Theoretical perspectives of quantitative physics education research

Lin Ding[*]

Department of Teaching and Learning, The Ohio State University, Columbus, Ohio 43210, USA

(Received 21 August 2018; published 3 July 2019)

[This paper is part of the Focused Collection on Quantitative Methods in PER: A Critical Examination.] The notion of quantitative physics education research (PER) so far has been mostly limited to the mere use of statistical methods or use of computational tools for analyzing numerical data. Little attention, in fact, has been given to the underpinnings of this research paradigm. To fill the gap, this theoretical paper addresses key and yet often tacit (or even misunderstood) principles of three commonly used quantitative genres in PER, which I, respectively, refer to as measurement (quantification of individual constructs), controlled exploration of relations (quantification of relationships between multiple constructs through controlled experiments), and data mining (quantification of new information from large datasets). For each genre, I elucidate the paradigmatic basis by focusing on its ontological assumptions (theories about the reality under quantitative investigation), epistemological commitments (views about the knowledge gained through quantitative investigation), and methodological implications for empirical investigations. Although framed in the context of physics education research, the discussions herein are applicable to other discipline-based or general education research employing quantitative methods.

DOI: 10.1103/PhysRevPhysEducRes.15.020101

I. INTRODUCTION

Since the dawn of physics education research (PER), the use of quantitative techniques in empirical studies for data production and interpretation has long been a tradition [1–5]. This tradition largely reflects the historical fact that many earlier researchers in the field were trained as physicists [2,5,6]. Their strong favor toward scholarly objectivity, coupled with their great number skills, has been a major blessing for the popularity of this paradigm within the PER community. As the field continues to grow and mature, other techniques derived from the qualitative and mixed-method paradigms also have rapidly garnered attention [7].

Despite its long standing in education research, the topic of quantitative methods is still frequently misunderstood [3,8]. Discussions about this line of inquiry have almost invariably limited to procedural and technical applications of statistics. Theoretical discourse that explores the deep underpinnings of the quantitative paradigm is rarely the aim of pursuit. On the one hand, empirical PER studies, although regularly invoking quantitative techniques, use them merely as tools; thereby bypassing their theoretical foundations. On the other hand, philosophical debates about research methodologies, which often take place in the broader context of educational and social sciences, have so far exclusively focused on the qualitative paradigm (primarily with an agenda to defend its legitimacy in the field) [9–11]. As a result, there remains a vacuum in the scholarly work, concerning the kind of inquiries that can speak *specifically and directly* to the theoretical foundations of quantitative methods in education research.

Here, by theoretical foundations, I mean the paradigmatic basis of quantitative methods, which in principle relates to issues about ontological assumptions (theories about the reality under investigation) and epistemological commitments (views about the knowledge gained through investigation) [11–13]. In this paper, I target this vacuum by looking into quantitative methods that are common in empirical PER studies. Specifically, I propose a new model of organizing quantitative PER (or any quantitative education research) into three genres of quantification; namely, measurement, controlled exploration of relations, and data mining (see Sec. IV). For each genre, I examine some key points from the ontological, epistemological, and methodological perspectives to highlight the unique features that set them apart.

It is worth emphasizing that the discussions presented in the paper are not meant to be comprehensive. In fact, issues about the nature of reality and the nature of knowledge can never be fully answered. Discussions about quantitative methods along this line are no exception. To that end, my paradigmatic inquiries should be taken as illustrative rather than all inclusive.

[*] ding.65@osu.edu

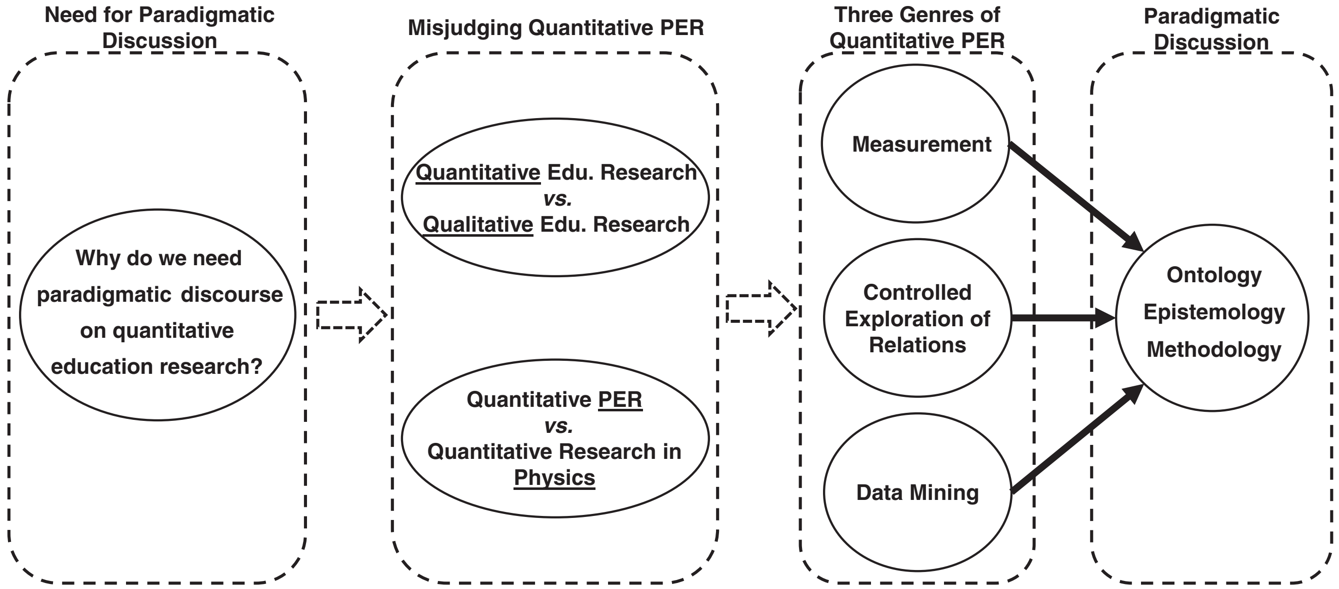
Need for Paradigmatic Discussion
Why do we need paradigmatic discourse on quantitative education research?
Misjudging Quantitative PER
Quantitative Edu. Research vs. Qualitative Edu. Research
Quantitative PER vs. Quantitative Research in Physics
Three Genres of Quantitative PER
Measurement
Controlled Exploration of Relations
Data Mining
Paradigmatic Discussion
Ontology
Epistemology
Methodology

FIG. 1. Overall structure of paradigmatic analysis of quantitative (physics) education research.

II. RESEARCH AIMS

In what follows, I address three related topics that build on each other and lead to the paradigmatic discussions about quantification.

