Physics Inventory of Quantitative Literacy: A tool for assessing mathematical reasoning in introductory physics

Suzanne White Brahmia,[1,*] Alexis Olsho,[1] Trevor I. Smith,[2] Andrew Boudreaux,[3]
Philip Eaton,[4] and Charlotte Zimmerman[1]

[1]*Department of Physics, University of Washington, Box 351560, Seattle, Washington 98195-1560, USA*
[2]*Department of Physics & Astronomy and Department of STEAM Education,*
Rowan University, 201 Mullica Hill Rd., Glassboro, New Jersey 08028, USA
[3]*Department of Physics & Astronomy, Western Washington University,*
516 High Street, Bellingham, Washington 98225, USA
[4]*School of Natural Sciences and Mathematics, Stockton University, Galloway, New Jersey 08205, USA*

One desired outcome of introductory physics instruction is that students will develop facility with reasoning quantitatively about physical phenomena. Little research has been done regarding how students develop the algebraic concepts and skills involved in reasoning productively about physics quantities, which is different from either understanding of physics concepts or problem-solving abilities. We introduce the Physics Inventory of Quantitative Literacy (PIQL) as a tool for measuring *Quantitative Literacy*, a foundation of mathematical reasoning, in the context of introductory physics. We present the development of the PIQL and evidence of its validity for use in calculus-based introductory physics courses. Unlike *concept* inventories, the PIQL is a *reasoning* inventory, and can be used to assess reasoning over the span of students' instruction in introductory physics. Although mathematical reasoning associated with the PIQL is taught in prior mathematics courses, pretest and post-test scores reveal that this reasoning is not readily used by most students in physics, nor does it develop as part of physics instruction—even in courses that use high-quality, research-based curricular materials. As has been the case with many inventories in physics education, we expect use of the PIQL to support the development of instructional strategies and materials— in this case, designed to meet the course objective that all students become quantitatively literate in introductory physics.

I. INTRODUCTION

Introductory physics uses familiar mathematics in distinct ways to describe the world and make meaning. To an expert, a physics equation "tells the story" of an interaction or process. For example, when reading the equation

$$+10 \text{ N} = -(20 \text{ N/m})x,$$

an expert, noting the units and the signs, may consider it to be an example of Hooke's law. They may then construct a mental story about the force exerted by a spring on a hand as the spring is stretched. The coordinate system is set by the (positive) sign in front of the spring force term which, in the case of a stretched spring, points toward the spring. Although the variable x typically represents the *position* of

an object in mechanics, experts recognize that it is being used here to represent the *displacement* of the end of the spring from its equilibrium position. The negative sign accounts for the fact that both the force and the displacement are one-dimensional vector quantities, and they are always in opposite directions for a spring. Putting these insights together, the expert might construct the story of a spring with a stiffness of 20 N/m being stretched by a hand in the negative direction and exerting a force on the hand of 10 N in the positive direction. Part of the challenge of learning physics is developing the ability to decode symbolic representations in this manner. For physicists, equations tell stories, but students do not always recognize these stories even when they "know" the prerequisite mathematics.

Instruction in college physics courses leans heavily on *Quantitative Literacy*, the interconnected skills, attitudes, and habits of mind that together support the sophisticated use of familiar mathematics for sensemaking [1,2]. Given the mathematical nature of even introductory-level courses in physics, *Physics Quantitative Literacy (PQL)*, i.e., Quantitative Literacy in the context of physics, has the potential to be an important learning outcome for all

*brahmia@uw.edu

students taking introductory physics. PQL is characterized by the blending of conceptual and procedural mathematics to generate and apply quantitative models of physical phenomena, a skill valued in all science, technology, engineering, and mathematics (STEM) fields.

While PQL is a desired outcome of physics instruction, there is a dearth of validated measures of reasoning about quantities and their relationships in physics contexts. Kalinowski and Willoughby developed a *reasoning* inventory to assess college students' scientific reasoning [3]. While their instrument includes several basic proportional reasoning and probability items, its primary focus is on verbal interpretations involving control of variables, hypothesis testing, and correlational reasoning. We have developed the Physics Inventory of Quantitative Literacy (PIQL) to broaden options for assessing quantitative reasoning in physics. The PIQL assesses quantification typically used in introductory physics, with a focus on the mathematical relationships between physics quantities. We intend use of the PIQL in introductory courses to raise awareness among instructors about the extent to which instructional goals surrounding PQL are being met, perhaps in a manner similar to how early concept inventories such as the Force Concept Inventory (FCI) [4] raised awareness that basic concepts were not being understood at the level many instructors assumed. An increased awareness that broad instructional goals are not being met can drive curriculum development to build deeper understanding [5–7]. To help facilitate administration, the PIQL is available online for widespread use [8].

After multiple iterations, the PIQL is in a steady state and we have constructed an argument for its validity in the context of calculus-based introductory physics courses at two selective research universities in the Pacific Northwest, referred to hereafter as *institution 1* and *institution 2*. In this paper, we provide evidence of the PIQL's validity as a measure of development of PQL in this population of students [9]. We recognize that this evidence is not representative of all introductory physics students and does not apply to the validity of the PIQL with all other populations. We acknowledge that physics education research more generally has not always been based on representative samples of students [10]—but like much of this previous work, the analysis and results are informative and useful nonetheless. In future work, we plan to establish evidence for the validity of the PIQL more broadly, modifying the instrument as necessary based on data collected with more diverse groups of introductory physics students.

In the next section, we present the theoretical underpinnings of the PIQL. Section III describes the methods used to develop individual items and collect them into a full instrument. The results of both qualitative and quantitative analyses to provide evidence for the validity of the PIQL for use in a calculus-based introductory physics sequence are presented in Sec. IV. We conclude with a discussion of plans to investigate the validity of the PIQL for use in all introductory physics courses and across multiple instructional settings.

II. THEORETICAL FOUNDATIONS

Physics Quantitative Literacy involves *proceptual* understanding of both the mathematics involved in creating physics quantities and in the quantitative models that relate the quantities to each other. In this section, we describe the role proceptual reasoning plays in the PIQL, how we use the framework of *conceptual blending theory* to interpret proceptual reasoning as a blend of procedural mathematics and physical reasoning, and discuss Sherin's symbolic forms as a poignant framing of proceptual reasoning in introductory physics [11].

Mathematics education researchers Gray and Tall define proceptual understanding as a combination of procedural mastery and conceptual understanding [12]. They explain that in the context of fractions, for example, "the symbol $\frac{3}{4}$ stands for both the process of division and the concept of fraction"; that is, a student with a proceptual understanding of fractions would move fluidly between the procedure of dividing 3 by 4, and the physical instantiation of the fraction $\frac{3}{4}$ as a precise quantification of portion. Similarly, a physics student with a proceptual understanding of the quantity torque would move fluidly between the procedure, finding the vector cross product of a position relative to an origin and a force, and conceptualizing the vector product $\vec{r} \times \vec{F}$ as a quantity unto itself (i.e., as simply $\vec{\tau}$), with its own, important emergent properties and consequences.

The PIQL assesses proceptual reasoning involving algebraic relationships between physics quantities. Such reasoning characterizes mastery of the content typically presented in introductory, calculus-based physics courses. The PIQL is designed to span a knowledge space [13] based on elements of mathematical reasoning ubiquitous in introductory physics. As illustrated in the example in the introduction to this paper, much of the mathematics used in introductory physics courses holds physical significance beyond its strict mathematical meaning. Many (or most) of the PIQL items require proceptual reasoning of the type described above. We regard such reasoning with familiar mathematics—in which "doing math" is not separate from "doing physics"—as meaningful quantitative understanding.

Conceptual blending theory (CBT) [14] provides a framework for understanding the integration of mathematical and physical reasoning involved in PQL [15–17]. In their theory, Fauconnier and Turner describe a cognitive process in which a unique mental space is formed from two (or more) separate mental spaces. This blended space can be thought of as a product of the input spaces, rather than a separable sum. According to CBT, development of expert mathematization in physics would occur not through a

simple addition of new elements (physics quantities) to an existing cognitive structure (arithmetic or algebra), but rather through the creation of new and independent cognitive spaces. These spaces, in which creative, quantitative analysis of physical phenomena can occur, involves a continuous interdependence of thinking about the mathematical and physical worlds. In the context of the PIQL, we consider proceptual reasoning about physical quantities and the relationships between them to be represented by the blended space, which is an inseparable blend of mathematics reasoning and physical meaning making.

As we discuss more fully below, the process of inventory development described by Adams and Wieman [18] guided the development of the PIQL. In this process, phase 1 of development involves delineating what the inventory is intended to measure, i.e., establishing the *test construct* [18]. We developed the test construct for the PIQL based on Sherin's theory of symbolic forms, which in turn was developed to explain how successful introductory physics students understand and construct equations [11]. Sherin's symbolic forms provides a framework for characterizing the reasoning targeted in PIQL.

> … *successful (physics) students learn to understand what equations say in a fundamental sense; they have a feel for expressions, and this understanding guides their work… We do students a disservice by treating (physics) conceptual understanding as separate from the use of mathematical notations [11].*

The symbolic forms framework hypothesizes that successful students can develop expertlike conceptual schema with which they associate certain symbol patterns in equations. Sherin developed a list of these symbolic forms, noting that it is not comprehensive, with the intention that subsequent research could help build up a library [11]. Figure 1 shows examples of symbolic forms that are represented in the PIQL, both from Sherin's original work and from more recent work by other researchers [19,20]. These forms can be considered a representation of proceptual reasoning in physics; they are the outcome of both a

procedural view of related symbols and the physics contexts in which they are relevant. Note that expert reasoning described in the spring example in the introduction to this paper, and in the PIQL item in Fig. 3 rely on the *quantity* and the *scaling* symbolic forms.

