such as when to close the two surveys, and to describe (1) the expected student population (e.g., number of students enrolled), (2) course descriptions, such as the level (intro or BFY), if the course is calculus based, (3) course goals (skills or concepts) and (4) a series of questions regarding the frequency students engage in various activities in the areas of agency, modeling skills, data analysis skills, and communication. A full version of the text of the CIS can be found in the Supplemental Material [76].

The E-CLASS itself is made up of 30 Likert-style questions to assess student epistimologies and expectations in comparison to experts. Student are asked to respond to each statement (from strongly agree to strongly disagree) both from their view and predict the view of experimental physicists. In some cases, the expertlike response is disagree. The data have been preprocessed to convert the Likert responses, so that all data are on the five-point scale of non-expert-like (indicated by 1 in the dataset) to expertlike (indicated by a 5 in the dataset). All of the research to date has been done by first collapsing the five-point scale to a three-point scale, but the full range is included in the public dataset. We warn researchers that the survey was not designed to reliably distinguish between the two outermost points on either end of the scale (i.e., "agree" and "strongly agree" or "disagree" and "strongly disagree"). Additionally, on the postsurvey only, students are asked about which items (23 out of the 30) were important for earning a good grade in the course. Finally, students were also asked a set of demographic, interest, and career plan questions. These questions can be found in the Supplemental Material [76]. We can also use the student responses to quantify response rates to the ECLASS.

The total number of students enrolled in each course is not typically known accurately. Instructors report the expected number of students on the CIS (question 19); however, this number is typically an estimate from the instructor and not the exact number of students enrolled. Therefore, it is difficult to assess the exact response rate of the ECLASS. However, we can determine an upper limit for the matched response rate. Using the total number of unique student responses across both pre- and postsurveys and the total number of matched responses, we can calculate the upper limit of the fraction of students who responded to both surveys:

$$P_{\text{upper limit},i} = \frac{|X_{i,\text{pre}} \cap X_{i,\text{post}}|}{|X_{i,\text{pre}} \cup X_{i,\text{post}}|}, \tag{1}$$

where, $X_{i,\text{pre or post}}$ is the set of students that completed the pre- and postsurvey for course i. This allows us to produce the histogram seen in Fig. 3. This analysis assumes every student had the chance to take both the pre- and postsurvey. In some cases, this is not true due to students dropping courses, entering courses later in the semester than the presurvey administration, etc.

Using the data in Figs. 2 and 3, we can conclude that there are typically more presurvey responses than postsurvey responses and there are considerably more introductory course responses. Overall, the distribution of the fraction of students responding to both the pre- and the postsurvey is similar for both introductory and BFY lab courses, with a slightly higher average response rate for BFY courses.

B. E-CLASS: Data schema

The E-CLASS dataset is organized into three comma-separated tables with an additional metadata table, which describes the questions presented in both the CIS that instructors respond to and the ECLASS itself (Fig. 4). Even though E-CLASS data are relatively straightforward, we have chosen to use this schema to illustrate how one might use relational databases for more complex datasets. For other datasets, a single spreadsheet may be more desirable. (A full schema can be found in the Supplemental Material [76].) The course information table contains data for each course that has offered the ECLASS, including an estimate of the number of enrolled students, level of physics being taught, and institution-level information, such as what is the highest degree granted by the intuition. The presurvey and postsurvey responses are located in their respective tables. Student demographic information is asked during the postsurvey and thus is stored in the postsurvey table.

The pre and postsurvey tables both have some additions and reductions.

- The pre- and postsurvey both have a derived column called "duration." This column uses the Qualtrics

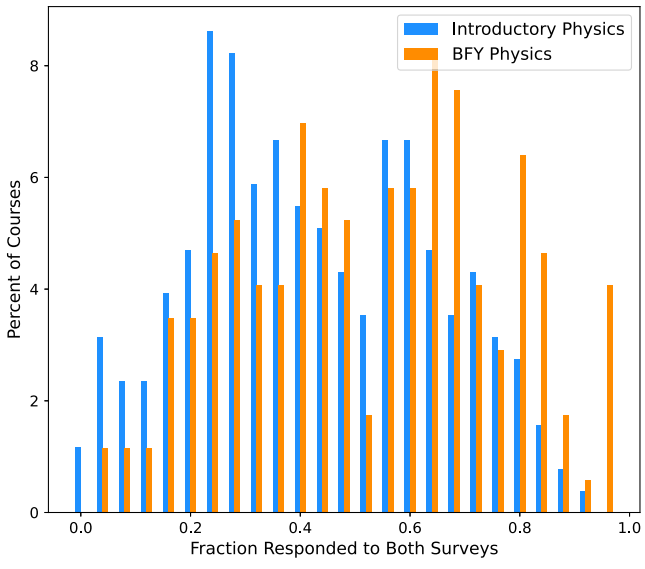

FIG. 3. The distribution of matched responses divided by the total number of unique responses from both pre- and postsurveys. The horizontal axis is calculated according to Eq. (1).