First, the question of why we should be aware of ontological and epistemological issues behind quantitative methods is addressed (Sec. III). This topic offers a justified rationale for the kind of theoretical investigations presented in the current paper. Next, drawing on this basis, I provide a closer look at quantitative methods (or more precisely, quantification practices) in education research by delving into their similarities to and differences from other methods of practices, such as qualitative methods in educational studies and quantitative methods in conventional physics (Sec. IV). As such, some oversimplified views about quantitative PER are refuted. Third, a more detailed discussion about different genres of quantification in PER is presented (Sec. V). In this topic, I classify commonly used quantitative methods in empirical PER into three genres (measurement, controlled exploration of relations, and data mining) and discuss each from the viewpoints of ontological assumptions, epistemological commitments and methodological implications. To provide readers with a roadmap to the key components of the manuscript, I outline in Fig. 1 the overall structure of the subsequent discussion.

III. WHY ONTOLOGICAL AND EPISTEMOLOGICAL ISSUES?

Why should we direct our attention to the underlying ontological and epistemological issues when it comes to discussions about quantitative methods? (Here, I refer to ontological and epistemological issues as those related to theories about the reality under investigation and theories about knowledge gained through investigation, respectively.) This is a foundational question for the thesis of the paper, which I address from the following perspectives.

First, it is important to note that discussions about quantitative methods in PER (or in any educational research) should not be restricted to considerations of methods as tools *per se*. Instead, we need to pay attention to the normative practices of how quantitative methods should be used and, more importantly, why they are used as such [3,8]. In other words, we need to think of quantitative methods as a set of quantification practices rather than just a set of technical algorithms. As Erickson and Wolcott [14,15] pointed out, methods in the sense of mere tools not only are the most uninteresting part of inquiries but also can muddle our understandings of the relationship between methods and research goals. Schwandt [16] echoed this sentiment and noted that what distinguishes various approaches to educational studies cannot be explained by the different methods. He continued to argue that the unique activities in different methods might be best defined by the purposes of investigators who choose to use them, and such purposes in fact are governed by one's ontological and epistemological stances [16]. Although Schwandt made this claim in the context of advocating a philosophical analysis of qualitative methods, his argument in essence is general and applicable to both qualitative and quantitative methods.

Second, an ultimate goal of studying quantitative methods is to push the field forward. If the focus of methodological inquiries remains only at the level of technical and procedural operations, our abilities to advance research practices are bound to be limited. As Kelly [17] argued in his proposal for a normative epistemology for educational research, a theory that taps into the ontological and epistemological challenges for generation of knowledge can improve research, strengthen scholarly communities, and reduce oppressions on the social science research community from the proponents of natural sciences. He further emphasized that "to the extent that epistemology has anything to say *a priori* about research methods, it ought to provide a prescription for the nature of debate, rather than to define validity for various methods of research." [17].

Third, from the practical standpoint, understanding key ontological and epistemological issues can help us better appreciate and more properly use different quantitative methods. As seen in Sec. IV, quantitative methods can be easily mischaracterized as equivalent to statistical procedures or as completely incommensurable with qualitative methods. These positions often lead to an unfair judgement of quantitative methods and distort their very nature. To a large extent, such mischaracterizations of quantitative methods can be attributed to the fact that the paradigmatic considerations of underlying ontological and epistemological issues are overlooked.

In addition, the abundance of paradigmatic debates on qualitative methods has already set up an excellent precedent for us, a precedent in which explorations of ontological and epistemological issues have contributed significantly to the rising status of qualitative research in today's educational and social sciences [7,11]. No doubt, similar conversations are equally necessary for quantitative methods. At the very least, such conversations can push the current practices of quantitative methods out of the purely technical state of just using statistical or computational algorithms. On that note, it is worth repeating that the fruitful discourse about paradigmatic foundations that abounds in qualitative education research has not easily spread to quantitative education research (largely due to its dissimilarity from the former but resemblance to natural sciences in terms of the use of quantitative data). Therefore, the latter requires an independent line of inquiries to unpack its unique, tacit, and often misunderstood theoretical underpinnings. To this end, theorizing around quantitative education research is both novel and critical.

IV. MISJUDGING QUANTITATIVE METHODS IN (PHYSICS) EDUCATION RESEARCH

Quantitative methods are frequently used in empirical PER studies. Ironically, they are also frequently misjudged for what they really are and what they can do [3]. On the one extreme, they can be demonized as representative of positivism or postpositivism that follows rigid doctrines to pursue a sanitized, absolute reality, a reality free of any personal values, beliefs, and theories. On the other extreme they can be idolized as *the* scientific way of doing educational studies if technical and procedural requirements are followed [18]. Although these are quite radical views, they do touch upon deep ontological and epistemological issues about quantitative methods.

In response to these viewpoints, I discuss quantitative methods as a whole by highlighting how they are similar to qualitative methods in PER and how they differ from quantitative methods in physics research. By doing so, I hope to sketch out a clearer picture of what is being investigated by quantitative methods and what is acquired through such investigation.

A. Quantitative methods in relation to qualitative methods in (physics) education research

One similarity shared by quantitative and qualitative methods in PER is what is being investigated; that is, abstract, socially constructed, and often elusive entities in the teaching and learning process [3,17]. These include, for example, learners' conceptual understanding, reasoning ability, self-identity, attitudes, and so forth. They are not directly observable but instead manifest themselves through performances of human subjects on certain tasks. In Schutz's [19] words, the reality under investigation in human sciences is a second-order construction, or "construction of construction." Here, Schutz considers first-order construction as the process by which humans make sense of the world (in PER, for example, it can be the process by which learners interpret the subject matter being learned or students make sense of the learning environment in which they are being placed) [16,19]. The second-order construction is the process by which we, the educational researchers, make sense of the first-order construction. Conceivably, investigators using either quantitative or qualitative methods are facing quite similar challenges in grasping how human beings construct meaning of the world [16,19].

It is in this perspective that giving a broad-brush label of positivism or post-positivism to quantitative research practices is inappropriate. [13] In fact, rarely in modern times do we see educational investigators practicing quantitative methods claim to pursue an absolute reality of teaching and learning, and rarely are their findings claimed to be unequivocally objective or free of personal values. This of course is not to say that quantitative and qualitative practices are no different in their ontological or epistemological stances. It is, however, important to note that practices of quantitative methods should not be crudely regarded as any less of constructivism than qualitative practices.

Similarly, quantitative and qualitative methods should not be viewed as a clear-cut dichotomy. For that, the newly emerged mixed-methods approach can be strong evidence [20]. As Campbell [21] contended, "there is no quantitative knowing without qualitative knowing." [22]. From a different angle, "whatever exists at all exists in some amount," said Thorndike [23]. In other words, because of the subtle connections between quantitative and qualitative practices, these two are destined to share, more or less, some commonalities [22,24].