Most introductory physics students are less sophisticated in their use of algebraic structures [21], and come from less privileged backgrounds [10] than the students in Sherin's study, who were in the last semester of calculus-based physics at an elite institution. As such, we consider symbolic forms to be a *learning objective* of an introductory physics course, rather than characteristic of how typical physics students think. PIQL is designed as a probe to help researchers and instructors meet this objective.

III. DEVELOPMENT OF THE PIQL

Development of the PIQL rested on two foundations: (i) existing literature in discipline-based education research in mathematics and physics, and (ii) our own research on student reasoning about ratio, about signs and signed quantities, and about covariation. The latter was conducted over an approximately five year period immediately preceding our explicit development of PIQL as a standardized assessment instrument. Many PIQL items have been pilot tested over the past decade as interview prompts, free-response written questions, and/or multiple-choice questions [22–25]. As we drew on these intellectual foundations, we followed an iterative cycle that led to the current version of the PIQL (see Fig. 2). This cycle consisted of assembling a working inventory of multiple-choice items, gathering data from introductory physics students, analyzing the results, identifying areas where the inventory could be improved (in terms of broader content coverage, items that were either too easy or too hard, redundant items, etc.), developing new items, adding and removing items from the working inventory, and repeating data collection and analysis.

Our work was guided in a general way by the four-phase process for developing assessment instruments proposed by Adams and Wieman [18]:

- Phase 1. Delineation of the purpose of the test and the scope of the construct or the extent of the domain to be measured;
- Phase 2. Development and evaluation of the test specifications;
- Phase 3. Development, field testing, evaluation, and selection of the items and scoring guides and procedures; and
- Phase 4. Assembly and evaluation of the test for operational use.

In particular, the process described by Adams and Wieman informed the collection of both qualitative and quantitative evidence for validity of the PIQL.
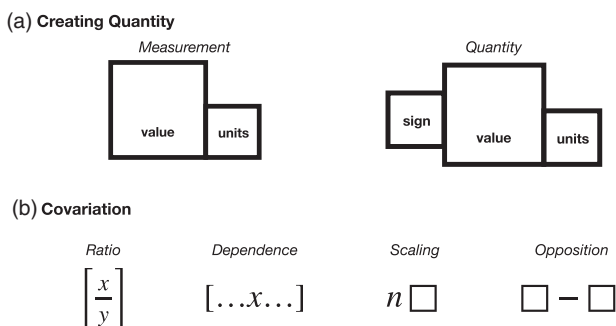

FIG. 1. Examples of symbolic forms represented in the PIQL. (a) Creating Quantity: Measurement [19] and Quantity [20], and (b) Covariation: ratio, dependence, scaling, and opposition [11].
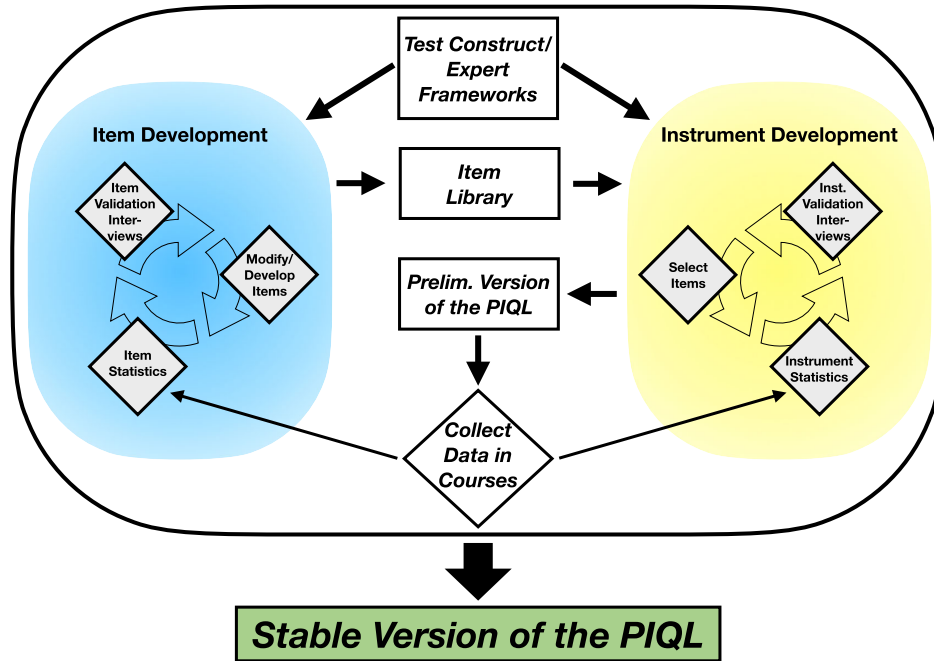
FIG. 2. Workflow for developing and revising items, piloting individual items using interviews, assembling versions of the PIQL, and collecting data with the instrument as a whole. We consider the most recent version of the PIQL to be stable because qualitative and quantitative data analyses provide strong evidence for its validity as a tool for measuring students' PQL (see Sec. IV).

A. Operationalizing PQL

Introductory physics courses present many new and abstract quantities, most of which are ratios, products, sums, or differences of other quantities. Quantities have associated units, and many can be positive or negative, where the sign carries physical meaning (e.g., negative work, positive charge). Quantities can also be vectors or scalars—which have different algebraic rules. Beyond additional meanings specific to the physical context, each of these aspects of quantity has associated *mathematical* reasoning that is rich, nuanced, and challenging, as evidenced by research in mathematics education [1,26,27].

Introductory physics primarily involves basic mathematics, such as arithmetic and algebra. A proceptual understanding of this mathematics is needed to support creative, quantitative reasoning with unfamiliar quantities. PQL thus involves using familiar mathematics in idiosyncratic ways, as exemplified in the spring example in the introduction. The PIQL was designed using three facets of quantitative modeling that are at the heart of quantification in introductory physics: proportional reasoning, reasoning with signed quantities, and covariational reasoning [1,11,27,28]. PIQL items involve reasoning that draws on a nuanced conceptual blend of mathematics and physics content within each facet. We emphasize that the items on the PIQL focus on quantitative reasoning rather than procedural mathematics. While there are specific procedural topics that are not covered (e.g., operations with vectors in

2 or 3 dimensions, or explicit differentiation or integration), many PIQL items probe students' understanding of the underlying conceptual mathematics associated with mathematical procedures (e.g., the meaning of a negative sign associated with a vector quantity, or interpretation of a slope or area under a curve). We believe the conceptual mathematics and quantitative reasoning covered by PIQL items are a foundation for the majority of mathematics used in college-level introductory physics courses.

1. Facet 1: Proportional reasoning

The use of ratios and proportions to describe systems and characterize phenomena is a hallmark of expertise in STEM fields, perhaps especially in physics. Boudreaux, Kanim, and White Brahmia developed a set of six proportional reasoning subskills, based on their analysis of college students' specific difficulties on proportional reasoning assessment items [23]. The subskills are consistent with the early work of Arnold Arons, who identified "underpinnings" to success in introductory physics, such as facility with the operational definitions that underlie familiar mathematical formulas (e.g., area = length × width), the verbal interpretation of ratios, including the slope of a graph, and reasoning with functional relationships (e.g., "if the distance between two point charges doubles, the force decreases by one-quarter") [29]. Three of the subskills—interpreting, applying, and constructing ratios—are particularly important in the PIQL [30].

2. Facet 2: Reasoning about sign

Negative pure numbers present more cognitive challenge than positive pure numbers for precollege mathematics students [31]. Mathematics education researchers have isolated a variety of "natures of negativity" fundamental to algebraic reasoning in the context of high school algebra—the many meanings of the negative sign that successful students distinguish and understand [32–34]. These various meanings of the negative sign form a foundation for scientific quantification, where the mathematical properties of negative numbers are well suited to represent natural processes and quantities. Physics education researchers report that a majority of students enrolled in a calculus-based physics course struggled to make meaning of positive and negative quantities in spite of completing Calculus I and more advanced courses in mathematics [25,35]. Developing "flexibility" with negative numbers, i.e., the recognition and correct interpretation of the multiple meanings of the negative sign, is a known challenge in mathematics education. There is mounting evidence that reasoning about negative quantity poses a significant hurdle for physics students at the introductory level and beyond [36–41].

3. Facet 3: Covariational reasoning

Covariational reasoning, i.e., the formal reasoning about how one variable changes due to small changes in another, related quantity, has been shown by mathematics education researchers to be strongly associated with student success in calculus [42,43]. Early PER encompassed nonlinear scaling and functional reasoning within proportional reasoning, which we recontexualize here using the language of covariational reasoning established by mathematics education researchers [44–46]. Covariational reasoning encompasses all functions that relate two or more quantities and considers the multifaceted ways one can think about those relationships. Through this lens, proportional reasoning could be considered a subset of covariational reasoning that focuses on linear relationships. Considering its prevalence in introductory physics, proportional reasoning will be treated as a separate facet of Quantitative Literacy in this work. Physics education research is only beginning to explore how covariational reasoning is used in introductory physics contexts. Preliminary work suggests that covariational reasoning of physics graduate students ("experts" in introductory physics contexts) differs in some ways from that of mathematics graduate students [47].

