Assessing physics quantitative literacy development in algebra-based physics

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

[1]*Department of Physics, University of Washington, Box 351560, Seattle, Washington 98195-1560, USA*
[2]*Department of Physics and Meteorology, United States Air Force Academy, 2354 Fairchild Drive, USAF Academy, Colorado 80840 USA*
[3]*Department of Physics and Astronomy and Department of STEAM Education, Rowan University, 201 Mullica Hill Road, Glassboro, New Jersey 08028, USA*
[4]*School of Natural Sciences and Mathematics, Stockton University, Galloway, New Jersey 08205, USA*

(Received 20 April 2025; accepted 5 June 2025; published 18 July 2025)

Quantitative reasoning is an essential learning objective of physics instruction. The Physics Inventory for Quantitative Literacy (PIQL) is a published assessment tool that has been developed for calculus-based physics courses to help instructors evaluate whether their students learn to reason this way. However, the PIQL is not appropriate for the large population of students taking physics who are not enrolled in, or have not completed, calculus. To address this need, we have developed the General Equation-based Reasoning inventory of QuaNtity (GERQN). The GERQN is an *algebra-based version* of the PIQL and is appropriate for most physics students; the only requirement is that students have taken algebra, so they are familiar with the use of variables, negative quantities, and linear functions. In this paper, we present the development and validation of the GERQN, and a short discussion on how the GERQN can be used by instructors to help their students learn.

DOI: 10.1103/lnd6-pxyt

I. INTRODUCTION

Reasoning quantitatively is a learning objective of many physics courses, and reasoning about mathematical models and their physical meaning is at the heart of what physicists do. Mathematical modeling of the physical world involves generating, translating, and interpreting the physical meaning of mathematical representations, and developing the skills to do so is a valued learning outcome of an introductory physics course [1–4].

In this work, we focus on one piece of modeling: quantitative literacy. The idea of quantitative literacy was introduced by mathematics education researchers, and is related to the practice of using familiar mathematics to represent the world [5]. The Physics Inventory of Quantitative Literacy (PIQL, pronounced "pickle") assesses *physics* quantitative literacy (PQL), or quantitative literacy situated in physical contexts. It is a reasoning inventory that can be used to measure the degree to which students learn to reason quantitatively as a result of taking physics courses [6]. The PIQL is written for a student population that has either completed or is co-enrolled in calculus, and therefore includes reasoning about vectors, limits, and changing rates of change. Research has shown that reasoning this way is challenging for many; students do not readily saturate the PIQL even after a year of college-level instruction, and scores are unlikely to change without direct instruction [6].

PQL relies largely on algebraic fundamentals typically taught in middle school and early high school, such as reasoning about ratio and rates of change. There is an opportunity for this kind of reasoning to be a learning objective of earlier physics courses. However, the PIQL is already difficult for students enrolled in college-level calculus-based physics [6]. A version of the PIQL that is specifically designed for algebra-prerequisite courses is needed. Such an instrument could then be used to help support the large number of students enrolled in algebra-based physics at both the precollege and college level develop PQL.

To address this need, we have developed an algebra-based version of the PIQL: the Generalized Reasoning-based inventory of QuaNtity (GERQN, pronounced "gherkin"). We intend the GERQN to expand the target population to include students who have completed algebra I; this necessarily also entails expanding the population of experts whom we expect to engage with the inventory. In this paper, we present the GERQN and describe its development and validation. This process included interviews with mathematics learning experts, physics teaching

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

professionals in high schools, and physics experts who teach at the college level. We conclude with a short reflection on curricular implications and how the assessment could be used by instructors to inform activity development.

II. BACKGROUND

The GERQN relies on the same theoretical framework as the PIQL [6], adjusted to an algebra-based level. In this section, we first summarize the facets of PQL that were used in the development of the PIQL for calculus-based courses, and aspects that are also prevalent in algebra-based courses. We then discuss our framework for test development, and theoretical basis for the statistical measures used to establish test validity.

A. Three facets of PQL

We define quantitative literacy as "the interconnected skills, attitudes, and habits of mind that together support the sophisticated use of familiar mathematics for sense making" [5–7]. *Physics* quantitative literacy (PQL) refers to quantitative literacy situated in physics contexts. In this way, PQL represents a conceptual blend between physics and mathematical reasoning [8–10]. The PIQL—the Physics Inventory of Quantitative Literacy—was developed to address the need for a validated reasoning inventory that can help instructors measure student development of PQL [6]. It was designed to measure three main facets of PQL: reasoning about signs, covariational reasoning, and proportional reasoning.

Reasoning in physics about the meaning of sign and signed quantities is nuanced [11]. For example, a negative sign could mean an amount of a quantity with respect to zero, a decrease in amount, a direction, or a type of charge—just to name a few. Mathematics education research has attended to the challenges students face when interpreting the negative sign; physics education research has expanded on this work to describe the additional difficulty of interpreting the meaning of negative physical quantities [12–15]. PQL associated with signed quantities includes recognizing the physical interpretation of the negative sign in various contexts, and distinguishing between negative quantities and negative rates of change. It also includes flexibly incorporating negative signs when translating between symbolic, graphical, and written representations of physical scenarios.