B. Quantitative methods in physics education research versus quantitative methods in physics

Using quantitative methods to interpret teaching and learning phenomena inevitably involves dealing with numbers (quantification). Owing to the standardized number systems in mathematics and statistics, quantitative practices in PER can become reminiscent of traditional physics research. In this sense, quantitative PER may be mistakenly viewed as holding a higher degree of scientific objectivity than qualitative PER.

But is it true that quantitative PER is comparable to quantitative physics research? Clearly, the answer is no. In fact, they differ significantly at least in the following three aspects.

The first difference lies in what is being studied. As mentioned before, the subjects of interest in PER require second-order constructions [19], which pertain to human agents, or more specifically their cognitive and affective responses to teaching and learning activities [3]. These topics are ill-defined, abstract, and even elusive, and the success of investigation lies at the mercy of placing the right human agents in the right settings. On the contrary, what is being studied in conventional physics remains at the level of first-order construction, relating directly to physical matters and their interactions. Because of this difference, PER investigators do not have the luxury that physicists enjoy of finding and manipulating identical agents for lab studies.

Another difference between the two fields is the process by which numerical information is produced. In traditional physics, there are standardized instruments available for quantitative measurement, and hence collecting and making of numerical readings seldom becomes controversial. This is true even for instruments created under different unit systems (e.g., SI or imperial systems), in which case measurement results still can be converted from one to another with relative ease and accuracy. In sharp contrast, quantitative measurement in PER lacks so-called “standardized” instruments [3]. This is because the creation of such instruments alone requires a second-order social construction [19], which is not only theory laden but also shaped by designers’ personal perspectives. This is why it is nearly impossible to have a globally agreed-upon instrument in PER or in any education research.

Here, one might disagree by bringing up the Force Concept Inventory (FCI) [25] as a counterexample and argue that it is possible to have a “standardized” instrument in PER. It is true that the FCI is broadly used to measure learners’ conceptual understandings of Newtonian concepts. However, this does not mitigate the fact that many researchers find it either problematic or inappropriate for certain student groups [26–31], or choose not to use it simply because they want to use their own self-developed instruments. If the latter, conversion of results still remains as an issue. In that sense, the ways quantification is practiced in PER are inherently different from those in traditional physics research.

In addition, the types of numerical information acquired from PER are different from those in physics. In conventional physics research, it is both feasible and natural for human beings to conceive the material world as comprised of continuous constructs (human-constructed concepts that can be placed along a continuum), for example length, time, and temperature [32]. On the other hand, when it comes to the objects of investigation in PER, it is rare for human cognitive and affective experiences (which are the subjects of interest in PER) to take on a numerical form spontaneously [33] (Chap. 9). Just imagine how difficult it is to think of, for example, someone’s “knowledge” in terms of a number value. This requires investigators to impose a numerical structure on their subject matters; but even so, they still face a challenge of treating their data as continuous. Strictly speaking, the kinds of numerical information generated in quantitative PER are nearly always noncontinuous [3]. Think of the quantitative data (student scores) collected from the FCI; for example, the numerical values in fact are discrete in nature. It therefore behooves the investigators to be mindful of this unique feature and properly justify their models of knowledge when using certain statistical techniques for data interpretation.

V. THREE GENRES OF QUANTIFICATION

The above discussion of quantitative methods as a whole provides only a glance of what this paradigm is. Because actual practices of quantification exist in many different forms, it is useful to examine them separately. To do so, I classify the quantitative methods that are commonly used for empirical PER studies into three categories, which I call, respectively, measurement, controlled exploration of relations, and data mining. Each category represents a genre, with its own unique ontological assumptions, epistemological commitments, and methodological implications.

It is worth noting that the way quantitative methods are categorized herein is not the only way. It is also important to note that the three genres presented below are not meant to cover the entire gamut of the quantitative paradigm. The reason the categorization is so carried out is to ensure that the research aims of highlighting different ontological and epistemological underpinnings are fulfilled while at the same time to keep the task manageable. Below I discuss the three genres in detail.

A. Measurement

One major genre of quantitative methods in PER is to generate numerical values for a certain construct of interest; for example, creating scores to indicate students’ understandings of a physics topic or assigning values to represent instructors’ pedagogical content knowledge.

For convenience, I call this genre measurement. Bear in mind that here the term of measurement is only referred to as quantifying individual constructs. As to quantification of relationships between different constructs, I defer it to Sec. V B.

1. Ontological assumptions

One fundamental question for measurement in PER concerns the nature of what is being measured. [34] This includes key assumptions that investigators need to make about common models of "reality" in order to achieve their measurement goals.

As discussed before, from a general perspective the subject of interest in PER measurement primarily concerns human cognition or affective experiences, which by nature are private to the human subjects themselves and hence are not directly observable [3]. However, thanks to experimental psychology, those internal, unobservable constructs can be externalized and made expressive through human performances on certain tasks [33]. In other words, one ontological assumption for quantitative PER measurement (or any other educational or psychological measurement) is the causal relation between internal cognition or affect and external human performance, with the former being a cause and the latter being the effect [33].

Another important but often tacit ontological assumption is that human subjects are composed of discrete and isolatable attributes, which only vary in degree (amount) from person to person [33,35]. Here, the notion of human attributes being discrete and isolatable, in theory, allows investigators to separate and study different attributes individually. For example, in order to measure students' conceptual understandings of a certain physics concept (an attribute under investigation), a researcher wishes to presume that other irrelevant attributes can be separable from and therefore are not part of the measurement results. As for the notion that human attributes only vary in amount from person to person, it essentially assumes that what is being measured does not change its identity across different human subjects but only changes its amount [33,35]. This, in theory, makes it possible to quantify (impose a numerical structure to) the constructs of interest and also makes between-person comparisons legitimate.

In addition, because quantitative measurement is a result of human-task interactions (in which human subjects receive prompts from certain tasks and then respond to them), it is assumed that the two aspects interact only in an additive manner [33]. This means that if the focus of investigation is human, then the tasks simply become stimuli to bring about human responses without themselves going through scrutiny. On the other hand, if the focus of investigation is the tasks, then human subjects retreat to the background and become irrelevant features. Simply put, the assumption of additive human-task relation allows researchers to study either of the two aspects separately of the other.

2. Epistemological commitments

Just as ontological assumptions can reveal important features of the modeled reality in measurement, epistemological discussions can tell us about characteristics of the knowledge gained through the measurement process.