B. Assessing reasoning blends

We consider the reasoning associated with each of our facets of PQL to be a product of well-formed *conceptual blends* [14]. In order to assess the reasoning engendered by these blended reasoning spaces, the item library includes several "multiple-choice-multiple-response" (MCMR)

items. Students are told that these items may have more than one correct response, and encouraged to choose all responses they believe are correct. An example MCMR item that we include on the PIQL is shown in Fig. 3. Complete understanding is demonstrated by selecting both D and G. In one study using an open-ended version of this question, White Brahmia observed that students were more accurate in describing the system energy (i.e., equivalent to selecting choice G in Fig. 3) if they also mentioned a statement similar to choice D [25]. A proceptually sophisticated response involves making sense of the negative sign by reasoning about both the orientation of the vector quantities as well as the physical ramifications of these vectors having opposite directions. We hypothesize that it is in the blended space that the reasoning becomes expertlike.

C. Instrument development

Like other inventories in physics, we originally intended the final version of the PIQL to have 20–30 items, and to take 30–40 min for students to complete. The prototype version of the PIQL ("protoPIQL") consisted of 18 items that focused primarily on two facets of PQL: ratios and proportions [22,24,25], and signs and signed quantities [24,36,48]. The protoPIQL also included two items on covariational reasoning taken with permission from the Precalculus Conceptual Assessment (PCA) [49]. The right side of Fig. 2 shows our workflow cycle for creating and pilot-testing preliminary versions of the PIQL. Iterative

---

The following problem is constrained to one dimension. A hand exerts a constant, horizontal force on a block as the block moves along a frictionless, horizontal surface. No other objects do work on the block. For a particular interval of the motion, the hand does $W = -2.7$ J of work on the block. Recall that for a constant force, $W = \vec{F} \cdot \Delta\vec{s}$.

Consider the following statements about this situation. Select the statement(s) that **must be true.** *Choose all that apply.*

   A. The work done by the hand is in the negative direction.

   B. The force exerted by the hand is in the negative direction.

   C. The displacement of the block is in the negative direction.

   D. The force exerted by the hand is in the direction *opposite to* the block's displacement.

   E. The force exerted by the hand is in the direction *parallel to* the block's displacement.

   F. Energy was added to the block system.

   G. Energy was taken away from the block system.

FIG. 3. PIQL MCMR item that exemplifies a proceptual understanding of the negative sign in a mathematics and physics blend. The correct responses are D and G.

Item developed from scratch begins here　　　　　Item developed previously begins here
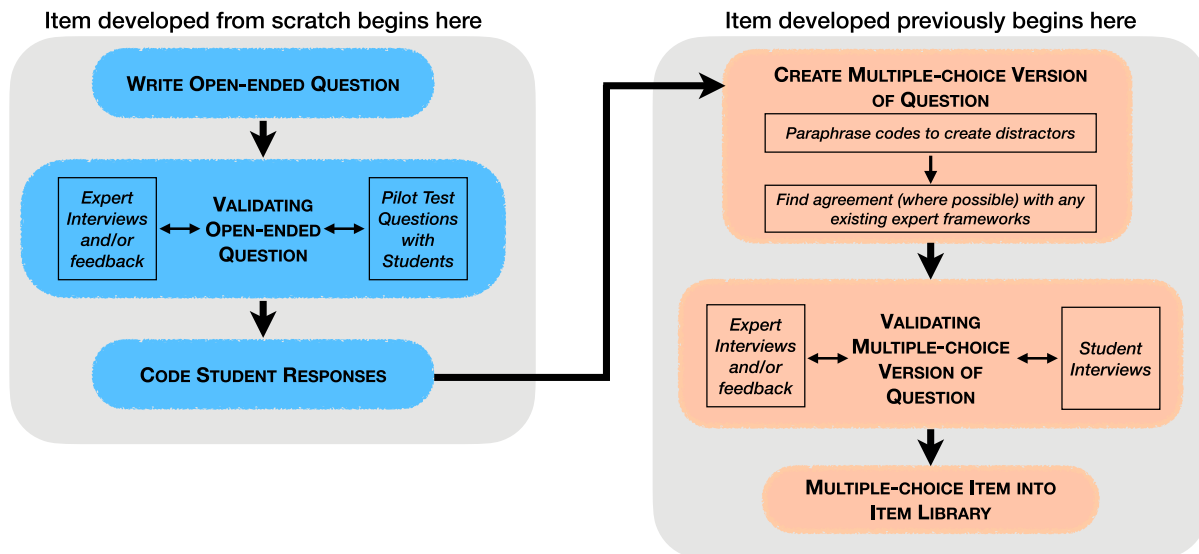

FIG. 4.　The initial development of items for the item library followed the processes shown above. The entire process applies to items that were developed from scratch. Items with existing evidence of validity, or those that had only a small finite number of possible answer options, enter the process at the top of the right side as indicated above.

revisions were made over several years that strengthen the evidence for the validity and reliability of the PIQL, reduce redundancies, and ensure that the three facets of PQL were all represented. Later versions of the PIQL include 20 or 21 items. Because of these iterative revisions, the items on the PIQL in each of the six datasets are slightly different; we label the datasets by their version of the PIQL: protoPIQL, v1.0, v1.1, v2.0, v2.1, and v2.2.

Quantitative evidence for the validity of the PIQL was obtained by administering each version to students in introductory physics courses at institution 1 and performing various statistical analyses of the data (see Sec. IV B for details). We used these results to identify items to remove from the PIQL and items to modify. Qualitative evidence for the validity of the PIQL was obtained through interviews with both physics students and expert physicists and mathematicians (see Sec. IV A for details). Interviews with experts helped to identify the need to develop additional items to meet our desired coverage of all three facets of PQL.

D. Item development

Before creating the PIQL, we assembled a library of multiple-choice items that assess student reasoning in each of our three facets of PQL. Our item library originally contained items created during previous work of some of the authors and their collaborators [22–25,50,51], and items from research on undergraduate mathematics education, including some from the Precalculus Concept Assessment (used with permission) [49]. Figure 4 shows, in detail, our process for developing and modifying items to include in our item library.

The need for developing new items was based on input from the interviews of physics and mathematics experts and our own judgment about the number of items involving each of our three facets of PQL. The need to modify items was often based on item-level statistical analyses of data collected from students at institution 1.

Items developed for the PIQL focus on reasoning rather than computation skill; most items require neither a calculator nor significant mental computation. Most of these items involve reasoning about signed quantities and about the covariation of quantities. Some items were developed "from scratch" and do not have a basis in earlier items; others grew out of an iterative process of item modification.

Most items chosen for the steady-state PIQL—both those that grew out of prior work and those developed specifically for the PIQL—have undergone changes over the course of administration. Modifications of the items took place through an iterative cycle of research, item administration, and item validation. Most modifications were rooted in our deepening understanding of aspects of quantification and quantitative modeling in introductory physics. Generally speaking, modifications to question stems were relatively minor; however, we did make significant changes to many answer choices to encompass not only observed patterns in student reasoning but also multiple natures of expert reasoning. For example, on questions probing student reasoning about the negative sign in a given context, we used results from earlier investigations of student reasoning about negativity to develop item distractors. The wording of these distractors was modified while retaining the student ideas to be consistent with the meanings of the negative sign described

by the framework of the natures of negativity in introductory physics. This framework categorizes various meanings and interpretations of the negative sign in introductory physics contexts [36]; its use as part of our process is shown on the right-hand side of Fig. 4. These distractors were then tested via student and expert interviews to ensure that the correct responses were consistent with expert reasoning and that the distractors were incorrect but consistent not only with common student reasoning but also with expert reasoning that might be correct in other contexts. Student validation interviews also led to improvements in the clarity and purpose of items. For further discussion of the validation process, see Sec. IV.

As an example of an item that underwent substantial iteration, we describe the development of the "Charged Spheres Question," which assesses student interpretation of the negative sign in the context of electric charge. Electric charge involves an idiosyncratic use of sign with quantity in physics—in this context, the sign is an indication of *type* rather than an indication that the quantity is mathematically negative or "less than zero" [52]. An item developed as a part of work investigating student understanding of negativity [24] in the context of the transfer of charge from one object to another was administered on the protoPIQL. When expert validation interviews with this item revealed ambiguities that could not be eliminated through item modification, we began to develop a new item involving the transfer of electric charge. The new item was developed as MCMR, to assess student understanding of the meaning of the negative sign in the context of electric charge, as well as identification of electrons as carriers of charge. Analyses of student responses indicated that the focus on charge carriers was unnecessary. We rewrote the item to focus on the meaning of the negative sign, and changed the wording to reduce ambiguity; the narrowed focus resulted in fewer answer choices, and we were able to rewrite the item as a single-response item. Although the new item shares a physics context (i.e., transfer of electric charge from one item to another) with the original item, the surface features of the current version of the item are distinct. Expert interviews indicate that the final version of the item, shown in Fig. 5, is well-aligned with expert understanding of the meaning of the negative sign in the context of electric charge. With these changes, the classical test theory (CTT) statistics for the item fall within the desired ranges (see Sec. IV for more information on quantitative validation of assessment items).