The use of vector quantities, and attention to their sign symbolically, varies considerably from algebra-based to calculus-based physics courses. For example, while dot products and cross products are common in calculus-based courses, algebra-based courses typically only include vector addition and decomposition [16]. Treatment of Hooke's law provides another common example—while calculus-based textbooks are likely to include the negative sign in the definition ($\vec{F} = -k\Delta\vec{x}$), algebra-based textbooks often treat the discussion of the negative sign as conceptual and provide a definition based on magnitudes ($F = k\Delta x$) [17,18].

Covariational reasoning refers to reasoning about the change in one quantity with respect to the change in another related quantity. It is commonly used in research on undergraduate mathematics education when studying student reasoning in precalculus and calculus courses [19–21]. In the physics education research literature, covariational reasoning is most often referred to as "scaling" [22–24]. However, recent work has leveraged the language of covariation to consider more specifically how students and experts describe changing quantities in physics and across STEM more broadly [25–28].

Mathematics education researchers have developed and iterated on a framework for covariation that clearly identifies specific types of covariation and activities associated with each type [19–21]. Recent work in physics education has built on this framework to operationalize covariational reasoning specifically with physics quantities [29]. The PIQL contains covariational reasoning items that ask students to analyze changing rates of change symbolically and graphically, reason about discrete covariation with multiple variables (e.g., "If this quantity doubles and another is tripled, what happens to a third quantity?"), and apply reasoning using limits.

Proportional reasoning has been studied in physics education research as ratio or as describing a function-based relationship between two quantities [30–32]. Research has described several specific ways that proportional reasoning is foundational to introductory physics, including proportion as unit rate (i.e., this amount of one quantity "for every" unit of that quantity) and using a ratio to determine an unknown quantity [33]. Historically, proportional reasoning has been used in physics education research to describe many different kinds of function-based relationships between quantities. In this work, we consider proportional reasoning to be the linear subset of covariational reasoning. That is, proportional reasoning describes covariational reasoning about quantities related by a linear function.

We include proportional reasoning as a facet of the same importance as reasoning about sign and covariational reasoning because of the high prevalence of linear relationships in introductory physics. This is especially true in algebra-based introductory physics, where more complex functions are often left for future courses. For example, it is common to treat drag forces that are nonlinear with respect to velocity only in calculus-based courses. It is also common practice to discuss changing rates of change qualitatively in algebra-based courses, but not expect students enrolled in such courses to solve symbolic expressions of nonlinear functions for their rate of change. For example, nonconstant acceleration is rarely taught using symbolic representations and procedures in an

algebra-based course [17,18]. However, interpreting the meaning of graphical representations of changing rates of change is often considered within the scope.

We use these facets as they represent a foundation of the kinds of mathematical reasoning essential in physics, they are identified by experts as valuable, and they are strongly represented in research conducted about mathematical reasoning in scientific contexts. We note, however, that prior research on the PIQL found evidence that for many students these facets are indistinguishable in student response patterns. There is reason to suspect that students do not readily distinguish between reasoning about sign, proportional reasoning, and covariation the way that an expert might [6].

Our aim in developing the GERQN was to create an assessment to measure reasoning about all three (expert) facets of PQL, but at an appropriate level for students who have completed algebra I. Research about precalculus and calculus resources has argued that much of the reasoning we associate with calculus stems from reasoning developed in algebra—namely, the meaning of symbols and reasoning about changes in quantities (covariational reasoning), with a particular focus on constant rates of change (proportional reasoning) [1,27,32,34–38]. Therefore, we consider our aim to be to measure *calculuslike* ideas without requiring the symbolic infrastructure, procedures, or reasoning about infinitesimal change that one may associate with a calculus-based course.

B. Test development

We used the protocol described by Adams and Wieman [39] to guide the development and validation process of the GERQN. The Adams and Wieman framework is intended for creation of a "formative assessment of instruction"—that is, an instrument that can be used to measure changes in student reasoning as a result of instruction *in order to inform instructional change*. The mixed-methods approach proposed by Adams and Wieman is commonly followed (see, for example, in Refs. [40–44]), and is particularly well-suited for guiding the development of instruments such as the GERQN that involves student and expert reasoning currently not well-represented in the physics education research literature. The approach includes multiple rounds of interviews with students, instructors, and other content experts, combined with statistical analyses of individual assessment items and the instrument as a whole. The framework describes a clear process for development of a valid and reliable instrument that can provide a measure of changes in students' PQL over the course of instruction in introductory physics. Moreover, collection of qualitative data during the validation process can ensure that the instrument will be valued by instructors, which can lead to strong uptake and therefore improvement in educational outcomes.

Statistical measures from classical test theory (CTT) can provide quantitative evidence for the validity and reliability of the GERQN [6,39,45–47]. Assembling a test containing items with a broad range of item difficulties and high values of item discrimination (or the point-biserial coefficient) provides evidence of validity that the test can accurately measure a broad range of students' reasoning. High values of test-level statistics such as Cronbach's α and Ferguson's δ provide evidence of the reliability of the test, as well as justification for interpreting a single-value test score as a meaningful representation of student reasoning. In addition to CTT statistics, exploratory and confirmatory factor analyses provide evidence for whether the test measures a single knowledge and reasoning construct or several distinct constructs [6,48,49].