One epistemological standpoint that becomes increasingly evident in the modern measurement literature is the recognition of and commitment to social construction. In his review of the history of cognitive psychology, Danziger [33] recounted how psychologists in the early 1920s invented and quantified various new categories of personality (such as "ascendance" and "submission") as a response to the then social needs of selecting suitable military personnel. Fast forwarding to our modern days, the pressing administrative needs for capturing the overall students' learning outcomes and holding schools accountable for student learning have promoted all sorts of educational measurements that target a variety of constructs of interest [35]. Take the recent new science standards [36,37], for example, it has created great enthusiasm and opportunities for researchers to seek measurement of students' core disciplinary ideas, crosscutting concepts, and scientific practices.

Another central epistemological issue concerns the nature of measurement results. On this matter, there are two major schools of thought in the contemporary field of educational measurement. One follows the classical test theory (CTT) to seek measurement that can reflect the "true" value of what is being investigated. Specifically, investigators in this camp conceptualize measurement results as a linear combination of a true value and error terms and hence direct their effort toward teasing apart the true value from errors [34,38]. Here, for the lack of a better term, the word true is mostly a legacy of conventional statistics, which is used to denote an assumed value to be sought and does not necessarily suggest realism on the part of CTT practitioners. In contrast, the other school of thought, represented by Rasch theory (RT), draws on the stochastic viewpoint and conceptualizes measurement results as a probabilistic outcome from the interactions between human subjects and tasks [3,34,39,40]. Because of the shift from the so-called true value to a probability-based value, investigators following RT claim their results are more generalizable than CTT results.

Regardless of the difference, both camps appear to acknowledge the tentative and error-prone nature of their measurement results. For one, this can be witnessed from how researchers' perspective of measurement validity evolved. As Kane [41] and Liu [34] each pointed out in their respective reviews of this topic, scholars have gone a long way evolving from pursuing algorithmlike definitive indicators of validity to nowadays recognizing validity being more of an argument than a property of measurement instruments or results. In other words, researchers come to realize that what they know from measurement is indeed subject to interpretation. What's also evident for both camps is that measurement results are context dependent [34]. This is why for both CTT and RT researchers sampling representative groups of human subjects and tasks is crucial. Given that it is impossible to include every human subject or every scenario into actual measurement, empirical results are never claimed to be definitive by investigators.

3. Methodological implications

From the above ontological and epistemological discussions, several methodological implications can be extracted.

One important implication directly relates to the design of measurement instruments. As pointed out earlier, constructs to be measured in quantitative PER are almost always unobservable. Therefore, one major challenge here is how to make the unobservables observable. This requires that investigators not only conceptually define what the construct of interest is but also operationally define how it can be carried out through observing human performance. Here, the operational definition is a key, as it directly guides the design of tasks (also known as questions or items in measurement instruments). Take the study by Ding and colleagues [42] for example, the investigators went great lengths to delineate their construct of interest, namely, learners' conceptual understandings of energy topics in introductory physics, and operationalized it as learners' flexible application of the energy principle (and its associated subtopics) in different contexts. Drawing on this definition, the investigators designed a set of isomorphic concept items to link the otherwise unobservable construct to the observable student performances on these questions.

Also critical here is the practice of imposing a numerical structure to the observed human performance. It is important for the numerical structure to be so designed that values assigned to the human performance can reflect the degree of what is being measured in a consistent way across different human subjects. Ideally, quantitative results from a measurement ought to be free of bias on any ground that is irrelevant to what is being measured. Only under this condition can subsequent analysis such as between-person comparisons become meaningful. Efforts along this line of work, although scarce in the past, have now become increasingly visible. Researchers including Ding [43], Henderson *et al.* [44], and Traxler *et al.* [45] have reexamined some popular PER assessment instruments to look for differential item functioning (a terminology coined to indicate inconsistent functioning of items) that might have given rise to the potentially biased numerical structures for students from different classes or of different genders.

In principle, development of PER measurement instruments requires a reasoned framework, and such a framework at the very least needs to include the following aspects. (a) Justification for why tasks so crafted can bring about observable human performance, (b) justification for why the observed human performance can allow investigators to make legitimate inferences about the target construct, and (c) justification for why the numerical structure imposed on the observed human performance can consistently reflect the degree (amount) of the unobserved construct. Note here that it is the inferred construct, not the observed performance that is being pursued and quantified. To that end, justified articulation of the definitions for the construct, both conceptual and operational, could not be overemphasized.

Another important implication derived from the above discussion regards how we deal with validity. As discussed before, validity should no longer be viewed as a fixed property of measurement instruments or measurement results. [41] Instead, it should be viewed as the legitimacy of arguments for the inferences that are drawn from the measurement results. In that respect, Lindell and Ding [46] illustrated various possible situations, in which the utility of assessment instruments could potentially be compromised if validity were taken as an inherent feature of the assessments. They further called for revalidation of instruments by engaging in systematic considerations of the specific contexts of any given study. In this epistemological perspective, it is imperative that investigators go beyond mere statistical calculations and enter into evidence-based argumentation to justify the link between data and inference. This, according to Liu [34], requires investigators to consider not only supporting evidence but also disconfirming evidence. As such, both confirmatory and falsification processes can be carried out, thereby strengthening the argument for validity.

One more methodological implication that can be extrapolated from the above discussion relates to the practical use of CCT or RT. While a full analysis of the pros and cons of each theory is beyond the scope of the paper, it is crucial to note that the two theories represent inherently different viewpoints about measurement results. Therefore, mixing the two approaches is not recommended and in fact is considered unacceptable [40]. In addition, there is no rigid rule governing which theory to use under what specific situation. Since both theories have advantages as well as drawbacks, it behooves investigators to justify their choice of theory.

B. Controlled exploration of relations

A second genre of quantification discussed below is to explore the relationships among different constructs through controlled studies. It differs from the above genre in that it builds on the measurement of individual constructs to examine the relationships thereof. It also differs from the third genre introduced in Sec. V C in that here relationships are examined through controlled (or quasi-controlled) studies.

1. Ontological assumptions

As with the previous case, the genre of controlled exploration of construct relationships assumes some unique positions about what is being investigated.

First, it is assumed that individually measured constructs relate to each other in some meaningful manner, be it a correlation, nonlinear association or causation. In PER, for example, quantitative studies are often designed to examine the relations between students' attributes and learning environments in which the students are placed. Here, a tacit and yet often misunderstood assumption is that the environments are composed of independent features that can be isolated and manipulated by investigators, and that student attributes are a function of the environmental features. [33,35] Thereby, a change in the environment, if systematically varied, can lead to a change in the students' attributes. In theory, this assumption provides a logical basis to make exploration of construct relationships possible through controlled studies.