E. Administration of the PIQL during development

We had two primary goals related to administration of the PIQL during its development: (i) collection of data from students to construct an argument for the validity of the PIQL as a whole test and individual PIQL items and (ii) development of methods for the administration of the

---

A student has two electrically neutral aluminum spheres, A and B. The student performs an experiment that causes charge to move from one of the spheres to the other. After the experiment, the charge of sphere A is $q_A = -5$ microcoulombs, and the charge of sphere B is $q_B = +5$ microcoulombs.
What are the meanings of the negative and positive signs in this context?

   I The signs imply that the charge on sphere A is less than that on sphere B.

  II The signs imply that the unbalanced charges on the two spheres are of opposite types.

 III The signs imply that charge was removed from sphere A and added to sphere B.

  A. I

  B. II

  C. III

  D. I and II

  E. I and III

  F. II and III

  G. I, II, and III

FIG. 5. Current version of the *Charged Spheres* item on the steady-state PIQL, which has withstood multiple challenges to its validity, intended to assess student reasoning about the negative sign in the context of electric charge. Answer B is correct.

PIQL that are scalable and sustainable, and improve its utility for physics instructors and researchers.

During initial development we administered the PIQL to all students enrolled in the 3-quarter, large-enrollment, calculus-based introductory physics sequence at institution 1. We ran versions of the PIQL over eight academic quarters. It was administered at the beginning of each term, before significant instruction, thus serving as a "pretest" for each course of the introductory sequence. Consistent with best practices, the PIQL was originally designed to be administered on paper and in-person (proctored) [53]; however, administration protocols were modified in response to the COVID-19 pandemic. Reference [54] contains a detailed description of both our in-person and online administration methods.

For most in-person administrations of the PIQL, students read items from a 5-page stapled packet and recorded their responses on a paper answer form as well as electronically, where the paper copy served as a backup if the electronic submission failed. Because students entered their responses electronically, we were able to download data in a format that allowed for statistical analyses without substantial cleaning. We developed this hybrid paper-and-electronic data collection system because the inclusion of several multiple-choice-multiple-response items on the PIQL

prevented the use of institution 1's existing system for scanning multiple-choice responses from bubble sheets.

Though our focus was on in-person, proctored administration of the assessment, we began to consider whether online, unproctored administration would better support broader validation and wide-spread dissemination. While existing research suggests little or no significant difference in student performance between proctored and unproctored administrations of some research-based assessments [55–57], researchers recommend that online, unproctored administration be validated separately [55]. We wanted to determine whether the PIQL could be administered online and unproctored by instructors who were reluctant or unable to allocate class time for administration.

The COVID-19 pandemic forced the issue. With both institution 1 and institution 2 moving to all online instruction over a very short time interval, in-person administration of the assessment became impossible. Although online "proctoring" services exist [58], the proctoring requirements do not align well with institution 1's policies regarding computer camera use during virtual instruction, and do not take into account possible limitations on students during such an uncertain and difficult period.

We ran the PIQL unproctored and entirely online using institution 1's existing survey and quiz platform. To mitigate student stress during the rapid shift to online learning, institution 1 suggested that no graded work be required during the first week of instruction. Because we do not grade students' responses to the PIQL for correctness, we decided to run the PIQL, as usual, during the first week of the term in each of the three courses of the calculus-based introductory physics sequence. As we were aware of the tendency of some students to place undue importance on such assessments, however, we presented the PIQL as a low-stakes survey.

We adhered to best practices for online administration of a low-stakes assessment instrument [56,59] as much as possible: we used a time limit substantially longer than the amount of time expected to complete the assessment; we embedded a short (less than 3 min) video at the beginning of the online PIQL that explained the purpose of the assessment and emphasized that course credit would be awarded based on completion rather than correctness; we sent students multiple reminders to complete the PIQL to improve overall participation rate; and we constructed the online version of the PIQL to discourage copying or saving of test items [54]. Little prior work has been done to establish best-practice recommendations regarding the administration of MCMR items in an online format. While Wilcox and Pollock report that student performance on coupled multiple-response items on the Colorado upper-division electrostatics (CUE) diagnostic was similar across online and in-person administration methods, an examination of possible differences was not a focus of the

work [60]. We also note that the coupled multiple-response items on the CUE are a different style of question than the MCMR items on the PIQL, and further, that the CUE and the PIQL are used with different populations of students (upper-division vs introductory level).

Six of the 20 items on the PIQL (v2.2) are MCMR. When the instrument was administered in person, there were multiple opportunities to remind students that they could choose more than one response on these items: both in writing on the instrument itself, and also verbally by the proctor. Validation interviews suggested that multiple reminders were necessary, as this variety of question is relatively rare on the assessments typically encountered by students. Because we recognized that a majority of students completing the survey for the first time would have little-to-no experience with MCMR items, we wanted to increase the likelihood that students would recognize that they could select multiple responses for those items when encountering them online (unproctored). All of the MCMR items were moved to the end of the survey for the online administration. After answering the last multiple-choice-single-response item, students saw a page containing only a statement that the remaining questions on the survey might have more than one correct response, and that students should choose all answers that they feel are correct. At the top of the page for each of the remaining items (all MCMR), students saw a reminder that the question might have more than one correct response. As with the paper version of the PIQL, we also prompted students to "choose all that apply" in the question stem for each MCMR item (see, for example, Fig. 3).

Response rates for the MCMR items suggest that the measures described above were effective; further, overall results from online administrations of the PIQL suggest that students take the online-version of the assessment seriously and perform at roughly the same level as for in-person administration [54]. This is consistent with work that indicates that low-stakes assessments can be successfully administered online [55–57]. We note that small but significant differences are seen in student performance on MCMR items [54]. A preliminary investigation of these differences suggest that the differences in performance on the MCMR items is due to the difference in administration method (i.e., in-person and on paper vs online), and is described in a manuscript in preparation [61].

Our iterative process of continually revising the PIQL combined with the sudden need to administer it online (due to the COVID-19 pandemic) prevented us from collecting student responses on identical versions of the PIQL being administered both in-person and online [62]; however, we have collected (and continue to collect) substantial evidence of the validity of the PIQL as an online instrument (see Sec. IV). Our ultimate goal is to disseminate the PIQL as an instrument that can be administered pre- or post-instruction either online or in person.

IV. EVIDENCE FOR THE VALIDITY AND RELIABILITY OF THE PIQL

One of the major efforts of our project involves establishing evidence for the validity of using the PIQL to measure students' PQL in introductory physics. We established qualitative evidence of validity by conducting interviews with physics experts and students. We established quantitative evidence of validity by administering the PIQL to students in the calculus-based introductory physics sequence and analyzing their responses using a variety of statistical and psychometric methods.

Our goals in establishing various pieces of evidence are to show that

- students interpret individual PIQL items and response choices as intended;
- physics experts agree on the correctness of individual items, and view the PIQL as a whole as a valuable tool for use in introductory physics courses;
- PIQL items represent a breadth of difficulties that are neither unattainable by incoming physics students nor trivial for students completing the introductory sequence;
- student performance on individual PIQL items roughly correlates with their performance on the test as a whole;
- students choose incorrect responses (a.k.a. "distractors") with nontrivial frequency; and
- student responses do not indicate PIQL items are redundant with each other.

The actions taken to analyze data may be seen in the gray diamonds on the sides of Fig. 2. We used the results from these analyses to inform revisions to individual items, the development of additional items, and revisions to the PIQL as a whole in terms of selecting which items to include. These efforts have resulted in a stable instrument with strong qualitative and quantitative evidence for its validity as a measure of students' PQL.

A. Qualitative evidence of validity

Face validity of the PIQL was assessed primarily through expert and student interviews. Interviews were used to collect evidence of the validity of the PIQL and individual PIQL items in three distinct ways:

1. Expert panel reviews were performed to verify that the correct answer choices are indeed correct, as well as to ensure that distractors are incorrect; experts also identified the mathematical content of each question, which allowed us to assess the face validity of the individual items.
2. Individual student interviews were performed to determine whether students are interpreting the questions as intended, to ensure that students are choosing the correct answer for the correct reasons, and to ensure that incorrect responses are chosen for consistent reasons.
3. Expert validation interviews were performed to verify that individual items and the assembled inventory are testing ideas that experts expect their students to learn.

During the expert panel reviews, physics education research faculty and graduate students worked in groups of 3–5 members, with each group seeing 6–8 questions. Panel members worked through each item individually to determine the correct answer and identify the specific mathematical construct(s) required to answer the question. The panels then discussed the items as a group, to come to a consensus about the correct answer and that the incorrect answers were indeed incorrect; together, they also rated the quality of the question in terms of clarity, ambiguity, and appropriateness. Researchers observed the conversations, took notes, and collected the materials afterwards.

For the individual student interviews, students were recruited in approximately equal numbers from each of the three courses in the calculus-based introductory physics sequence at institution 1. During interviews lasting 30–60 min, the students were asked to work through the questions to be validated, following the same "think-aloud" protocol used during the initial development of the assessment items, described above in Sec. III D. The focus of these interviews was not to explore student reasoning about the material, but to ensure that students were interpreting the items as intended, and choosing the correct answer(s) for the right reasons.

A small number of interviews following this protocol were also performed with students enrolled in the calculus-based introductory physics sequence at institution 2. The student validation interviews informed changes to the questions to improve their coherence with the target population.