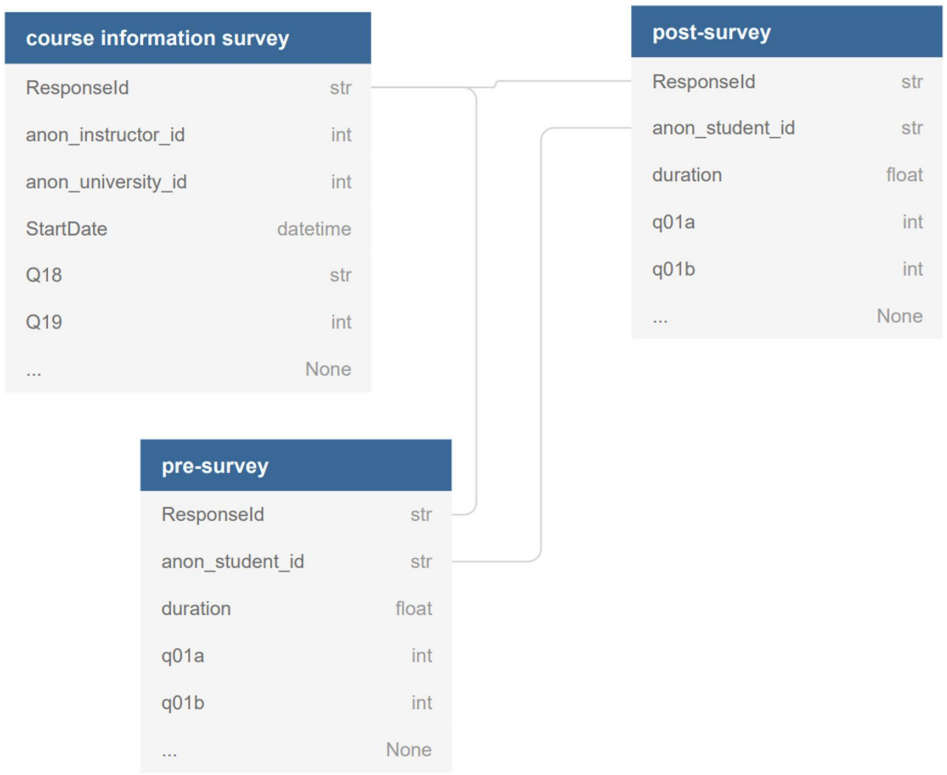

FIG. 4. An abbreviated data schema for the E-CLASS dataset showing a subset of the variable names and type of data (e.g., str for string, int for integer, etc.). The E-CLASS dataset has three tables. The course information table (upper left) contains data for each course that has offered the E-CLASS survey and includes course information such as an estimate of number of enrolled students, level of physics being taught, and institution-level information. The presurvey (lower left), and postsurvey (upper right) responses are located in their respective tables. Student demographic information is asked during the postsurvey and thus is stored in the postsurvey table. A full schema can be found in the Supplemental Material [76].

recorded start and submission times and the pandas TimeDelta function to calculate the number of seconds from when the student opened the survey to when they finished the survey.

- The original gender response included a text box if the student chose to respond "other" from the available options (woman, man, other). As the text box information could contain identifiable information, these data are changed to be NULL values.
- The choices for students' race or ethnicity follows the Department of Education IPEDS definitions of race [77]. Thus, the data are stored in multiple columns since IPEDS recommends letting students select multiple races or ethnicities.

A full description of each column can be found in the metadata table "question_lookup.xlsx" in the github repository [71].

C. E-CLASS: Data anonymization

Anonymization of human subjects data often needs to happen at different levels and E-CLASS data are no exception. E-CLASS data contain information about the specific students who reply, the instructors who teach the courses, the courses themselves, and the institutions that these courses are offered at. In each case, the dataset has been reduced in some way to protect the students, instructors, and institutions represented in the dataset. There have been no student data removed from the E-CLASS dataset for deidentification purposes, except for open response data, student names, and student entered ID numbers. Institutional identifiers have all been replaced with an anonymous institution index. The same is true for instructor and student identifiers.

The data are deidentified in the following ways:

1. All student names are removed.
2. All student IDs are tokenized using an increasing integer renamed to be anon_student_id.
3. All instructors are tokenized using an increasing integer renamed to be anon_instructor_id.
4. All institutions are tokenized using an increasing integer renamed to be anon_university_id.
5. Text responses to all questions have been removed from the dataset. Thus, there are no written descriptions of courses nor are any of the demographic questions that allow for an "other" textbox included.

We also considered if our dataset could have any "outliers" that could be used to identify a student. An example of

an outlier is if we asked students to report their age and had a choice of 80+. Any student who selected that box would likely be the only student in the class in that age range and thus could possibly be identified. We do not ask such a question or any similar low probability questions, and thus did not remove data for being classified as an outlier.

We consulted extensively with the University of Colorado's IRB office for this process to make sure we were creating a dataset that is no longer considered human subjects data and thus can be shared. To be clear, once the data have been deidentified, it is no longer human subjects data, and thus our approved protocol no longer applies.