Another important assumption is that the ways different constructs relate to each other do not readily manifest themselves under natural conditions. Rather, they are easily dispersed or hidden if left in an uncontrolled environment and hence require investigators to provide specific interventions as a means to separate and manipulate target constructs. [47] By doing so, the relationships between key constructs can be teased out with less interference from other environmental features or human attributes.

A third point, which originates from social statistics and yet is crucial for this type of work, concerns the quantifiability nature of the relationships among constructs. It is postulated that human conduct (including observable behaviors as well as unobservable attributes) follows, in Danziger's [33] words, "quantitative scientific laws," and that the lawfulness of human conduct becomes visible *if* individual observations are aggregated. Here, the putative existence of quantitative patterns in human conduct justifies this genre of practice whose aim is to extrapolate quantitative relationships between various human-related constructs. More importantly, the quantitative patterns are postulated to be a result of aggregation. This, in some sense, is similar to the idea of ensemble averages in thermodynamics, where an average of different microstates can lead to new information about the macrostate of a system.

2. Epistemological commitments

To better appreciate this genre, it is crucial that we discuss some key epistemological issues regarding the nature of the knowledge derived from this type of work.

First, just as measurement of individual constructs (as discussed in Sec. V A) is shaped by social values, the results that investigators get from controlled studies of relationships are also theory laden and influenced by the investigators' perspectives [48]. In empirical PER (or education research in general), selecting target relationships for study involves human decisions about, for example, what constructs to choose or eliminate and what relations to pursue or ignore. Even for a seemingly simplistic case of comparing different student groups exposed to different instructional methods, investigators must consider *a priori* how to conceptualize the relationship between instructional methods and student learning outcomes.

That said, researchers working in this genre do strive for a certain level of objectivity (particularly in the sense of replicability) and seek not only descriptive patterns but also causal relations. As Cook and Sinha [48] mentioned, even if controlled quantitative experiments in educational research cannot provide so-called facts, they still can offer propositions that can be confidently accepted as correct. The key point here is not so much about making a claim of knowledge as a synonym for reality, but instead it is the pattern that "stubbornly reoccurs across multiple researcher predilections" [48] that captures investigators' attention. Similarly, as Danziger [33] pointed out, it was "the repeated demonstration of striking regularities in social statistics that first convinced a large public that human conduct was subject to quantitative scientific laws." Indeed, these "striking regularities" are the kinds of knowledge claims that researchers in this area hope to make.

Another important epistemological issue for this genre concerns the aggregate nature of the study results. In controlled quantitative experiments, the "lawfulness" of human conduct often emerges in aggregate and may or may not appear in observations of individual human subjects [33]. In other words, the aggregate results that this genre of work produces refer to an "abstract, statistically constructed individual" and do not *need* to correspond to *any* of the actual human subjects who comprise such a "collective individual." As Cobb [35] pointed out, the epistemological focus here is "to distinguish between the abstract, collective individual to whom knowledge claims refer and the individual students who participate in experiments." In fact, the construction of this abstract "aggregate student" or "average student" [33] allows investigators to avoid the challenges of dealing with individual students.

3. Methodological implications

Based on the above discussions, it is clear that quantification practices in this genre hinge on the careful design of experiments to draw out the relationships among different constructs. Since literature on experimental design abounds, I leave out technical details on this matter. Instead, I highlight the role of conceptual frameworks that need be created to describe and explain the target relationships for empirical studies. Here, the frameworks are referred to as models (however preliminary) that investigators build even before data collection to represent the relationships being studied. It is important to note that the initially built frameworks need not be complete or fully refined; in fact, it is impractical to hold such an expectation. However, investigators do need to strive for a framework that has the following features. First, it conceptually delineates key relations between target constructs. This can clarify what is being investigated and help investigators connect research questions to practical plans of action. Second, the framework justifies why the relationships are conceptualized as such (for example, evidence from prior studies, support from learning theories, or perhaps both). This can provide reasoned evidence for the construction of the framework and can also serve as an opportunity for investigators to make modifications to the framework if necessary. Consider, for instance, Ding's [49] work on investigating the causal relationships among three constructs (learners' conceptual understanding of Newtonian force topics, scientific reasoning skills, and learning attitudes toward physics). The investigator synthesized a whole host of literature on learning science and used it as bedrock to design a conceptual framework that not only outlined the links among the three constructs but also explained the rationale thereof. As a result, quantitative testing of this framework in the subsequent analysis became meaningful.

Another implication worth noting here concerns how to properly practice and fairly judge this genre of quantification. As previously pointed out, the knowledge claims acquired from this tradition often refer to an abstract, statistically aggregated person, whose attributes need not reflect the individual persons participating in the studies. This, in principle, is analogous to what is known to physicists as micro- and macrostates in thermodynamics, where a macrostate is an average result of many microstates, and the macrostate does not need to correspond to any specific microstate. In educational research, it is important for investigators to note that their results may not be mapped onto any individual student. In fact, such a mapping is not only unrequired for this genre of work but also unexpected. Therefore, when it comes to making knowledge claims, investigators need be mindful not to overstretch inferences from data or make claims too specific to any student. On the other hand, when it comes to judging the merits of this type of work (for example, peer review of research manuscripts), it is important to remember its ontological and epistemological boundaries and hence make appropriate judgment by the norms of this research tradition. As Cobb [35] argued, "[j]udgments of theoretical depth do not transcend research traditions but are instead conceptually relative to the norms and values of particular research communities."

C. Data mining

Of the three genres of quantification, data mining is the newest line of work; emerging nearly as an independent field known as educational data mining [50]. As with the previous case, data mining can be used to explore quantitative relationships between different constructs. However, it differs from the previous in that investigators in this tradition often have little control of data collection. In fact, the data being examined often have been previously collected; hence the types of research questions that can be addressed may be limited [50,51]. That said, this genre has its unique advantages. The most striking of them perhaps is its access to and utilization of large sets of data (such as regional or national repositories) which few controlled experiments in local settings can produce. Because of this feature, this genre has its own ontological and epistemological emphases.

1. Ontological assumptions

As with before, data mining assumes an additive nature of human attributes and environmental features (meaning human subjects and environments presumably are composed of separable attributes or features) [33]. It is also assumed that human conduct in aggregate follows quantitative patterns. However, differing from the previous case, investigators in data mining have little or no control of data collection. To that end, the ontological emphasis now is being placed heavily on the quantity of data. In other words, a large compilation of information becomes a necessary condition for data miners to identify the so-called reality about human patterns [52,53]. In some sense, the large scale of information required for data mining is to make up for the lack of control in the collection of such data information.