As we approached a steady-state version of the PIQL we performed validation interviews with both mathematics and physics experts using the PIQL as a whole. Our two mathematics experts are faculty members in the departments of mathematics at two different public universities, both with experience teaching college-level calculus. The validation interview with the mathematics experts was performed as a group interview with both experts attending virtually. Our seven physics experts are faculty members of the department of physics at institution 1 and were chosen for their extensive experience teaching in the introductory physics sequences. We held an in-person group session for the validation interviews with our physics experts. Before both of the group interviews, the experts were asked to complete the PIQL and submit their responses to us. At both group interviews, we discussed each item to ensure that there was consensus for the correct responses and to check that the items were clearly worded. We also discussed in detail the reasoning each item addresses, and the appropriateness of the required reasoning for our target population of students. These expert interviews

served as a final check of the individual items, and of the PIQL as a whole. Experts agreed that, overall, the PIQL represents reasoning that they expect their students to develop during introductory physics courses and that is important in physics generally. We removed one item experts felt did not represent reasoning central to introductory physics. Physics and mathematics experts agreed that administering the PIQL to their students would give important information about the students' quantitative reasoning, and how that reasoning changed over instruction. These results provide evidence for the validity of the PIQL as a whole.

B. Quantitative evidence of validity using classical test theory

We have previously published a detailed account of various quantitative analyses to measure the validity and reliability of each version of the PIQL [63]. Here we present the highlights of this work, emphasizing evidence that the most recent version of the PIQL consistently meets our expectations for quantitative measures of validity and reliability.

Using classical test theory we calculated the difficulty and discrimination parameters for each item; we want to have a wide range of difficulty values with most items between 0.2 and 0.8 (representing the fraction of students who answer each item correctly), and we want most discrimination values to be above 0.3 (representing the difference in CTT difficulty between the top and bottom 27% of students) [64]. We also calculated Cronbach's α as a measure of reliability: a value of at least 0.7 indicates that the test is reliable for measuring the performance of groups of students on a single-construct test, and a value of at least 0.8 indicates that the test is reliable for measuring the performance of individual students [65].

Figure 6 shows the distribution histograms of the CTT difficulty and discrimination parameters for each version of the PIQL. One can see that, with each successive version (ordered from bottom to top in the figure), more items fall within the desired difficulty range (between the red dashed lines). Moreover, the discrimination values get progressively higher, with all 20 items in v2.2 being higher than the desired threshold (red dashed line). Additionally, Cronbach's α went from 0.67 on the protoPIQL to 0.80 on v2.2.

Figure 7 shows the score distribution for each course in which data were collected. As mentioned above, data were collected at the beginning of each course. We consider these to be three snapshots of the introductory course sequence: before mechanics "PreMech," after mechanics but before electricity and magnetism "PostMech," and after electricity and magnetism but before thermodynamics and waves "PostEM." The data in Fig. 7 combine all four administrations of PIQL v2.2, including 2758 data sets and 2078 individuals [66]. A one-way analysis of variance showed no significant difference between the four academic terms, $F(3, 2754) = 1.6$, $p = 0.2$.

Figure 7 shows that students entering the calculus-based introductory physics sequence at institution 1 have moderate levels of PQL before instruction: the PreMech mean and standard deviation are 9.6 ± 3.7 out of 20. Students' PQL increases slightly with instruction, with mean scores 10.8 ± 4.3 for PostMech and 11.3 ± 4.5 for PostEM, but these score distributions show that many students have not achieved high levels of PQL after two introductory courses. Recent studies have shown that differences between student data from pre- and post-instruction research-based conceptual physics assessments typical result in a Cohen's d effect size of about $d = 1$ [67]. Effect sizes from PIQL data are significantly lower: $d = 0.3$ from PreMech to PostMech and $d = 0.1$ from PostMech to PostEM, despite data being
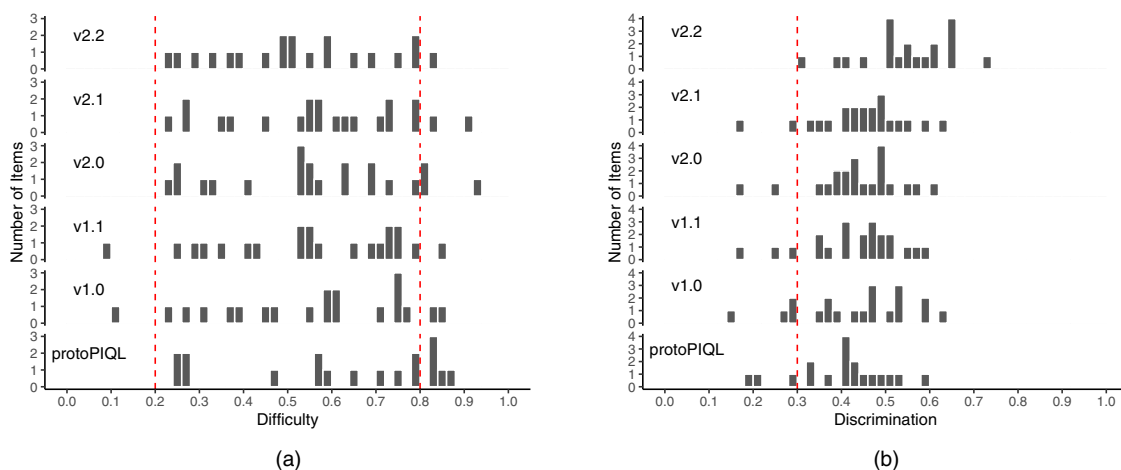

FIG. 6. CTT difficulty (a) and discrimination (b) parameter distributions for all versions of the PIQL. Successive versions are ordered from bottom to top. The desired range of difficulty values is between 0.2 and 0.8 (shown by dashed red lines). The desired range for discrimination is above 0.3.
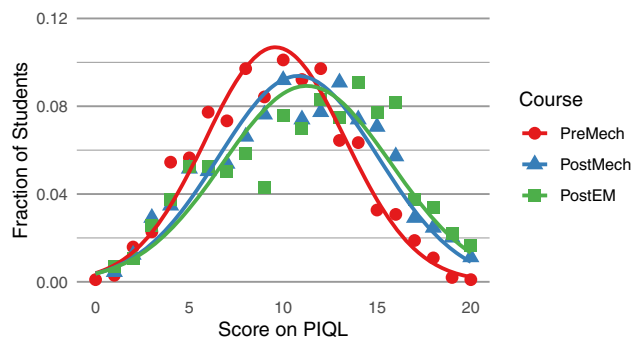
FIG. 7. Distribution of PIQL scores for each course in which data were collected. Data include 2758 response sets from 2078 individuals who responded to PIQL v2.2 in four different academic terms. (Some students provided data in multiple courses at various times.)

collected in courses that make heavy use of research-based instructional materials. This suggests that PQL development, while an implicit goal of many introductory physics courses, may be limited without specific targeted instruction.

C. Exploring the substructure of the PIQL

The PIQL was initially developed to probe student reasoning about three facets of PQL that were defined from an expert's perspective: ratios and proportions, covariation, and signed quantities or negativity (see Sec. III A). In the language of factor analysis, this would imply that the PIQL was originally intended to elicit student responses that are consistent with a three-factor structure. Because the intended factor structure of the PIQL was well understood at the beginning of its development, confirmatory factor analysis (CFA) of student responses was used at the onset, in conjunction with exploratory factor analysis (EFA). CFA is a model-driven statistical method whose goal is to identify the adequacy of a proposed factor model for explaining patterns in response data from the instrument being analyzed [68]. In our work, CFA helps reveal whether or not student response patterns align with the facet-driven model predicted by experts. EFA is a data-driven statistical method whose goal is to uncover the underlying dependencies between observed variables [69]. For all versions of the PIQL, CFA determined that the proposed, facet-driven, expert factor model was not an adequate representation of the latent trait structure of *student* PIQL responses [63,70]. The target threshold for CFA is to have goodness-of-fit statistics such as the confirmatory fit index (CFI) and Tucker-Lewis fit index (TLI) above a threshold of 0.9 [71]. For all versions of the PIQL the CFI and TLI were below 0.8 when using the facet-driven factor model; therefore, students' response patterns did not match our expectations of reasoning developing differently in the areas of ratio and proportion, covariation, and signed quantities or negativity.

Given that the CFA results do not fit with the proposed factor model, we moved on to a more in-depth investigation using EFA. The goal of using EFA was to determine if the PIQL has any substructure, and how closely any substructure aligns with the three facets of PQL. EFA results from preliminary versions of the PIQL showed factor structures that did not align with our three facets of PQL. These results also allowed us to identify redundant items that loaded strongly onto a common factor, excluding all other items; we chose to eliminate redundant items to produce a more efficient instrument.

Analyses of student responses to the most recent versions of the PIQL (v2.0, v2.1, and v2.2) suggest unidimensionality, with no strong substructure among the items [63]. Results from EFA parallel analysis and CFA suggest that these versions may be adequately described by a single factor. Goodness-of-fit analyses show that the CFI and TLI were above 0.93 for both versions under CFA using a unidimensional model. Additionally, the standardized root mean square of the residuals (RMSR) and the root mean square of the error of approximation (RMSEA) were both below 0.04, with a model being considered an adequate representation of the data with values below 0.06 and 0.08, respectively. Collectively these results show strong support for the claim that students answer PIQL items in a manner consistent with a unidimensional model [71]. We consider this to be evidence that PQL is a somewhat coherent construct for students that does not cleanly separate into various facets in the ways expected by experts.