III. METHODS

We designed the GERQN by modifying PIQL items, developing new items as needed, and removing items that were beyond the scope of the intended audience. We followed the procedure described by Adams and Wieman and used in the development of the PIQL [6,39].

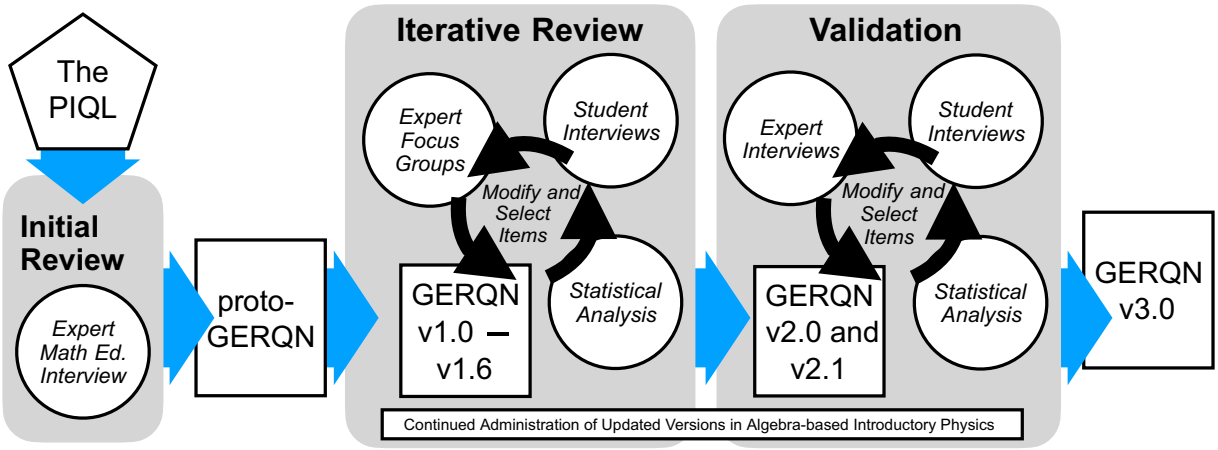


FIG. 1. Development and validation process of the GERQN.

Figure 1 describes the three stages of development and validation:

1. Initial review: We examined and modified the PIQL through the lens of algebra-based reasoning, based on prior research and an interview with a middle school mathematics education researcher. The result was a prototype or "protoGERQN."
2. Iterative review: We modified, removed, and developed new items based on student interviews, an expert focus group, and item statistics. This process resulted in versions 1.0–1.6.
3. Validation: We conducted student validation interviews, expert validation interviews, and a final review of the item and inventory statistics. This process resulted in versions 2.0–3.0.

The majority of our data were collected at a large, R1 university in the Pacific Northwest, which we will refer to as the main institution. Three courses make up the year-long, introductory, algebra-based physics sequence at that institution. We refer to these courses as "mechanics," "electricity and magnetism," and "waves and optics." We identify student data based on these course names. For example, student interviews are labeled by the course students were either enrolled in or most recently completed at the time of the interview. We collected quantitative data at the start of each term; thus, we identify the administration datasets as "PreMech" (which corresponds to data collected in the first two weeks of instruction in mechanics), "PostMech" (which corresponds to data collected in the first two weeks of instruction in electricity and magnetism), and "PostEM" (which corresponds to data collected in the first two weeks of instruction in waves and optics). A summary of the data collected in each stage can be found in Table I.

TABLE I. Data sources and N values for each type of data collected during iterative review and validation. Versions 1.1 and 1.4 were each further split into two versions, A and B, equally and randomly across the student populations listed to test early drafts of item revisions. The bold represents the total and distinguishes the values as such.

	Data type	Version	N
Iterative review	**Test administered**	**1.1A/B**	**708**
	PreMech		488
	PostMech		157
	PostEM		63
	Faculty focus group	**1.1A**	**1 group**
	R1 research (physics)		2
	R1 teaching (physics)		4
	Test administered	**1.4A/B**	**726**
	PreMech		231
	PostMech		425
	PostEM		70
	Student interviews	**1.5**	**12**
	PreMech		7
	PostMech		5
	Test administered	**1.6**	**2778**
	PreMech		1116
	PostMech		864
	PostEM		798
Validation	**Student interviews**	**1.6**	**17**
	PreMech		6
	PostMech		5
	PostEM		6
	Faculty interviews	**2.0–2.1**	**6**
	High school (physics)		3
	Two-year college (physics)		1
	HBCU research (physics)		1
	R1 research (math education)		1
	Test administered	**3.0**	**1453**
	PreMech		746
	PostMech		573
	PostEM		134

A. Student interviews

We conducted two rounds of individual interviews with students enrolled in the introductory algebra-based physics sequence at the main institution. Round 1 interviews were analyzed during iterative review, and round 2 interviews were used for validation purposes. We used the same protocol for both rounds. Student interviews gave insight into whether students chose the correct answers for the correct reasons, whether common answer options were missing, and what reasoning students used.