D. E-CLASS: Data sharing

The E-CLASS dataset is available freely without the need to go through an IRB process, nor to request any permission from the authors of the dataset. The E-CLASS data are available through a github repository [71]. It can be downloaded and used for research purposes. It cannot be used for commercial purposes. To publish a research result using the E-CLASS data, authors need only to cite the dataset itself per the guidelines in the repository and possibly this paper.

The repository includes a DataHelper python module [71] that can help researchers load the dataset and split the dataset for different groups such as introductory students or BFY students. The DataHelper python module is built on the pandas Python Data Analysis Library [78]. ECLASS data users familiar with pandas will find that this helper class responds the same. The DataHelper module is open source. Researchers are encouraged to use whatever tools they are familiar with when investigating E-CLASS research questions.

Although the goal of this paper and data release is to help researchers get access to a large and interesting dataset, this does not imply the authors of this paper are necessarily available for technical or scientific support.

V. DISCUSSIONS AND CONCLUSIONS

We presented a framework that can enable data sharing across quantitative PER research areas. It has used the E-CLASS dataset as an example of how to use the framework. In summary, this paper makes the following recommendations for sharing quantitative data collected in PER:

1. The data collection process should be clearly articulated, including the original research questions and limitations the data may have in identifying new research questions.
2. The data schema should be clearly articulated to understand how multiple tables and other data connect with each other.
3. Data should be deidentified according to local IRB expectations, such that it can be shared as non-human-subjects data.
4. Data should be shared freely and openly without barriers, such as cost to access or other restrictions on whom can apply for and receive data.

Using this model, PER datasets can be made widely available. This sharing of data freely and openly supports validation of results, increases the value of investment in scientific research, and helps to democratizes PER by providing large datasets to all researchers.

In addition to a model for sharing large quantitative datasets in PER, this paper presents the 70 000 response E-CLASS dataset. This dataset has a large response rate across many different kinds of institutions, physics laboratory instruction practices, and curriculum levels. It is our hope that these data can be used to reproduce previous results motivating pedagogical changes in laboratory classrooms. It is also our hope that this dataset will drive new research questions within the context of laboratory pedagogy. Additionally, the size of the dataset can promote advanced statistical examination of the survey instrument itself. Ultimately, we hope that sharing this dataset will help researchers promote changes broadly that make labs more inclusive and effective.

The model presented in this paper is the first step towards a more open data sharing culture in PER. It mirrors that of the statistical evaluation framework presented in Aiken *et al.* [79], which argues for an open and rigorous procedure in assessing statistical models presented in PER. By not only sharing data freely, but also articulating an explicit framework for data sharing, we create a common language in the research community. This common language can be used to reinforce results across different studies, increase the value of the investment of time and grant money on specific research topics, and create a more open and accessible research community for everyone.

The E-CLASS data are available through a github repository [71].

ACKNOWLEDGMENTS

We would like to thank the many people who have worked over the years to help collect the data in the E-CLASS dataset. We would also like to thank Coline Bouchayer who created Fig. 1. This project has received support from the INTPART project of the Research Council of Norway (Grant No. 288125) and the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education (DIKU), which supports the Center for Computing in Science Education. Additional support was provided by the National Science Foundation (PHY-1734006).

APPENDIX: THE FACTORIZATION OF THE E-CLASS

One of the limitations of the E-CLASS dataset is that it was not designed to measure latent variables, such as affect.

TABLE I. Summary statistics describing the number of responses for both introductory labs and BFY labs. "Unique student" is defined as a student's first response to the survey for that semester and course. Unique institutions and unique instructors do not sum up across introductory and BFY courses because the set of instructors teaching introductory and BFY courses intersects.

	Introductory	BFY	Total
Number of unique courses	363	236	599
Number of unique institutions	88	70	133
Number unique students responding to presurvey	30 067	5313	35 380
Number unique students responding to postsurvey	24 465	3817	28 282
Number of unique instructors	123	103	204
Number of matched pre- and postresponses	19 445	3096	22 541

Describing these limitations of the dataset is important in the data-sharing process (Sec. III). However, it is possible that groups of questions may end up measuring latent factors. In this Appendix, we demonstrate, using the dataset provided in this paper, that the E-CLASS does not produce measurable latent factors and that exploratory factor analysis of the E-CLASS is not useful. This is done in two ways: first, using a simple linear correlation between each question, and second, using principal component analysis (also known as exploratory factor analysis).

1. Factor analysis of survey data

The primary goal of factor analysis is to account for the maximum amount of variance explained by the minimum number of factors (in the case of exploratory factor analysis) and a set number of factors (in the case of confirmatory factor analysis) [80]. With survey data, these factors are assumed to be latent variables such as affect, conceptual understanding, or attitude towards a particular topic. A block of questions may align with a particular factor, for example, the first five questions of a survey may measure a student's affect around laboratory science, whereas the next five questions may measure a student's understanding of a particular lab practice. If the number of factors that explain the variance is close to the number of questions in a survey, the survey is said to *not factor*.