V. ANALYZING DATA FROM MULTIPLE-CHOICE-MULTIPLE-RESPONSE ITEMS

A notable feature of the PIQL is our inclusion of multiple-choice-multiple-response items in which students are instructed to "select all statements that must be true" from a given list, and to "*choose all that apply*" (see Fig. 3 for emphasis used in the original text). As described in Sec. II we consider PQL to be a conceptual blend between physics concepts and mathematical reasoning [14,72]. The MCMR item format has the potential to reveal more information about students' thinking than standard single-response items, and to measure the complexity of ideas that students bring from both of these input spaces [73]. The MCMR item format also poses challenges with data analysis, as typical analyses of multiple-choice tests (such as CTT, EFA, and CFA) assume single-response items.

For MCMR items, dichotomous scoring methods require a student to choose *all* correct responses and *only* correct responses to be considered correct. For example, item 18 on PIQL v2.2 has two correct answer choices: D and G (see Fig. 3). In a dichotomous scoring scheme a student who picks only answer D would be scored the same way as a student who chooses answers E and F (incorrect). This ignores the nuance and complexity of students' response
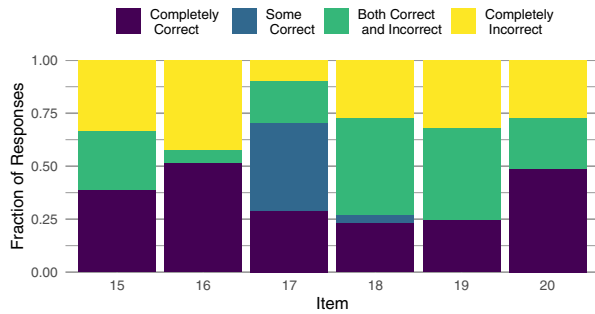
FIG. 8.   Fraction of student responses in each category of our four-level scoring scheme for MCMR items. Items 17 and 18 have multiple correct answers. These results are from PIQL v2.2.

patterns within (and between) items. As such, the CTT results for these items are not entirely representative of students' responses.

In an effort to move beyond the constraints of dichotomous scoring for MCMR items, we have developed a four-level scoring scale in which we categorize students' responses as Completely Correct, Some Correct (if at least one but not all correct response choices are chosen), Both Correct and Incorrect (if at least one correct and one incorrect response choices are chosen), and Completely Incorrect [74,75]. Figure 8 shows the results of using this four-level scoring scale to categorize student responses to the six MCMR items on PIQL v2.2. The dark purple Completely Correct bars are equivalent to CTT difficulty; however, Fig. 8 also shows us that at least 60% of students provide at least one correct response to each item (Completely Correct, Some Correct, and Both Correct and Incorrect combined), although this is often coupled with an incorrect response (6%–45% of students categorized as Both Correct and Incorrect). This tells a very different story than the CTT results, which group the Some Correct, Both Correct and Incorrect, and Completely Incorrect categories together into a broad incorrect category.

These four-level scoring results also reveal differences hidden by dichotomous scoring. For example, on PIQL v2.2 two items (Items 17 and 18) have more than one correct answer choice. Figure 8 shows that roughly equal numbers of students answer these items completely correctly, but item 17 has a much higher fraction of students in the Some Correct category. Students are much more likely to include one of the incorrect responses to item 18 than they are for item 17. Items with multiple correct answers also present a new question: is it better for a student to choose some correct answers or both correct and incorrect answers? The answer may depend on the specifics of each item and the associated answer choices. Future work will include analyzing data from MCMR items to develop a more sophisticated scoring scheme.

To further examine the responses students give to individual PIQL items we use item response curves (IRCs), which show the fraction of students who choose each answer choice as a function of the students' overall score on the PIQL [76–79]. IRCs have been used with single-response tests to rank incorrect responses and to compare different student populations with regard to both correct and incorrect answer choices [78,79]. We find IRCs particularly helpful for examining student responses to items with multiple correct answers.

Figure 9 shows three IRCs with different behavior. Item 14 is a single-response item with correct answer B. Even fairly high-scoring students persist in choosing a particular incorrect answer F [80]. Item 17 has three correct responses (A, C, D), with A being the most commonly chosen, and C being the least commonly chosen. Few students at any score level choose E, and fewer than 20% of students who score above average (10.8) choose either incorrect response (B, E). Item 18 is particularly interesting in that all responses are chosen by 20%–60% of students in the middle score range (8–12), which represents 40% of the students (see Fig. 3 for the text of item 18 and Fig. 7 for the score distribution). This supports the results
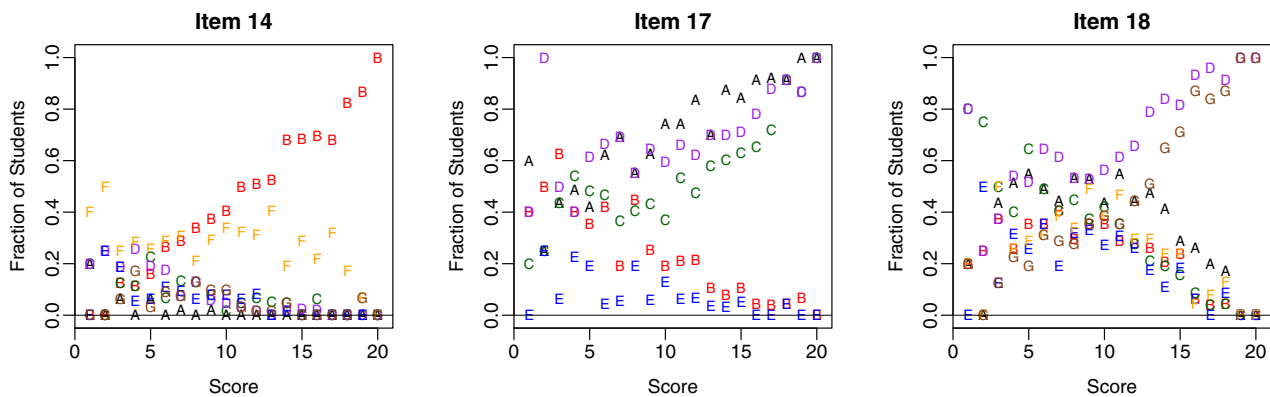

FIG. 9.   Item response curves for three items on PIQL v2.2. Each plot shows the fraction of students who chose each response out of the students who earned each score on the total test. Item 14 has correct answer B, item 17 has correct answers A, C, and D, and item 18 has correct answers D and G.

from Fig. 8 that students are likely to choose both a correct and an incorrect response to item 18.

Both the four-level scoring scheme and the IRCs provide more information than traditional CTT analyses and allow us to see patterns in students' responses that go beyond typical dichotomous scoring methods. We have used these to gain a deeper qualitative picture of student performance on each PIQL item, and these have been very valuable for deciding which items and answer choices to keep, eliminate, or modify. The MCMR item format also opens possibilities for future work to develop rigorous quantitative analyses that can supplement traditional analyses and provide richer information about student understanding and learning.

VI. CONCLUSION

We have developed the Physics Inventory of Quantitative Literacy, an instrument to assess quantitative reasoning in introductory physics contexts. Instruction in introductory physics includes a greater amount and a deeper level of quantitative reasoning than introductory courses in most other STEM disciplines. We have presented expert interview data confirming that the promotion of such reasoning is an important goal of introductory physics courses, which are required for most STEM majors.

We have described the development of the PIQL and presented evidence of its validity. We demonstrated that experts believe that the PIQL is a valid assessment of PQL. Results from quantitative analyses show strong evidence for the validity and reliability of the PIQL at institution 1, and suggest that student responses align with a coherent unidimensional model of PQL. We have also provided evidence that students' PIQL scores do not improve much even after having taken two quarters of calculus-based physics; PQL is not a byproduct of current common research-based instructional practices.

We have supplemented rigorous psychometric analyses with four-level scoring methods for MCMR items and IRCs, which provide additional information about students' choices of both correct and incorrect responses. These analyses informed our revisions of the PIQL. Future work will include developing more sophisticated analyses that can include the nuance of MCMR data into CTT-style analyses.

We continue to collect evidence for validation of the PIQL for use as an online reasoning inventory. We recognize that to construct a more compelling argument for the validity of comparing paper and online versions of the PIQL, we must learn more about how students interact with MCMR test items when using a computer or other internet-capable device. To this end, we intend to develop interview protocols to investigate how students interact with the PIQL, especially its MCMR items, when it is administered online. A more in-depth analysis of student performance on MCMR items for different administration methods is in progress.

Next steps involve collecting data to construct an argument of the PIQL's validity to assess PQL across more diverse student populations. Historically, physics education research studies oversample [10] from large research universities that are similar to institutions 1 and 2, where the work we describe in this paper has been done. The broader population of introductory physics students is a more racially and socioeconomically diverse group, attending a variety of geographically diverse post-secondary institutions [10].

The use of the PIQL can raise awareness in undergraduate physics education, and STEM education more broadly, across varied post-secondary learning environments. We believe that a broader awareness has the potential to inform instructional practices and help drive change that will deepen student PQL. It is essential that materials and methods developed with a relatively homogeneous student sample be validated for the breadth of learning environments in which students take introductory physics. A PIQL with broader evidence of validity will serve as a new tool for improving the ability of instruction to promote Quantitative Literacy.