We recruited students via announcements within course-management software and selected participants on a first-come-first-served basis. We continued to recruit students until we had a representative sample across each course in the sequence and across demographic lines including gender, race and ethnicity, and prior experience in math and physics courses. Students were offered $15 gift cards for participating.

Student interviews were semistructured and consisted of one participant and one interviewer. The participant was asked to complete the GERQN while thinking out loud, and the interviewer only spoke to prompt them to remember to talk out loud if they fell silent. After the participant indicated they were finished, the interviewer asked follow up questions to clarify the participant's reasoning. Interviews were audio-recorded and transcribed using Otter.ai software [50]. These transcripts were subsequently hand-corrected. Any written work produced by students during the interview was also collected.

Interviews were analyzed by a subset of the research team. For each item, a subset of the research team coded responses based on whether the item was answered correctly, and the general reasoning approach used by the student to arrive at their answer. These codes were confirmed by a separate research team member who coded a subset of the data individually, and discussed with the initial coders until uniform consensus was reached.

The results were discussed across the research team to determine whether the question prompt and/or answer choices needed to be revised, or whether to remove the item completely. In instances where the research team disagreed about how to amend the item, two versions were administered. Responses by students who had not taken calculus were more heavily considered in evaluating the mathematics level of the items.

B. Expert interviews

We conducted two rounds of expert interviews: one focus group during iterative review, and a set of individual interviews during validation. For the focus group, faculty who had experience teaching in the algebra-based sequence at the main institution were recruited. We included both research-focused and teaching-focused faculty. Members of the research team reached out to candidates directly; all of those who were interested and willing were invited to participate. Participants were asked to complete the GERQN on their own before the interview, and then met together at one time with two members of the research team. The faculty focus group was intended to provide insight into whether the inventory as a whole was considered valuable by instructors, and whether it would be a useful measure of their students' reasoning. We also asked instructors for feedback on the scope of individual items for the student population they teach.

In the validation phase, we conducted individual faculty interviews with experienced instructors across a variety of institutions. Interview candidates were identified by their experience teaching, implementation of research-validated methods, and experience with teacher preparation. Our aim was to populate the expert pool with the breadth of instructors we expect to use the GERQN. Therefore, we included several high-school physics teachers (in private and public institutions), a community college physics faculty member, a physics faculty member situated at a historically black university in the American South, and a mathematics education researcher who specializes in teacher preparation around algebra and calculus reasoning. None of the experts who were interviewed for validation had been previously interviewed about the GERQN. We did not reinterview faculty at the main institution individually, as they had already been interviewed and been administering updated versions of the GERQN in their courses—we consider this evidence they already perceived the GERQN as valuable.

Faculty were asked to read the GERQN before their interview. The interviews consisted of 1–3 members of the research team and one faculty member and were conducted and recorded over Zoom. Participants were asked if the test aligned with learning outcomes of their courses, whether there were any individual items that were too difficult or did not align with their goals of the course, and for feedback with respect to readability. Most of the experts we interviewed also provided, unprompted, item-by-item feedback on how their students would interact with the item and suggestions for possible improvements.

C. Test administration

The test was administered online using Qualtrics, as part of normal course activities. Students were expected to complete the assessment for a small amount of course credit based on participation. Each item was displayed on its own page, and students could move forward and backward freely with no time constraints. As revisions were made to the GERQN during the iterative review and validation stages, the most up-to-date version was administered to students each term. When there was a dispute about a particular change among the research team, two versions of the GERQN were administered such that half the students in a given course saw one version and the other half saw the other.

During the iterative review stage, we examined the frequency with which students selected each response option to each item. We often removed or modified rarely selected response options, but kept response options that showed up prominently in either student or expert interviews. Additionally, we performed a variety of statistical and psychometric tests of reliability and validity using dichotomously-scored data. These included calculating item-level statistics (classical test theory item difficulty and item discrimination, using the point biserial correlation), calculating test-level statistics (Cronbach's α and Ferguson's δ), and performing both exploratory and confirmatory factor analyses.

IV. RESULTS FROM INITIAL AND ITERATIVE REVIEW

In this section, we describe the outcomes from initial and iterative review during the development of the GERQN.

A. Initial review

Compared to students enrolled in calculus-based physics, the target population for the GERQN is on average at an earlier stage of their academic career and has taken fewer prior physics and mathematics courses. To determine an appropriate mathematics and reading level for the GERQN, we interviewed a middle school mathematics education researcher. She was asked to read the original PIQL and identify items containing content beyond the scope of algebra I. The discussion between the expert and the research team members resulted in removing items with reference to physics content knowledge or vectors; adjusting items with nonlinear relationships to linear contexts, single variable contexts, or to prompts about global behavior; and adapting item context to reflect more everyday scenarios. These changes are summarized in Table II. In altering the contexts of the items, the aim was to remove direct references to unfamiliar physics quantities while

TABLE II. A summary comparison of the protoGERQN and the PIQL. A sample of representative items of the final GERQN and the PIQL can be found in the Appendix.