2. Factor analysis of the E-CLASS

We assess E-CLASS factorization using two methods. First, we examine the spearman correlations between questions on the presurvey and postsurvey. If there are groups of questions that highly correlate ($\rho > \pm 0.5$), then this might indicate that there are latent variables that the survey is measuring. Second, we use principal component analysis to directly measure whether there are latent factors measured by the E-CLASS. To perform exploratory factor analysis of the E-CLASS, we used the scikit learn principal component analysis package in PYTHON [81]. We do not limit the number of factors since there is no assumption that the E-CLASS is measuring latent variables. We use the

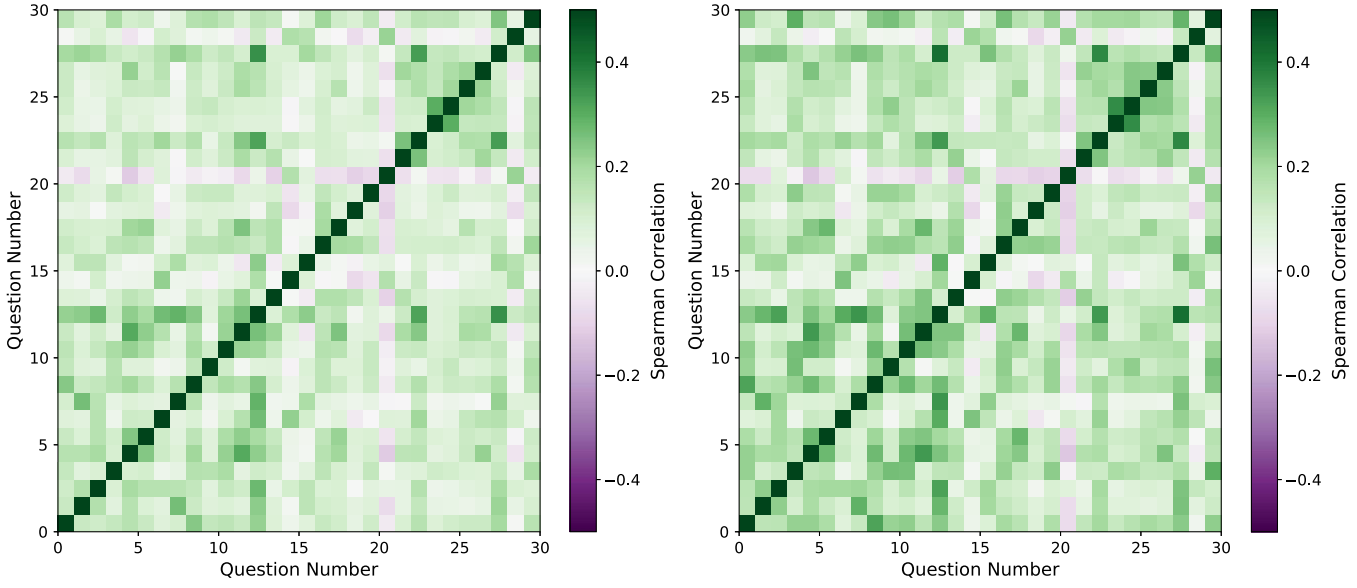

FIG. 5. Spearman correlation coefficients for the E-CLASS survey questions. The left panel shows the presurvey correlations. The right panel shows the postsurvey correlations. Data have not been separated between introductory and BFY students. Instead, they are combined across both groups. If questions measured similar constructs, there would be a high ($\rho > 0.5$) correlation between different questions. This is not the case for either the pre- or postsurveys.

TABLE II. Demographic responses for the ECLASS dataset. Data are reported as matched or unmatched because demographics questions are asked on the postsurvey only. In each case, demographic numbers are reported for unique student responses. "Unique" is defined as the first time a student responds to a survey at a particular level. If the same student has taken both introductory courses and BFY courses they will show up, at most, twice. Once in each category. (Note this is just for demographic counts, all student data still exist for all courses in the dataset.) We include both "matched" and "unmatched" categories. Matched is defined as the student has replied to both pre- and postsurveys. Unmatched means the student has replied to only the postsurvey because the demographic questions are asked in the postsurvey. Additionally, the student's declared major is included in the ECLASS dataset.

	Introductory		BFY	
Gender	Unmatched	Matched	Unmatched	Matched
Female	11 485	8240	1249	899
Male	14 741	10 568	2595	2067
Other	253	197	62	52
Not reported	608	440	100	78
Race or ethnicity				
American Indian or Alaskan Native	67	40	8	5
Asian	5521	4106	782	652
Black or African American	1615	1002	93	64
Hispanic or Latino	1792	1260	176	127
Native Hawaiian or other Pacific Islander	93	57	5	2
White	13 688	9889	2314	1776
Other race or ethnicity	550	379	119	88
Race not reported	1735	1248	279	217
Multirace	2026	1464	230	165
Majors				
Physics	1089	782	1769	1420
Chemistry	786	518	39	24
Biochemistry	937	669	15	5
Biology	3614	2105	252	114
Engineering	7324	5177	808	646
Engineering Physics	196	145	315	250
Astronomy	90	69	8	7
Astrophysics	215	175	88	71
Geology or Geophysics	243	187	7	4
Math or Applied Math	649	494	65	56
Computer Science	1981	1430	35	26
Physiology	519	262	39	9
Other Science	3118	1676	274	119
Nonscience Major	1595	979	40	20
Open option or Undeclared	1975	1301	38	26

TABLE III. Average number of hours students spend working in the lab and outside of lab on lab activities for each type of institution. Additionally, the average number of total lab experiments per term is shown. The standard deviation of these distributions is shown in parentheses.