We also seek to use the PIQL as a tool for characterizing the *development* of students' PQL using additional psychometric analysis methods. We will use methods such as item tree analysis to characterize students' knowledge states at various points within the physics curriculum and explore the potential hierarchy of skills measured by the PIQL [81–83]. We will also modify polytomous models of item response theory for analysis of responses to MCMR items [84–86]. Such methods will extend our use of item response curves, and could lead to a more robust multilevel scoring scheme from which reasoning development patterns are likely to emerge.

Lastly, we note that poorly developed PQL affects students in subsequent courses. Recent research suggests that students enrolled in advanced physics courses, which make use of mathematics such as linear algebra and differential equations, do not necessarily understand the underlying rationale for the mathematical procedures used to solve problems [87–89]. Many students learn procedures effectively. Instructors, as well as the students themselves, would like for these procedures to also be conceptualized [87]. For many students, however, Quantitative Literacy is not a strong outcome of prior physics instruction. Our hope is that the PIQL will serve as a valuable tool for tracking and improving the development of Physics Quantitative Literacy throughout the sequence of undergraduate physics instruction.

[1] P. W. Thompson, Quantitative reasoning and mathematical modeling, in *New Perspectives and Directions for Collaborative Research in Mathematics Education: Papers from a Planning Conference for WISDOM^e*, edited by S. A. Chamberlain, L. L. Hatfield, and S. Belbase (University of Wyoming College of Education, Laramie, WY, 2010), pp. 33–57.

[2] B. Ojose, Mathematics literacy: Are we able to put the mathematics we learn into everyday use?, Journal of Mathematics Education **4** (2011), https://educationforatoz.com/journalofmatheducation/backissues.html.

[3] S. T. Kalinowski and S. Willoughby, Development and validation of a scientific (formal) reasoning test for college students, J. Res. Sci. Teach. **56**, 1269 (2019).

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

[5] 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).

[6] J. Von Korff, B. Archibeque, K. A. Gomez, 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).

[7] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).

[8] The PIQL is available online at https://www.physport.org/assessments/PIQL.

[9] We recognize that no instrument is ever "valid" or "validated," as validity is a quality of the claim being made, not the instrument itself; nor can any claim be *proved* valid. We have attempted to use language that reflects this.

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

[11] B. L. Sherin, How students understand physics equations, Cognit. Instr. **19**, 479 (2001).

[12] E. M. Gray and D. O. Tall, Duality, ambiguity, and flexibility: A "proceptual" view of simple arithmetic, J. Res. Math. Educ. **25**, 116 (1994).

[13] J.-C. Falmagne, M. Koppen, M. Villano, J.-P. Doignon, and L. Johannesen, Introduction to knowledge spaces: How to build, test, and search them, Psychol. Rev. **97**, 201 (1990).

[14] G. Fauconnier and M. Turner, *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities* (Basic Books, New York, 2002).

[15] T. J. Bing and E. F. Redish, The cognitive blending of mathematics and physics knowledge, AIP Conf. Proc. **883**, 26 (2007).

[16] D. Hu and N. S. Rebello, Understanding student use of differentials in physics integration problems, Phys. Rev. ST Phys. Educ. Res. **9**, 020108 (2013).

[17] S. Van Den Eynde, B. P. Schermerhorn, J. Deprez, M. Goedhart, J. R. Thompson, and M. De Cock, Dynamic conceptual blending analysis to model student reasoning processes while integrating mathematics and physics: A case study in the context of the heat equation, Phys. Rev. Phys. Educ. Res. **16**, 010114 (2020).

[18] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2010).

[19] A. Dorko and N. Speer, Calculus students' understanding of area and volume units, Investigations Math. Learn. **8**, 23 (2015).

[20] S. White Brahmia, Quantification and its importance to modeling in introductory physics, Eur. J. Phys. **40**, 044001 (2019).

[21] J. Tuminaro, A cognitive framework for analyzing and describing introductory students' use and understanding of mathematics in physics, Ph.D. thesis, University of Maryland, 2004.

[22] S. White Brahmia, Developing expert mathematization in the introductory physics course: an impedance mismatch, in *Proceedings of the 2nd International Conference On Research, Implementation And Education Of Mathematics And Sciences (2nd ICRIEMS)* (Faculty of Mathematics and Natural Sciences Yogyakarta State University, Karangmalang, Yogyakarta, 2015), ISBN 978-979-96880-8-8.

[23] A. Boudreaux, S. Kanim, and S. White Brahmia, Student facility with ratio and proportion: Mapping the reasoning space in introductory physics, arXiv:1511.08960.

[24] S. White Brahmia and A. Boudreaux, Exploring student understanding of the negative sign in introductory physics contexts, in *Proceedings of the 19th Annual Conference on Research in Undergraduate Mathematics Education, Pittsburgh, Pennsylvania*, edited by T. Fukawa-Connelly, N. Infante, M. Wawro, and S. Brown (2016).

[25] S. White Brahmia, A. Boudreaux, and S. E. Kanim, Developing mathematical creativity with physics invention tasks, arXiv:1602.02033 [Am. J. Phys. (to be published)].

[26] J. Vlassis, Making sense of the minus sign or becoming flexible in "negativity", Learn. Instr. **14**, 469 (2004).

[27] P. W. Thompson and L. Saldanha, Fractions and multiplicative reasoning, in *Research Companion to Principles and Standards for School Mathematics*, edited by J. Kilpatrick, W. G. Martin, and D. Schifter (National Council of Teachers of Mathematics, Reston, VA, 2003) Chap. 7, pp. 95–113.

[28] P. W. Thompson, M. P. Carlson, C. Byerley, and N. Hatfield, Schemes for thinking with magnitudes: A hypothesis about foundational reasoning abilities in algebra, in Epistemic Algebraic Students: Emerging Models of Students' Algebraic Knowing Papers from an Invitational Conference (University of Wyoming College of Education, Laramie, WY, 2014), p. 1.

[29] A. B. Arons, Student patterns of thinking and reasoning, Phys. Teach. **21**, 576 (1983).

[30] A. Boudreaux, S. E. Kanim, A. Olsho, S. White Brahmia, C. Zimmerman, and T. I. Smith, Toward a framework for the natures of proportional reasoning in introductory physics, in *Proceedings of the Physics Education Research Conference, virtual conference*, edited by S. Wolf, M. Bennett, and B. Frank (2020), pp. 45–50, 10.1119/perc.2020.pr.Boudreaux.

[31] J. P. Bishop, L. L. Lamb, R. A. Philipp, I. Whitacre, B. P. Schappelle, and M. L. Lewis, Obstacles and affordances for integer reasoning: An analysis of children's thinking and the history of mathematics, J. Res. Math. Educ. **45**, 19 (2014).

[32] A. Gallardo and T. Rojano, School algebra. Syntactic difficulties in the operativity, in *Proceedings of the XVI International Group for the Psychology of Mathematics Education, Baton Rouge, LA*, edited by D. Kirshner, North American Chapter 1, 265 (1994).

[33] P. W. Thompson and T. Dreyfus, Integers as transformations, J. Res. Math. Educ. **19**, 115 (1988).

[34] T. Nunes, *Schools, Mathematics, and the World of Reality*, edited by R. B. Davis and C. A. Maher (Allyn and Bacon, Boston, MA, 1993), ISBN 9780205134458.

[35] S. White Brahmia, A. Boudreaux, and S. E. Kanim, Obstacles to mathematization in introductory physics, arXiv:1601.01235.

[36] S. White Brahmia, A. Olsho, T. I. Smith, and A. Boudreaux, Framework for the natures of negativity in introductory physics, Phys. Rev. Phys. Educ. Res. **16**, 010120 (2020).

[37] K. Hayes and M. C. Wittmann, The role of sign in students' modeling of scalar equations, Phys. Teach. **48**, 246 (2010).

[38] R. R. Bajracharya, T. M. Wemyss, and J. R. Thompson, Student interpretation of the signs of definite integrals using graphical representations, AIP Conf. Proc. **1413**, 111 (2012).

[39] T. Huynh and E. C. Sayre, Blending mathematical and physical negative-ness, in *Rethinking Learning in the Digital Age: Making the Learning Sciences Count, 13th International Conference of the Learning Sciences (ICLS) 2018*, Vol. 2, edited by J. Kay and R. Luckin (International Society of the Learning Sciences, London, 2018), pp. 957–960 [arXiv:1803.01447].

[40] M. Eriksson, C. Linder, and U. Eriksson, Towards understanding learning challenges involving sign convention in introductory level kinematics, *Proceedings of the 2018 Physics Education Research Conference, Washington DC* (2018), 10.1119/perc.2018.pr.Eriksson.

[41] S. Ceuppens, L. Bollen, J. Deprez, W. Dehaene, and M. De Cock, 9th grade students' understanding and strategies when solving $x(t)$ problems in 1D kinematics and $y(x)$ problems in mathematics, Phys. Rev. Phys. Educ. Res. **15**, 010101 (2019).

[42] L. A. Saldanha and P. W. Thompson, Re-thinking covariation from a quantitative perspective: Simultaneous continuous variation, in *Annual Meeting of the Psychology of Mathematics Education—North America*, Vol. 1, edited by S. B. Berenson and W. N. Coulombe (North Carolina State University, Raleigh, NC, 1998), pp. 298–304.