	PIQL	protoGERQN
Construct	Calc-based 20 items 45 min	Algebra-based 17 items 30 min
Covariation	Changing rates of change Multivariable	Linear rates of change Single variable
Sign reasoning	Scalars and vectors	Scalars only
Contexts	Physics and everyday	Everyday

maintaining the core reasoning ubiquitous to solving physics problems. The prototype that came out of the initial review is named the "protoGERQN."

The "Bottle" item is an example of an item originally in a nonlinear context that we converted to a linear context (Fig. 2). The PIQL version of this item presents the test-taker with a spherical bottle that is being filled with water. Students are asked to select the graph that correctly relates the volume of water in the bottle as a function of the height of water in the bottle. On the GERQN, the bottle has straight sides and neck.

The "Jogger" item is an example of an item for which we simplified the language. "Jogger" asks students to compare the distance traveled by two joggers; the PIQL version also asks students to identify the reasoning needed to determine which jogger went farther. For the GERQN, we removed the reasoning aspect of the responses with the intent of reducing the effort required to complete the task. In addition, we changed a numeric value in the question statement from a decimal to an integer; the middle school mathematics education expert suggested that this was more appropriate for the target population of the GERQN. This change is also aligned with research in physics education about student reasoning about integers and decimals [51].

Finally, the "Internal Energy" and "Money" items on the PIQL and GERQN respectively provide an example of how we removed physics content from an item, while keeping the required reasoning the same. The PIQL item was intended to probe student understanding of symbolizing with signed quantities, using the context of the first law of thermodynamics ($\Delta U = Q + W$). It does not require understanding of the relevant quantities, meaning that work, heat, and internal energy are defined in the question stem. However, we are aware that students likely to take the GERQN include biology and chemistry majors who may be familiar with different sign conventions [52]. The target population may also include students for whom new physical quantities are distracting. Therefore, we created a similar item in the "real world" context of money, relating the change in money in a wallet, the money spent or earned by paying for or doing a job, and the money spent or earned by selling or buying something. The aim was to write an item that captures the reasoning used in physics—in this case, reasoning about symbolizing signed quantities—without burdening students with unfamiliar quantities or definitions. A more detailed comparison of the PIQL and GERQN items described in this section can be found in the Appendix.

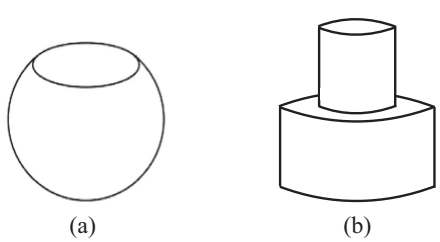


FIG. 2. Bottle figure from (a) the PIQL and (b) the GERQN.

B. Iterative review

Iterative review took place over two years, during which time we conducted interviews with students, conducted a faculty focus group, and administered versions of the assessment to algebra-based introductory physics courses (see Table I).

Student interviews resulted in revisions to several items, either to clarify the prompt, adjust answer options, or change the context of the problem. The final outcome of these revisions was version 1.6. The faculty focus group provided evidence that instructors consider the reasoning required by the GERQN valuable and aligned with the learning objectives of their courses.

One unexpected outcome of the faculty focus group was a discussion of whether the "Growth" item (see Fig. 3) was

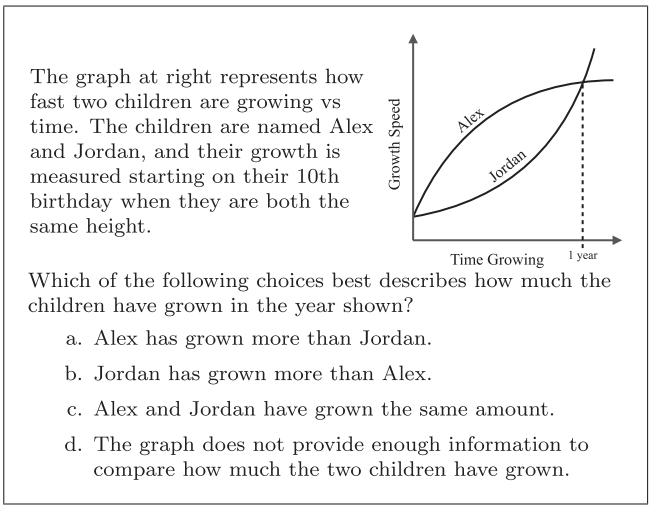


FIG. 3. GERQN item "Growth," where students are asked to compare the growth of two children using a graph of growth rate.