	2-year	4-year	Master's	Ph.D.
Introductory				
Number of hours working in lab	2.6(0.7)	2.8(0.9)	3.2(1.3)	2.4(0.8)
Number of hours working outside lab	0.23(0.4)	0.42(0.8)	0.90(3.2)	0.06(4.2)
Number of lab experiments	9.5(3.1)	9.5(2.8)	10.4(1.7)	8.9(4.2)
BFY				
Number of hours working in lab	⋯	5.5(7.3)	3.0(0.8)	2.95(2.0)
Number of hours working outside lab	⋯	2.5(2.8)	3.1(7.5)	0.68(1.3)
Number of lab experiments	⋯	6.6(3.3)	7.9(3.4)	6.0(3.8)

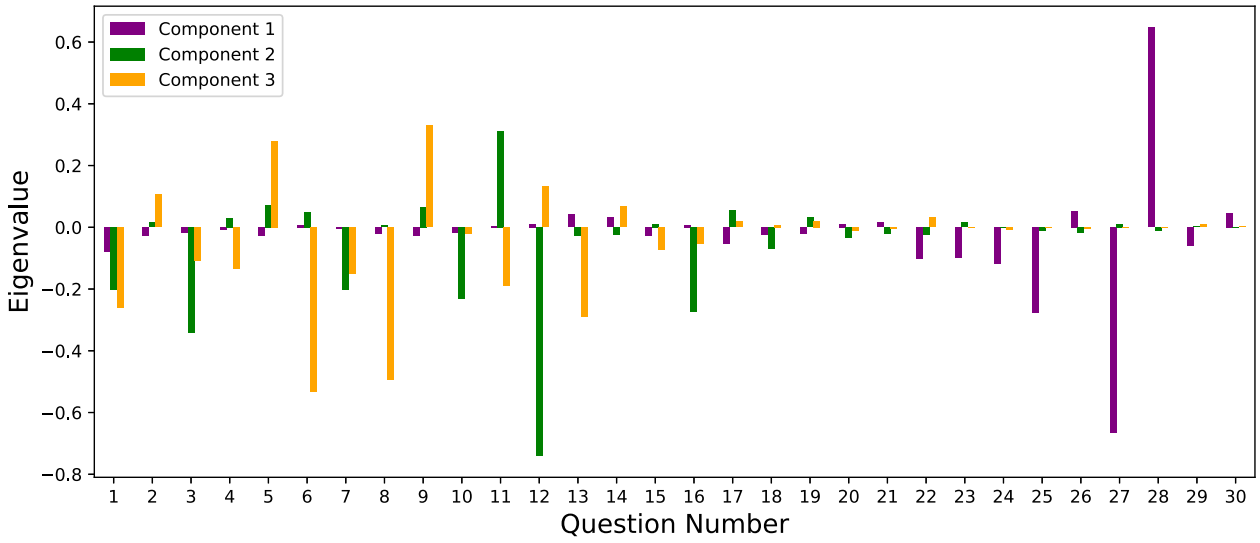

FIG. 6. The eigenvalues of the first three principal components, which explain approximately 30% of the variance. The first component is dominated by a single question, whereas the second two components are controlled by a collection of unrelated questions. Since the goal of PCA is to maximize the amount of variance controlled by the minimum number of variables in the first component and then forward to the next components, this is additional evidence there is no latent structure being evaluated by the E-CLASS.

entire dataset of unique replies provided in this paper, which includes 35 249 student responses for the presurvey and 28 222 for the postsurvey. We do not separate students between introductory and BFY courses.

Using the spearman correlation, we find that there are no questions that correlate on the presurvey or the postsurvey with ρ greater than 0.5 or less than -0.5 (Fig. 5). Therefore, we conclude that there are no direct linear correlations in the E-CLASS questions that would indicate there are latent factors being measured by the E-CLASS.

We also calculate the principal components to assess the factorization of the E-CLASS. As seen in Fig. 7, there is an obvious "elbow," but the total amount of variance accounted for with those three factors is only 32.1% for the presurvey and 34.3% for the postsurvey, and thus the explained variance is not captured by a small (< 5) number of components. In fact, to account for 80% of the variance, one needs to include 15 factors, which is only a reduction of about a factor of 2 from the number of questions on the survey. Additionally, there are no controlling questions for the first three components (Fig. 6). The first component is dominated by two questions (Q27 and Q28). The second two are controlled by a collection of unrelated questions. Therefore, we believe that this presents evidence that there are no latent factors being measured by the E-CLASS.