Another important ontological assumption for data mining is that once constructs of interest (such as human attributes) are empirically defined, they can be transferred across different datasets (representing different human subjects and under different conditions), although such transfer may not always be trivial [54]. This in essence makes it possible to combine data collected from different contexts for aggregate investigations. This assumption is rooted in two fundamental notions; one from experimental psychology that the identities of human attributes remain more or less stable but the amounts thereof can vary, and the other notion from social statistics that individual deviations cancel out in large data sets [33].

2. Epistemological commitments

Although investigators of this genre may have the least opportunity to inject their personal values into the data during its production, they have control of how it is to be examined. The fact that researchers act upon previously collected information does not preclude the need for conceptual models or frameworks that precede any quantitative analysis. In fact, the models and frameworks that investigators bring with them to data mining are their interpretative constructions of the data and of the research questions, which are inevitably influenced by their personal values. These personal values, in turn, can strongly shape the ways the data are structured, evaluated, and interpreted. From this perspective, the seemingly objective practices of data mining are not free of theories or personal views. Fortunately, this important epistemological feature has been recognized and acknowledged by practitioners working in this tradition [53,54].

On the other hand, just as with the previous case, investigators in this tradition do seek broader generalizability in their knowledge claims (in an aggregate sense). However, differing from the above case in which relationships are studied in controlled (or quasicontrolled) environments, data mining places less emphasis on experimental design but more on the fit between large-scale data and computational models. [55]. Here, computational models (see, for example, Refs. [56,57]) are referred to as statistical or computational algorithms used for performing data analysis (which are different from the term "model" used elsewhere in the manuscript). With the help from these computational models as well as large-scale data sets, data miners take pride in the replicability of their results by pointing to the persistent patterns in aggregate information of human conduct. Nonetheless, they also come to an epistemological realization that such replicability mostly is about descriptive regularities in the data and is not so much about underlying causality.

3. Methodological implications

From the above ontological and epistemological discussions, it is clear that data mining is not just robotic applications of computational or statistical algorithms. Instead, data miners need to build, before mining the data, conceptual frameworks that can delineate and justify the relationships in question. Typically, a good conceptual framework needs to have the following functions. (a) Define (both conceptually and operationally) what constructs and relationships are under investigation; (b) justify why the relationships are defined as such (for example, based on relevant literature, prior empirical studies or learning theories), and (c) if needed, consider alternative relationships and explain how the alternative relationships are being considered or dismissed.

Also important and unique to this genre of work is that investigators should use caution when combining large-scale data sets collected from different contexts. Although theoretically defined constructs can be transferred across different sets of data, the actual systems or contexts in which the datasets are collected may vary from case to case. For instance, information generated in a local classroom environment may or may not be well aligned with the information at the institutional or administrative level. This largely stems from the lack of consistency in certain difficult-to-define data variables, including student demographics and course information. Consider a case of examining learners' course grades. What is being marked as a passing grade in one class may likely be labeled as a failing grade at the institutional level or at a different administrative level [58]. Similarly, data collected from various institutions or regions by using different standards or criteria (for example, what determines gender and race, or what types of teaching qualify to be interactive instruction) can all potentially become problematic when they are combined for data mining [58]. This, therefore, can lead to what Cope and Kalantzis [51] called an issue of lacking interoperability or commensurability. In such cases, investigators should at the very least lay out arguments for their decisions, if the datasets to be combined are not intuitively interoperable or commensurable.

VI. DISCUSSION

The three genres of quantification practices as summarized in Table I, although not comprehensive, capture the main traditions of quantitative physics education research or other discipline-based educational research. Specifically, they represent three different types of approaches to producing and interpreting numerical information about human-related activities in teaching and learning. Although they are discussed separately in the current paper to highlight each of their unique ontological and epistemological underpinnings, by no means should we consider that they could only be used independently of each other [13]. In fact, investigators frequently cross the boundaries of these genres to engage in multiple practices of empirical analysis. For instance, scholars in PER often use research-based assessment instruments to measure student conceptual understandings of certain physics concepts, and use these measurement results in experimental or quasi-experimental settings to evaluate the effectiveness of reformed curricula, instructional materials or classroom activities (see, for example, Refs. [30,31,59]). In these cases, measurement provides the essential threads of information (such as student learning outcomes and curricular or classroom characteristics), while controlled (or quasicontrolled) investigations weave these threads into a logical and defensible web of relational systems (i.e., how student learning outcomes relate to curricular or classroom characteristics). It is worth noting here that the measurement and (quasi-)controlled investigations in the above example should not be thought of as occurring in a sequential order. Instead, these two aspects are closely intertwined in actual studies, and the design, implementation, or analysis of one aspect cannot happen without taking the other into consideration.

It is also important to note that the three genres discussed herein should not be taken as three discrete types of practices. In fact, they can be conceptualized as locating along a continuum of various shades of quantification practices. For instance, the boundary between controlled investigations of quantitative relationships and data mining may not be so clearly defined. For the vast majority of PER studies, randomized experiments are rare due to logistical

TABLE I. Paradigmatic summaries of the three quantification genres.

	Measurement	Controlled study of relations	Data mining
Ontological assumptions	Cause-effect relations between internal human attributes and external human conducts;	Human attributes as a function of environments;	Humans and environments composed of isolatable and manipulable attributes and features;
	Discrete and isolatable human attributes, varying in degree but invariant in identity;	Environments composed of independent and manipulable features;	Manifestation of quantitative patterns in human conducts through large aggregation;
	Additive nature of human-task interactions	Latent but quantifiable relations between human constructs and environments	Human attributes transferrable across datasets
Epistemological commitments	Sociocultural influences on constructs of interest;	Theory-laden results with influences from investigators' personal values;	Results influenced by investigators' personal values;
	Measurement results not as a perfect reflection of absolute reality;	Results aimed toward some degree of generalizability;	Results aimed toward a higher level of generalizability with emphasis on large aggregation
	Tentativeness and error-prone nature of measurement	Aggregate nature of results	
Methodological implications	Measurement to be anchored in frameworks that define constructs conceptually and operationally and that guide consistent use of scoring;	Experimental (or quasiexperimental) designs to be anchored in frameworks that articulate constructs of interests and justifiable relations thereof;	Mining practices to be anchored in frameworks that define and justify relations among constructs of interests;
	Validity not as an inherent property of measurement tools but as legitimacy of connecting evidence to claims	Proper use and fair judgement of aggregate results (versus individual results)	Caution for combining datasets collected at different institutions or administrative levels

and ethical restrictions, and quasicontrolled experiments can only tease apart a very few human attributes and environmental features. On the other hand, large datasets collected from different contexts may still be uniform in some aspect and hence are not so far apart from those generated in controlled experiments. To that end, there can be cases where some quantification practices lie between data mining and controlled investigations (see, for example, Refs. [60,61]). Therefore, researchers need to be sensitive to the important ontological and epistemological issues in both genres so as to take the most suitable actions for data production and interpretation.