[43] P. W. Thompson, Images of rate and operational understanding of the fundamental theorem of calculus, Educ. Studies Math. **26**, 229 (1994).

[44] M. Carlson, S. Jacobs, E. Coe, S. Larsen, and E. Hsu, Applying Covariational Reasoning While Modeling Dynamic Events: A Framework and a Study, J. Res. Math. Educ. **33**, 352 (2002).

[45] K. C. Moore, T. Paoletti, and S. Musgrave, Covariational reasoning and invariance among coordinate systems, J. Math. Behav. **32**, 461 (2013).

[46] M. Oehrtman, M. Carlson, and P. W. Thompson, Foundational reasoning abilities that promote coherence in students' function understanding, in *Making the Connection: Research and Teaching in Undergraduate Mathematics Education*, Vol. 73 (Mathematical Association of America, 2008) 1st ed., pp. 27–42.

[47] C. Zimmerman, A. Olsho, S. White Brahmia, M. Loverude, A. Boudreaux, and T. I. Smith, Toward understanding and characterizing expert physics covariational reasoning, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT* (2019), 10.1119/perc.2019.pr.Zimmerman.

[48] S. White Brahmia and A. Boudreaux, Signed quantities: Mathematics based majors struggle to make meaning, in *Proceedings of the 20th Annual Conference on Research in Undergraduate Mathematics Education, San Diego, California*, edited by A. Weinberg, C. Rasmussen, J. Rabin, M. Wawro, and S. Brown (2017).

[49] M. Carlson, M. Oehrtman, and N. Engelke, The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings, Cognit. Instr. **28**, 113 (2010).

[50] E. Cohen and S. E. Kanim, Factors influencing the algebra "reversal error", Am. J. Phys. **73**, 1072 (2005).

[51] S. White Brahmia, Mathematization in introductory physics, Ph.D. thesis, Rutgers University-Graduate School-New Brunswick, 2014.

[52] A. Olsho, S. White Brahmia, T. Smith, and A. Boudreaux, When negative is not "less than zero": Electric charge as a signed quantity, Phys. Teach. **59**, 253 (2021).

[53] A. Madsen, S. B. McKagan, and E. C. Sayre, Best practices for administering concept inventories, Phys. Teach. **55,** 530 (2017).

[54] A. Olsho, S. White Brahmia, T. I. Smith, C. Zimmerman, A. Boudreaux, and P. Eaton, Online administration of a reasoning inventory in development, in *Proceedings of the 2020 Physics Education Research Conference, virtual conference*, edited by S. Wolf, M. Bennett, and B. Frank (2020), pp. 382–387, 10.1119/perc.2020.pr.Olsho.

[55] S. Bonham, Reliability, compliance, and security in web-based course assessments, Phys. Rev. ST Phys. Educ. Res. **4,** 010106 (2008).

[56] 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).

[57] B. R. Wilcox and S. J. Pollock, Investigating students' behavior and performance in online conceptual assessment, Phys. Rev. Phys. Educ. Res. **15,** 020145 (2019).

[58] Many of these systems can be described more accurately as surveillance—they cannot interact with the students or provide a physical presence.

[59] 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).

[60] B. R. Wilcox and S. J. Pollock, Validation and analysis of the coupled multiple response Colorado upper-division electrostatics diagnostic, Phys. Rev. ST Phys. Educ. Res. **11,** 020130 (2015).

[61] A. Olsho, T. Smith, S. White Brahmia, P. Eaton, C. Zimmerman, and A. Boudreaux, Effect of administration method on student performance on multiple-choice/multiple-response items, arXiv:2106.02137.

[62] The first version of the PIQL administered online (v2.2) differs by only one item from the final version administered in-person (v2.1).

[63] T. I. Smith, P. Eaton, S. White Brahmia, A. Olsho, A. Boudreaux, and C. Zimmerman, Toward a valid instrument for measuring Physics Quantitative Literacy, *Proceedings of the 2020 Physics Education Research Conference*, virtual conference, edited by S. Wolf, M. Bennett, and B. Frank (2020), pp. 496–502, 10.1119/perc.2020.pr.Smith_T.

[64] W. Wiersma and S. G. Jurs, *Educational Measurement and Testing*, 2nd ed. (Allyn & Bacon, Boston, MA, 1990).

[65] R. L. Doran, *Basic Measurement and Evaluation of Science Instruction.* (National Science Teachers Association, Arlington, VA, 1980).

[66] Because data from four academic terms are combined, some individuals provided multiple response sets at different times in their academic sequence.

[67] J. M. Nissen, R. M. Talbot, A. N. Thompson, and B. Van Dusen, Comparison of normalized gain and Cohen's *d* for analyzing gains on concept inventories, Phys. Rev. Phys. Educ. Res. **14,** 010115 (2018).

[68] T. A. Brown, *Confirmatory Factor Analysis for Applied Research*, 2nd ed. (The Guilford Press, New York, 2015), pp. 72–75.

[69] D. N. Lawley and A. E. Maxwell, *Factor Analysis as a Statistical Method* (Butterworths, London, 1963), pp. viii.

[70] T. I. Smith, P. Eaton, S. White Brahmia, A. Olsho, A. Boudreaux, C. DePalma, V. LaSasso, S. Straguzzi, and C. Whitener, Using psychometric tools as a window into students' quantitative reasoning in introductory physics, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT*, edited by Y. Cao, S. Wolf, and M. Bennett (2019), 10.1119/perc.2019.pr.Smith_T.

[71] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14,** 010124 (2018).

[72] S. White Brahmia, A. Olsho, A. Boudreaux, T. I. Smith, and C. Zimmerman, A Conceptual Blend Analysis of Physics Quantitative Literacy Reasoning Inventory Items, in *Proceedings of the 23rd Annual Conference on Research in Undergraduate Mathematics Education, Boston, MA*, edited by S. S. Karunakaran and A. Higgins (2020), pp. 862–867.

[73] We note that items 4 and 14 on the PIQL are single-response items that could be written as MCMR; these items had few enough answer choices that we could predefine combinations and use a single-response format.

[74] T. I. Smith, S. White Brahmia, A. Olsho, A. Boudreaux, P. Eaton, P. J. Kelly, K. J. Louis, M. A. Nussenbaum, and L. J. Remy, Developing a reasoning inventory for measuring Physics Quantitative Literacy, arXiv:1901.03351.

[75] T. I. Smith, S. White Brahmia, A. Olsho, and A. Boudreaux, Developing a reasoning inventory for measuring Physics Quantitative Literacy, in *Proceedings of the 22nd Annual Conference on Research in Undergraduate Mathematics Education, Oklahoma City, Oklahoma*, edited by A. Weinberg, D. Moore-Russo, H. Soto, and M. Wawro (2019), pp. 1181–1182.

[76] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, Testing the test: Item response curves and test quality, Am. J. Phys. **74,** 449 (2006).

[77] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the Force Concept Inventory, Am. J. Phys. **80,** 825 (2012).

[78] P. J. Walter and G. Morris, Assessing student learning and improving instruction with transition matrices, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA*, edited by D. L. Jones, L. Ding, and A. Traxler (2016), pp. 376–379, 10.1119/perc.2016.pr.089.

[79] M. Ishimoto, G. Davenport, and M. C. Wittmann, Use of item response curves of the Force and Motion Conceptual Evaluation to compare Japanese and American students' views on force and motion, Phys. Rev. Phys. Educ. Res. **13,** 020135 (2017).

[80] For the discussion about using IRCs we have purposely omitted the details of the context of items 14, 17, and 18. We want to highlight the use of IRCs to help identify different trends in student answers and see how these different between single-response and multiple-response items. A detailed discussion of why students answer these particular items in these particular ways is beyond the scope of this paper.

[81] N. T. Young and A. F. Heckler, Observed hierarchy of student proficiency with period, frequency, and

angular frequency, Phys. Rev. Phys. Educ. Res. **14,** 010104 (2018).

[82] M. Schrepp, On the empirical construction of implications between bi-valued test items, Math. Soc. Sci. **38,** 361 (1999).

[83] A. Sargin and A. Ünlü, Inductive item tree analysis: Corrections, improvements, and comparisons, Math. Soc. Sci. **58,** 376 (2009).

[84] R. D. Bock, Estimating item parameters and latent ability when responses are scored in two or more nominal categories, Psychometrika **37,** 29 (1972).

[85] R. D. Bock and I. Moustaki, Item response theory in a general framework, in *Handbook of Statistics*, Vol. 26, edited by C. R. Rao and S. Sinharay (Elsevier, London, 2007), Chap. 15, pp. 469–514.

[86] Y. Suh and D. M. Bolt, Nested logit models for Multiple-Choice Item Response Data, Psychometrika **75,** 454 (2010).

[87] M. D. Caballero, B. R. Wilcox, L. Doughty, and S. J. Pollock, Unpacking students' use of mathematics in upper-division physics: Where do we go from here?, Eur. J. Phys. **36,** 065004 (2015).

[88] T. Smith, J. Thompson, and D. Mountcastle, Student understanding of Taylor series expansions in statistical mechanics, Phys. Rev. ST Phys. Educ. Res. **9,** 020110 (2013).

[89] T. I. Smith, D. B. Mountcastle, and J. R. Thompson, Student understanding of the Boltzmann factor, Phys. Rev. ST Phys. Educ. Res. **11,** 020123 (2015).