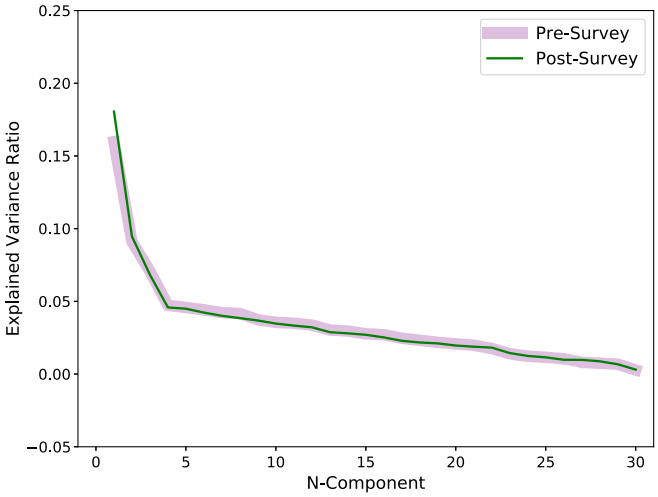

FIG. 7. The explained variance ratio of the question items (also known as a "scree" plot [82]) for both the pre- and postsurveys. The explained variance ratio is the fraction of variance explained per component as opposed to the raw variance given per component. If the E-CLASS measured latent factors, the amount of variance explained prior to the elbow would be much higher. In this plot, we see that the components before the elbow account for about 30% of the variance, and the explained variance does not reach a critical amount (> 0.8) before including at least 15 components.

[1] M. Spiro, Open data policy and data sharing in astroparticle physics: The case for high-energy multi-messenger astronomy, J. Phys. Conf. Ser. **718**, 022016 (2016).

[2] C. Draxl and M. Scheffler, The nomad laboratory: From data sharing to artificial intelligence, JPhys Mater. **2**, 036001 (2019).

[3] U. S. House of Representatives, H.r. 3427/s.1701 fair access to science and technology research act of 2017 (2017).

[4] M. MillsPlan S—what is its meaning for open access journals and for the JACMP?, J. Appl. Clinical Med. Phys. **20**, 4 (2019).

[5] National Science Foundation, Dissemination and sharing of research results—NSF data management plan requirements, https://www.nsf.gov/bfa/dias/policy/dmp.jsp.

[6] American Institute of Physics, https://publishing.aip.org/resources/researchers/open-science/.

[7] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[8] B. Fecher, S. Friesike, and M. Hebing, What drives academic data sharing?, PLoS One **10**, e0118053 (2015).

[9] B. R. Wilcox and H. J. Lewandowski, Based assessment of students' beliefs about experimental physics: When is gender a factor?, Phys. Rev. Phys. Educ. Res. **12**, 020130 (2016).

[10] B. Wilcox and H. J. Lewandowski, Impact of instructional approach on students' epistemologies about experimental physics, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA* (AIP, New York, 2016), pp. 388–391.

[11] B. R. Wilcox and H. J. Lewandowski, Open-ended versus guided laboratory activities: Impact on students' beliefs about experimental physics, Phys. Rev. Phys. Educ. Res. **12**, 020132 (2016).

[12] B. R. Wilcox, B. M. Zwickl, R. D. Hobbs, J. M. Aiken, N. M. Welch, and H. J. Lewandowski, Alternative model for administration and analysis of research-based assessments, Phys. Rev. Phys. Educ. Res. **12**, 010139 (2016).

[13] PLOS ONE, https://journals.plos.org/plosone/s/data-availability.

[14] S. E. Maxwell, M. Y. Lau, and G. S. Howard, Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?, Am. Psychol. **70**, 487 (2015).

[15] J. M. Aiken, R. Henderson, and M. D. Caballero, Modeling student pathways in a physics bachelor's degree program, Phys. Rev. Phys. Educ. Res. **15**, 010128 (2019).

[16] J. M. Aiken and M. D. Caballero, Methods for analyzing pathways through a physics major, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. Jones, L. Ding, and A. Traxler (AIP, New York, 2016), pp. 28–31.

[17] C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame, Data sharing by scientists: practices and perceptions, PLoS One **6**, e21101 (2011).

[18] J. M. Wicherts, D. Borsboom, J. Kats, and D. Molenaar, The poor availability of psychological research data for reanalysis, Am. Psychol. **61**, 726 (2006).

[19] Y. LeCun, The MNIST database of handwritten digits, 10.1109/MSP.2012.2211477 (1998).

[20] S. Kanim and X. C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).

[21] J. Von Korff, B. Archibeque, K. A. Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50 k-student study, Am. J. Phys. **84**, 969 (2016).

[22] G. V. Glass, Primary, secondary, and meta-analysis of research, Educ. Res. **5**, 3 (1976).

[23] J. M. Nissen, M. Jariwala, E. W. Close, and B. Van Dusen, Participation and performance on paper-and computer-based low-stakes assessments, Int. J. STEM Educ. **5**, 21 (2018).