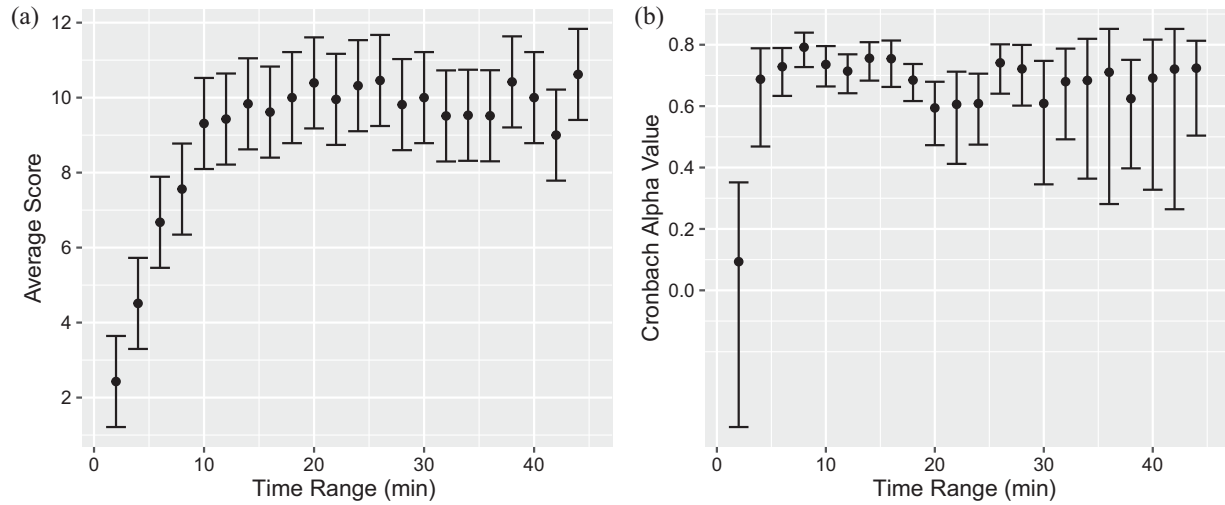


FIG. 4. (a) Average score with respect to time on task and (b) Cronbach's α with respect to time on task, both for GERQN version 3.0 ($N = 2118$). Students who took between 0 and 45 min are shown [These plots show results from 1610 students who completed the test within 45 min. Data beyond this time are too sparse to reliably group into 2-min bins]. Error bars represent 95% confidence intervals.

appropriate for an algebra-based course, or whether it required calculus to solve. After discussion, it was agreed that this item could be solved by noticing that one child grew faster than the other over the entire duration. This reasoning was confirmed in student interviews, where several students used that line of thinking to arrive at the correct answer.

Quantitative measures were used throughout the iterations to ensure that changes made did not negatively impact established statistical results from the PIQL, and as a check that all the answer options were useful to include. For example, distractors that were not chosen frequently were critically examined in student interviews, and those that were never chosen were removed. How item and inventory statistics changed across versions is described in more detail in Sec. V. A comparison of representative items from the GERQN and the PIQL can be found in the Appendix.

V. VALIDATION RESULTS

In this section, we describe how we validated the version of the GERQN that emerged from iterative review. Validation took place over one year, during which time we conducted individual student interviews, conducted individual interviews with experts, and administered version 3.0 in an algebra-based introductory physics sequence (see Table I). Together, these data form the basis from which we sought evidence of validity of the instrument and individual items, both qualitatively and quantitatively.

A. Data-driven cut-off time

Due to the out-of-class and online nature of the assessment, there was an open question of how seriously students approach the items. We binned student responses ($N = 2118$ for v3.0) into 2-min intervals based on the time the students took on the assessment and compared the average score and Cronbach's α of these groups (Fig. 4). These data suggest that, for students who spend fewer than 10 min engaging with the test, there is a strong relationship between time on task and student score, as well as a distinct relationship between time on task and Cronbach's α for very low times. After about 10 min, we see no relationship between either of these quantities and time on task. We use this as evidence that the data from students who spend less than 10 min engaging with the assessment are not representative of the overall population. The statistical analysis we present in this paper was therefore conducted with students who took at least 10 min on the test and left 5 items or fewer unanswered ($N = 1612$ for v3.0 and $N = 2778$ for v1.6) [53].

B. Inventory validation

During individual validation interviews, all physics instructor experts expressed that the items on the GERQN collectively represented desired learning outcomes for their introductory physics courses. They uniformly agreed that student improvement on the assessment would be a valued outcome of completing their courses, and that their students are unlikely to answer all items correctly at the very start of their courses. The emphasis of the different instructors' learning outcomes varied; some were more focused specifically on proportional reasoning than covariation more generally, for example. We interpret this spread as an indication that the test as a whole represents quantitative reasoning that is valued across different institutional contexts and different levels of physics instruction.

The interviews also provided context for the inventory at different levels of mathematics preparation. For example,

one expert teaches physics to freshman in high school in two different courses: one in which students have typically completed algebra I and one in which students are typically co-enrolled in algebra I. This expert expressed that the GERQN would be suitable for both, but their students who have completed algebra I would be more comfortable with the function notation used on several items. The expert noted that students co-enrolled in algebra I may misinterpret functions such as $N(t)$ as $N \times t$. Another expert who also teaches a physics course in which students are co-enrolled in algebra I noted that they found the GERQN less valuable for that population than for their students who had completed algebra I. We take these expert insights as a view into validity for the earliest physics course likely to use the instrument, and conclude that the GERQN is appropriate for students who have completed algebra I and beyond. We suggest the GERQN be used with caution in high-school physics courses for which algebra I is not a prerequisite. Across all experts, we confirmed that for the target population, the mathematics notation and functions used are at an appropriate level.