As readers may have already noted, this theoretical paper is not centered around technical applications of statistical algorithms. Instead, it provides theoretical discussions concerning the topics of what is being investigated and what is acquired from investigations, topics that directly relate to paradigmatic (ontological and epistemological) underpinnings of quantification practices. These discussions help us form better and fairer views toward what quantitative PER is and what it can do. For example, from the above discussions it is clear that just because we separate, control, and quantify observations does not mean quantitative PER is more objective than other methods of inquiries. In fact, investigators' personal values and perspectives can penetrate into every stage of empirical work even in the most seemingly objective practices of data mining. Therefore, quantitative methods should not be idolized as "the scientific way" of doing PER or other educational research.

On the other hand, we should not demonize quantitative methods as all about statistics or computational algorithms. It is true that statistical or computational techniques are an integral part of quantitative educational research. It is also true that investigators often overlook or downplay the paradigmatic foundations of the research methods they practice. However, these should not become a cause for degrading the value of the quantitative paradigm or any component thereof. In fact, it behooves us—quantitative researchers—to bring those tacit ontological and epistemological underpinnings to the spotlight of scholarly discourse. It is only by doing so that quantitative PER (or quantitative educational research in general) can be properly judged for its merits.

Looking forward, quantitative PER is expected to advance continuously together with qualitative PER. This will likely create a great opportunity for investigators from both paradigms to engage in conversations about possible collaborations. As of now, there already is the emerging mixed-methods approach to PER. However, it is still viewed and practiced as a set of technical procedures to be followed and therefore awaits us, the education researchers, to go beyond the surface level and delve deeper into its theoretical realm. After all, what we are pursuing is not just recipelike guidelines for empirical investigations. Instead, what we are after is something that can tell us why and how we do what we do. That is why understanding the theoretical perspectives of these methods we practice becomes critical.

ACKNOWLEDGMENTS

The work is partially supported by the Ohio State University (OSU) Center on Education and Training for Employment faculty fellowship, by the OSU Office of International Affairs 2018 Academic Enrichment Grant, and by the OSU-Zhejiang Normal University research collaboration initiative grant.

[1] E. F. Redish, *Teaching Physics with the Physics Suite* (John Wiley & Sons, Hoboken, NJ, 2003).

[2] R. Beichner, An introduction to physics education research, in *Reviews in PER: Getting Started in PER*, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2009), Vol. 2.

[3] L. Ding and X. Liu, Getting started with quantitative methods in physics education research, in *Reviews in PER: Getting Started in PER*, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2012), Vol. 2, pp. 1–33.

[4] J. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).

[5] National Research Council, *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering* (National Academies Press, Washington, DC, 2012).

[6] 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).

[7] V. K. Otero and D. B. Harlow, Getting started in qualitative physics education research, in *Reviews in PER: Getting Started in PER*, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2009), Vol. 2.

[8] M. J. Sanger, Using inferential statistics to answer quantitative chemical education research questions, in *Nuts, and Bolts of Chemical Education Research*, edited by D. M. Bunce and R. Cole (American Chemical Society, Washington, DC, 2007), pp. 101–148.

[9] J. W. Creswell, *Qualitative Inquiry & Research Design: Choosing among Five Approaches*, 3rd ed. (Sage Publications, Thousand Oaks, CA, 2013).

[10] J. Corbin and A. Strauss, *Basics of Qualitative Research*, 3rd ed. (Sage Publications, Thousand Oaks, CA, 2008).

[11] E. G. Guba and Y. S. Lincoln, Paradigmatic controversies, contradictions, and emerging confluences, in *Handbook of Qualitative Research*, 3rd ed., edited by N. K. Denzin and Y. S. Lincoln (Sage Publications, Thousand Oaks, CA, 2005), pp. 191–216.

[12] N. K. Denzin and Y. S. Lincoln, Paradigms and perspectives in contention, in *Handbook of Qualitative Research*, 3rd ed., edited by N. K. Denzin and Y. S. Lincoln (Sage Publications, Thousand Oaks, CA, 2005), pp. 183–190.

[13] M. L. Smith, Multiple methodology in education research, in *Handbook of Complementary Methods in Education Research*, edited by J. L. Green, G. Camilli, and P. B. Elmore (Lawrence Erlbaum Associates, Mahwah, NJ, 2006), pp. 457–475.

[14] H. F. Wolcott, Ethnographic research in education, in *Complementary Methods for Research in Education*, edited by R. M. Jaeger (American Educational Research Association, Washington, DC, 1988), pp. 187–249.

[15] H. F. Wolcott, Posturing in qualitative inquiry, in *Handbook of Qualitative Research in Education*, edited by M. D. LeCompte, W. L. Millroy, and J. Preissle (Academic Press, New York, 1992), pp. 3–52.

[16] T. A. Schwandt, Constructivist, interpretivist approaches to human inquiry, *Handbook of Qualitative Research*, 2nd ed., edited by N. K. Denzin and Y. S. Lincoln (Sage Publications, Thousand Oaks, CA, 1994), pp. 118–137.

[17] G. J. Kelly, Epistemology, and educational research, in *Handbook of Complementary Methods in Education Research*, edited by J. L. Green, G. Camilli, and P. B. Elmore (Lawrence Erlbaum Associates, Mahwah, NJ, 2006), pp. 33–56.

[18] N. K. Denzin and Y. S. Lincoln, The discipline and practice of qualitative research, *Handbook of Qualitative Research*, 3rd ed., edited by N. K. Denzin and Y. S. Lincoln (Sage Publications, Thousand Oaks, CA, 2005), pp. 1–32.

[19] A. Schutz, *Collected Papers I: The Problem of Social Reality* (Kluwer Boston, Hingham, MA, 1967).

[20] M. H. Towns, Mixed methods designs in chemical education research, in *Nuts, and Bolts of Chemical Education Research*, edited by D. M. Bunce and R. Cole (American Chemical Society, Washington, DC, 2007), pp. 135–148.

[21] D. T. Campbell, Qualitative knowing in action research, in *The Social Contexts of Method*, edited by M. Brenner,

P. E. Marsh, and M. Brenner (Croom Helm, London, 1978), pp. 184–209.

[22] K. R. Howe, *Closing Methodological Divides: Toward Democratic Educational Research* (Kluwer Academic Publishers, Dordrecht, Netherlands, 2003).