[24] B. Van Dusen, M. Shultz, J. M. Nissen, B. R. Wilcox, N. Holmes, M. Jariwala, E. W. Close, and S. Pollock, Online administration of research-based assessments, arXiv:2008.03373.

[25] L. A. Alliance, Lasso data sharing policy, https://learningassistantalliance.org/modules/lasso/LASSO-FAQ.php#researcherDataAccess.

[26] S. B. McKagan, L. E. Strubbe, L. J. Barbato, B. A. Mason, A. M. Madsen, and E. C. Sayre, Physport use and growth: Supporting physics teaching with research-based resources since 2011, Phys. Teach. **58**, 465 (2020).

[27] PhysPort, Physport data sharing policy, https://www.physport.org/DataExplorer/SecurityFAQ.cfm#2q2.

[28] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking, Phys. Rev. Phys. Educ. Res. **15**, 010135 (2019).

[29] H. Dataverse, https://dataverse.harvard.edu/.

[30] D. W. Bank, https://data.worldbank.org/, retreived=08-04-2021.

[31] U. S. Government, https://www.data.gov/, retreived=08-04-2021.

[32] CERN, https://opendata.cern.ch/, retreived=08-04-2021.

[33] O.S. Framework, https://osf.io/.

[34] O. K. Foundation, https://opendatacommons.org/licenses/odbl/1-0/, retreived=08-04-2021 ().

[35] O. K. Foundation, https://opendata.cern.ch/, retreived=08-04-2021 ().

[36] L. Sweeney, k-anonymity: A model for protecting privacy, Int. J. Uncertainty, Fuzziness Knowledge-Based Systems **10**, 557 (2002).

[37] B. M. Zwickl, T. Hirokawa, N. Finkelstein, and H. J. Lewandowski, Epistemology and expectations survey about experimental physics: Development and initial results, Phys. Rev. ST Phys. Educ. Res. **10**, 010120 (2014).

[38] B. R. Wilcox and H. J. Lewandowski, Students' epistemologies about experimental physics: Validating the Colorado Learning attitudes about Science Survey for experimental physics, Phys. Rev. Phys. Educ. Res. **12**, 010123 (2016).

[39] B. R. Wilcox and H. J. Lewandowski, Developing skills versus reinforcing concepts in physics labs: Insight from a survey of students' beliefs about experimental physics, Phys. Rev. Phys. Educ. Res. **13**, 010108 (2017).

[40] B. R. Wilcox and H. J. Lewandowski, Students' views about the nature of experimental physics, Phys. Rev. Phys. Educ. Res. **13**, 020110 (2017).

[41] B. Wilcox and H. J. Lewandowski, Impact of grading practices on students' beliefs about experimental physics, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH*, edited by D. Jones, L. Ding, and A. Traxler (AIP, New York, 2017), pp. 367–370.

[42] B. Wilcox and H. J. Lewandowski, Correlating students' beliefs about experimental physics with lab course success, in *Proceedings of the 2015 Physics Education Research Conference, College Park, MD*, edited by D. Jones, L. Ding, and A. Traxler (AIP, New York, 2015), pp. 367–370.

[43] N. G. Holmes and H. J. Lewandowski, Investigating the landscape of physics laboratory instruction across North America, Phys. Rev. Phys. Educ. Res. **16**, 020162 (2020).

[44] B. R. Wilcox and H. J. Lewandowski, Improvement or selection? A longitudinal analysis of students' views about experimental physics in their lab courses, Phys. Rev. Phys. Educ. Res. **13**, 023101 (2017).

[45] D. Dounas-Frazer and H. J. Lewandowski, Correlating students' views about experimental physics with their sense of project ownership, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC* (AIP, New York, 2018).

[46] B. Pollard and H. J. Lewandowski, Transforming a large introductory lab course: Impacts on views about experimental physics, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC* (AIP, New York, 2018).

[47] N. Sulaiman, B. Pollard, and H. J. Lewandowski, Impact on students' views of experimental physics from a large introductory physics lab course, in *Proceedings of the 2020 Physics Education Research Conference*, virtual conference (AIP, New York, 2020), pp. 533–538.

[48] L. E. Strubbe, J. Ives, N. G. Holmes, D. Bonn, and N. K. Sumah, Developing student attitudes in the first-year physics laboratory, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. Jones, L. Ding, and A. Traxler (AIP, New York, 2016), pp. 340–343.

[49] C. Walsh, M. M. Stein, R. Tapping, E. M. Smith, and N. G. Holmes, Exploring the effects of omitted variable bias in physics education research, Phys. Rev. Phys. Educ. Res. **17**, 010119 (2021).

[50] D. Hu, B. M. Zwickl, B. R. Wilcox, and H. J. Lewandowski, Qualitative investigation of students' views about experimental physics, Phys. Rev. Phys. Educ. Res. **13**, 020134 (2017).

[51] E. M. Smith, M. M. Stein, C. Walsh, and N. G. Holmes, Direct measurement of the impact of teaching experimentation in physics labs, Phys. Rev. X **10**, 011029 (2020).