Quantitative validation of the instrument was established along the following lines:

- test-retest stability via the Pearson correlation coefficient,
- internal consistency and reliability via Cronbach's α,
- discriminatory power via Ferguson's δ, and
- alignment with the single factor observed on the PIQL [6] via exploratory factor analysis.

The results of these analyses are summarized in Table III.

Test-retest stability was established through a Pearson correlation coefficient between two quarters of the same version, administered at the beginning of the same course. The data were aggregated by score and the percentage of students who received each score compared across quarters. The correlation indicates how similar the score frequencies are across quarters, where a 0 represents no correlation and a 1 represents a perfect correlation. Adams and Wieman report Pearson correlations are expected to be above 0.9 [39]; we see that all courses are near or exceed this benchmark.

TABLE III. Inventory statistics for v3.0. Pearson correlations were calculated between all combinations of quarters and averaged; uncertainty represents the standard deviation across these combinations. Cronbach's α is calculated in aggregate across quarters; uncertainty represents the width of the 95% confidence range. Ferguson's δ was calculated across all quarters.

Course	N	Average Pearson correlation coefficient	Average Cronbach's α	Average Ferguson's δ
PreMech	746	0.95	0.67 ± 0.05	0.95 ± 0.01
PostMech	573	0.91	0.69 ± 0.06	0.96 ± 0.01
PostEM	293	0.89	0.73 ± 0.07	0.95 ± 0.01

Cronbach's α is an indicator of internal consistency. A high score indicates that item scores are related to one another. In general, a Cronbach's α of 0.7 or above is recommended; each course meets this benchmark within a 95% confidence range.

Ferguson's δ can be used as a measure of test-wide discrimination. A high Ferguson's δ indicates that there is a wide spread of students across all score options. For example, Ferguson's δ is 1 if an equal proportion of students earn each possible score; it is 0 if all of the students earn the same score. It is distinct from standard deviation in that Ferguson's δ provides a measure of how students are distributed across the possible scores, regardless of the mean. There is speculation as to whether Ferguson's δ is a good measure of discrimination because it is population dependent [54], but we include it here as all of the data presented in this paper are reflective of the student population at the main institution. We consider validating the GERQN at other institutions an area of future work.

Prior research found that while the PIQL was built on three facets of physics quantitative literacy (reasoning about sign, proportional reasoning, and covariational reasoning), exploratory factor analysis found it to be a single-factor assessment. This finding provided evidence that PQL is a way of reasoning [6]. We confirmed this result with GERQN v3.0 through exploratory factor analysis. We used the Kaiser-Meyer-Olkin (KMO) criterion to determine whether the data were suitable for factor analysis. We found a KMO value of 0.887, which suggests that each item is sufficiently correlated with the others such that one or more factors can be extracted [55]. Similarly, we found the Bartlett's test of sphericity was significant with an α value of 0.05 ($\chi^2(120) = 5697.75,\ p < 0.001$). We used the Kaiser-Guttman criterion and found that 1 or 4 factors could be extracted. We modeled the data using 1 and 4 factors. For the single-factor model, we found a confirmatory fit index of 0.94 (RMSEA = 0.07), which aligns with recommended cut-off values [56–59]. While the four factor model also had a high confirmatory fix index (CFI = 0.98, RMSEA = 0.04), the factor groupings did not reasonably align with any theoretical, expert perspective. We also performed confirmatory factor analysis using the expert determined factors (reasoning about sign, proportional reasoning, and covariational reasoning) and found a lower confirmatory fit index of 0.87 (RMSEA = 0.04). These statistics suggest that the test is well described by a single factor, confirming our interpretation: for students at the introductory, algebra-based physics level, the facets of PQL are not clearly separable. It is also possible that there are underlying features of reasoning that are not aligned with current interpretations of the reasoning required for these items; we leave this investigation for future research.

C. Item analysis

We examined individual items during student and expert interviews, as well as using classical test theory, to validate that the items appear appropriate to expert instructors, are understood by students, are of the appropriate level, and have discriminatory power.

We calculated the classical test theory difficulty and discrimination of each item for administrations of v1.6 and v3.0, seeking to meet the same standard as established on the PIQL [6]. We aim for a wide spread in difficulties between 0.2 and 0.8, and discrimination values above 0.3. Here, the word "difficulty" is a misnomer and refers to the fraction of students who choose the correct answer on a particular item; "discrimination" is a measure of how correlated choosing the correct answer on that item is with the overall test score (point biserial correlation). Discrimination is high when students who choose the correct answer are also likely to score high on the assessment as a whole; it is low when the correlation is not as strong. The results for v1.6 and 3.0 across the introductory sequence are shown in Figs. 5 and 6. We used these quantitative measures for v1.6 and v3.0 to inform final decisions about items to keep and items to remove.