[23] E. L. Thorndike, The nature, purposes and general methods of measurements of educational products, in *The Seventeenth Yearbook of the National Society for Study of Education*, Vol. Part II: The Measurement Of Educational Products, edited by G. M. Whipple (Public School Publishing Company, Bloomington, IL, 1918), pp. 16–24.

[24] D. C. Phillips and N. C. Burbules, *Postpositivism and Educational Research* (Rowman & Littlefield Publishers, Lanham, MD, 2000).

[25] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[26] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure?, Phys. Teach. **33**, 138 (1995).

[27] D. Hestenes and I. A. Halloun, Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[28] P. Heller and D. Huffman, Interpreting the Force Concept Inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).

[29] J. T. Laverty and M. Caballero, Analysis of the most common concept inventories in physics: What are we assessing?, Phys. Rev. Phys. Educ. Res. **14**, 010123 (2018).

[30] M. D. Caballero, E. F. Greco, E. R. Murray, K. R. Bujak, M. Jackson Marr, R. Catrambone, M. A. Kohlmyer, and M. F. Schatz, Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study, Am. J. Phys. **80**, 638 (2012).

[31] L. Ding and M. Caballero, Uncovering the hidden meaning of cross-curriculum comparison results on the Force Concept Inventory, Phys. Rev. ST Phys. Educ. Res. **10**, 020125 (2014).

[32] A. Agresti and B. Finlay, *Statistical Methods for the Social Sciences*, 4th ed. (Pearson Education, Upper Saddle River, NJ, 2009).

[33] K. Danziger, *Constructing the Subject: Historical Origins of Psychological Research* (Cambridge University Press, New York, 1990).

[34] X. Liu, *Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach* (Information Age Publishing, Charlotte, NC, 2010).

[35] P. Cobb, Putting philosophy to work: Coping with multiple theoretical perspectives, in *Second Handbook of Research on Mathematics Teaching and Learning*, edited by F. K. Lester, Jr. (MacMillan, New York, 2007).

[36] National Research Council, *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas* (National Research Council, Board on Science Education, Division of Behavioral and Social Sciences and Education, Washington, DC, 2012).

[37] National Research Council, Next Generation Science Standards, http://www.nextgenscience.org/ (National Research Council, Washington, DC, 2013).

[38] P. V. Engelhardt, An introduction to classical test theory as applied to conceptual multiple-choice tests, in *Reviews in PER: Getting Started in PER*, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2009), Vol. 2.

[39] L. Ding, Applying Rasch theory to evaluate the construct validity of Brief Electricity and Magnetism Assessment, AIP Conf. Proc. **1413**, 175 (2012).

[40] W. J. Boone, J. Staver, and M. S. Yale, *Rasch Analysis in the Human Sciences* (Springer, New York, 2014).

[41] M. T. Kane, Validation, in *Educational Measurement*, edited by R. L. Brennan (Praeger Publishers, Westport, CT, 2006), pp. 17–64.

[42] L. Ding, R. Chabay, and B. Sherwood, How do students in an innovative principle-based mechanics course understand energy concepts?, J. Res. Sci. Teach. **50**, 722 (2013).

[43] L. Ding, Seeking missing pieces in science concept assessments: Reevaluating the Brief Electricity and Magnetism Assessment through Rasch analysis, Phys. Rev. ST Phys. Educ. Res. **10**, 010105 (2014).

[44] R. Henderson, P. Miller, J. Stewart, A. Traxler, and R. Lindell, Item-level gender fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **14**, 020103 (2018).

[45] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).

[46] R. Lindell and L. Ding, Establishing reliability and validity: An ongoing process, AIP Conf. Proc. **1513**, 27 (2012).

[47] Institute of Education Sciences and National Science Foundation, Common guidelines for education research and development (Institute of Education Sciences, U.S. Department of Education, and National Science Foundation, Washington, DC, 2013).

[48] T. D. Cook and V. Sinha, Randomized experiments in educational research, in *Handbook of Complementary Methods in Education Research*, edited by J. L. Green, G. Camilli, and P. B. Elmore (Lawrence Erlbaum, Mahwah, NJ, 2006), p. 551.

[49] L. Ding, Verification of causal influences of reasoning skills and epistemology on physics conceptual learning, Phys. Rev. ST Phys. Educ. Res. **10**, 023101 (2014).

[50] C. Romero and S. Ventura, Data mining in education, WIREs Data Mining and Knowledge Disc. **3**, 12 (2013).

[51] B. Cope and M. Kalantzis, Big data comes to school: Implications for learning, assessment, and research, AERA Open **2**, 1 (2016).

[52] M. A. Waller and S. E. Fawcett, Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management, J. Bus. Logistics **34**, 77 (2013).

[53] B. K. Daniel, Big data and data science: A critical review of issues for educational research, Br. J. Educ. Technol. **50**, 101 (2019).

[54] R. J. Mislevy, J. T. Behrens, K. E. Dicerbo, and R. Levy, Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining, J. Educ. Data Mining **4**, 11 (2012); https://jedm.educationaldatamining.org/index.php/JEDM/article/view/22.

[55] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. d. Baker, *Handbook of Educational Data Mining* (CRC Press Taylor & Francis Group, Boca Raton, FL, 2010).

[56] K. Kim and P. M. Bentler, Data modeling: Structural equation modeling, in *Handbook of Complementary Methods in Education Research*, edited by J. L. Green, G. Camilli, and P. B. Elmore (Lawrence Erlbaum Associates, Mahwah, NJ, 2006), pp. 161–176.

[57] D. Harrison and S. W. Raudenbush, Linear regression and hierarchical linear models, in *Handbook of Complementary Methods in Education Research*, edited by J. L. Green, G. Camilli, and P. B. Elmore (Lawrence Erlbaum Associates, Mahwah, NJ, 2006), pp. 411–426.

[58] E. Wagner and D. Yaskin, Predictive models based on behavioral patterns in higher education, in *Data-Intensive Research in Education: Current Work and Next Steps*, edited by C. Dede (Computing Research Association Washington, DC, 2015), pp. 23–31.

[59] M. Kohlmyer, M. Caballero, R. Catrambone, R. Chabay, L. Ding, M. Hauganet *al.*, Tale of two curricula: The performance of 2000 students in introductory electromagnetism, Phys. Rev. ST Phys. Educ. Res. **5**, 020105 (2009).

[60] R. P. Springuel, M. C. Wittmann, and J. R. Thompson, Applying clustering to statistical analysis of student reasoning about two-dimensional kinematics, Phys. Rev. ST Phys. Educ. Res. **3**, 020107 (2007).

[61] A. Pawl, R. E. Teodorescu, and J. D. Peterson, Assessing class-wide consistency and randomness in response to true or false questions administered online, Phys. Rev. ST Phys. Educ. Res. **9**, 020102 (2013).