[52] K. Funkhouser, W. M. Martinez, R. Henderson, and M. D. Caballero, Design, analysis, tools, and apprenticeship (DATA) lab, Eur. J. Phys. **40**, 065701 (2019).

[53] B. R. Wilcox and H. J. Lewandowski, A summary of research-based assessment of students' beliefs about the nature of experimental physics, Am. J. Phys. **86**, 212 (2018).

[54] A. S. Downey, S. Olson *et al.*, *Sharing Clinical Research Data: Workshop Summary* (National Academies Press, Washington, DC, 2013).

[55] P. Samarati and L. Sweeney, Generalizing data to provide anonymity when disclosing information, Computer Science Laboratory, SRI International, Technical Report No. SRI-CSL-98-03, 1998.

[56] J. M. Aiken, Ph.D. thesis, University of Oslo, 2020.

[57] L. Ding, Theoretical perspectives of quantitative physics education research, Phys. Rev. Phys. Educ. Res. **15**, 020101 (2019).

[58] G. Chung, Toward the relational management of educational measurement data, Teachers College Record **116**, 1 (2014).

[59] R. P. Springuel, M. C. Wittmann, and J. R. Thompson, Reconsidering the encoding of data in physics education research, Phys. Rev. Phys. Educ. Res. **15**, 020103 (2019).

[60] J. M. Aiken, R. De Bin, M. Hjorth-Jensen, and M. D. Caballero, Predicting time to graduation at a large enrollment American University, PLoS One **15**, e0242334 (2020).

[61] A. S. Bryk and S. W. Raudenbush, Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model, in *Multilevel Analysis of Educational Data* (Elsevier, New York, 1989), pp. 159–204.

[62] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker, and M. Warschauer, Mining big data in education: Affordances and challenges, Rev. Res. Educ. **44**, 130 (2020).

[63] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart, Using machine learning to predict physics course outcomes, Phys. Rev. Phys. Educ. Res. **15**, 020120 (2019).

[64] S.-Y. Lin, J. M. Aiken, D. T. Seaton, S. S. Douglas, E. F. Greco, B. D. Thoms, and M. F. Schatz, Exploring physics students' engagement with online instructional videos in an introductory mechanics course, Phys. Rev. Phys. Educ. Res. **13**, 020138 (2017).

[65] R. Solli, J. Aiken, R. Henderson, and M. Caballero, Examining the relationship between student performance and video interactions, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC* (AIP, New York, 2018).

[66] A. Silberschatz, H. F. Korth, S. Sudarshan *et al.*, *Database System Concepts*, Vol. 4 (McGraw-Hill, New York, 1997).

[67] C. Batini, M. Lenzerini, and S. B. Navathe, A comparative analysis of methodologies for database schema integration, ACM Computing Surveys (CSUR) **18**, 323 (1986).

[68] E. F. Codd, Relational database: A practical foundation for productivity, in *Readings in Artificial Intelligence and Databases* (Elsevier, New York, 1989), pp. 60–68.

[69] O. Jaquette and E. E. Parra, Using ipeds for panel analyses: Core concepts, data challenges, and empirical applications, in *Higher Education: Handbook of Theory and Research* (Springer, New York, 2014), pp. 467–533.

[70] R. J. Bayardo and Rakesh Agrawal, Data privacy through optimal k-anonymization, in *21st International Conference on Data Engineering (ICDE'05)* (2005), pp. 217–228, 10.1109/ICDE.2005.42.

[71] https://github.com/Lewandowski-Labs-PER/eclass-public.

[72] U.S. Census, https://www.census.gov/data/developers/data-sets.html.

[73] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay *et al.*, Jupyter notebooks-a publishing format for reproducible computational workflows, in ELPUB (2016), pp. 87–90.

[74] H. Shen, Interactive notebooks: Sharing the code, Nature (London) **515**, 151 (2014).

[75] J. M. Perkel, Why jupyter is data scientists' computational notebook of choice, Nature (London) **563**, 145 (2018).

[76] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.17.020144 for course information survey.

[77] IPEDS, IPEDS definitions, https://nces.ed.gov/ipeds/report-your-data/race-ethnicity-reporting-changes, accessed: 10-23-2018.

[78] W. McKinney, Data structures for statistical computing in Python, in *Proceedings of the 9th Python in Science Conference*, edited by S. van der Walt and J. Millman (2010), pp. 56–61, 10.25080/Majora-92bf1922-00a.

[79] J. M. Aiken, R. D. Bin, H. Lewandowski, and M. D. Caballero, A new framework for evaluating statistical models in physics education research (to be published).

[80] R. L. Gorsuch, Exploratory factor analysis, in *Handbook of Multivariate Experimental Psychology* (Springer, New York, 1988), pp. 231–258.

[81] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. **12**, 2825 (2011), http://jmlr.org/papers/v12/pedregosa11a.html.

[82] A. G. Yong, S. Pearce *et al.*, A beginner's guide to factor analysis: Focusing on exploratory factor analysis, Tutorials Quant. Methods Psychology **9**, 79 (2013).

Correction: Some labels in the previously published Fig. 1 were rendered improperly during the production phase and have been remedied.