Student interviews were conducted to establish evidence that:

- the students interpreted the questions and answer choices as intended, and that
- there were no commonly desired answers that were not already multiple choice options.

We also sought to confirm prior characterization of student reasoning associated with incorrect answer choices, toward helping instructors interpret the ways that their students might be reasoning. Across all interviews, students understood the items as intended and chose answers for the reasons we expected. Only one item was revised as a result of these interviews; the change was made halfway through the interviews, and confirmed with the second half.

Several experts had item-specific suggestions to improve readability. They also provided feedback on which items would be too difficult or confusing for their students, so that they did not expect valuable data on those items if the inventory were to be administered in their classes. This feedback led to some revisions of some items and the

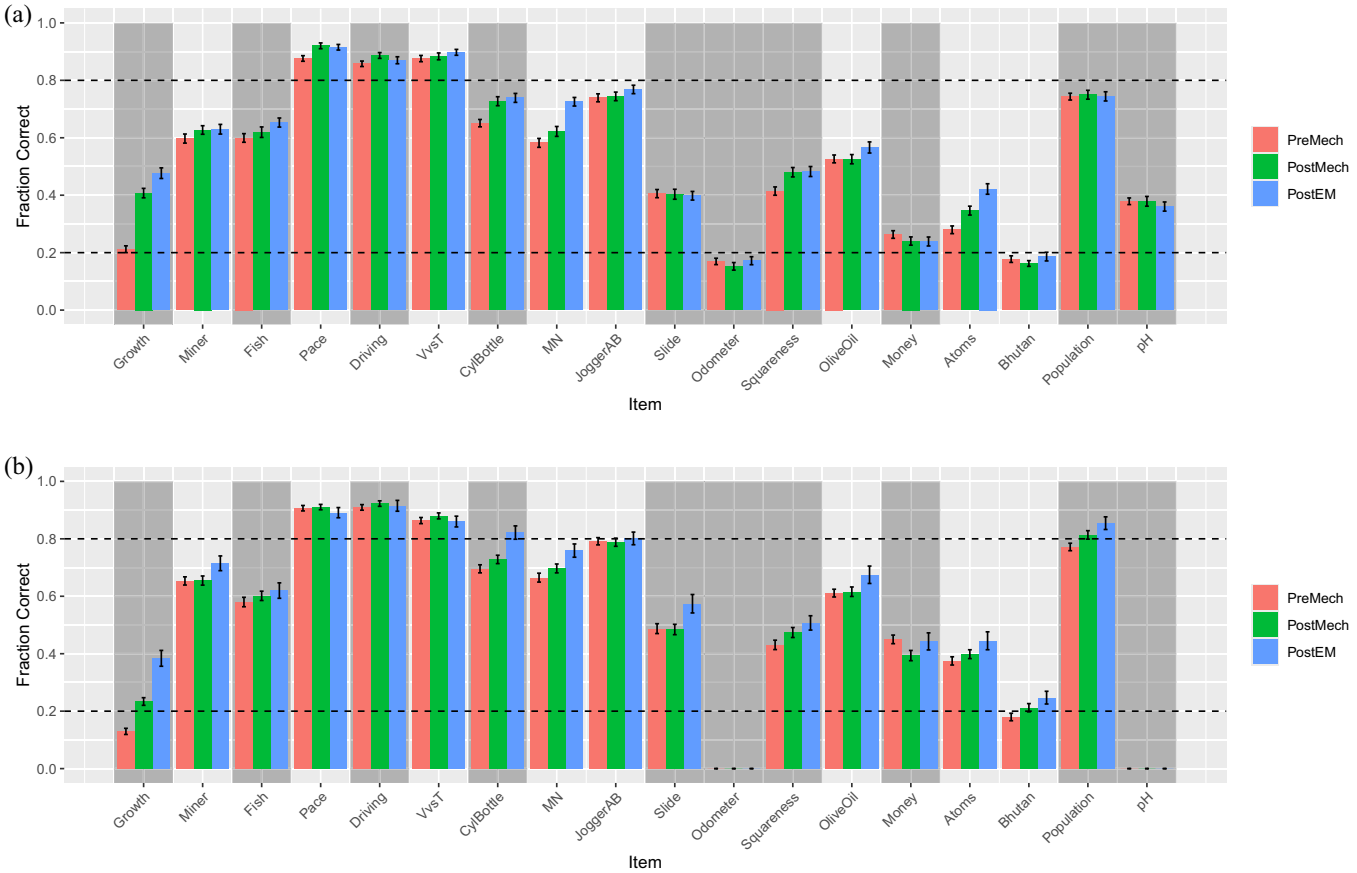


FIG. 5. Difficulty measures for each item on (a) version v1.6 and (b) v3.0 of the GERQN. Here, difficulty means the fraction of students who answered each item completely correctly. The horizontal lines represent the ideal range for inventory items. These data were collected over six quarters at the main institution ($N = 2778$ for v1.6 and $N = 1612$ for v3.0). Gray highlight indicates items that were changed between versions. Odometer and pH were removed in v3.0.

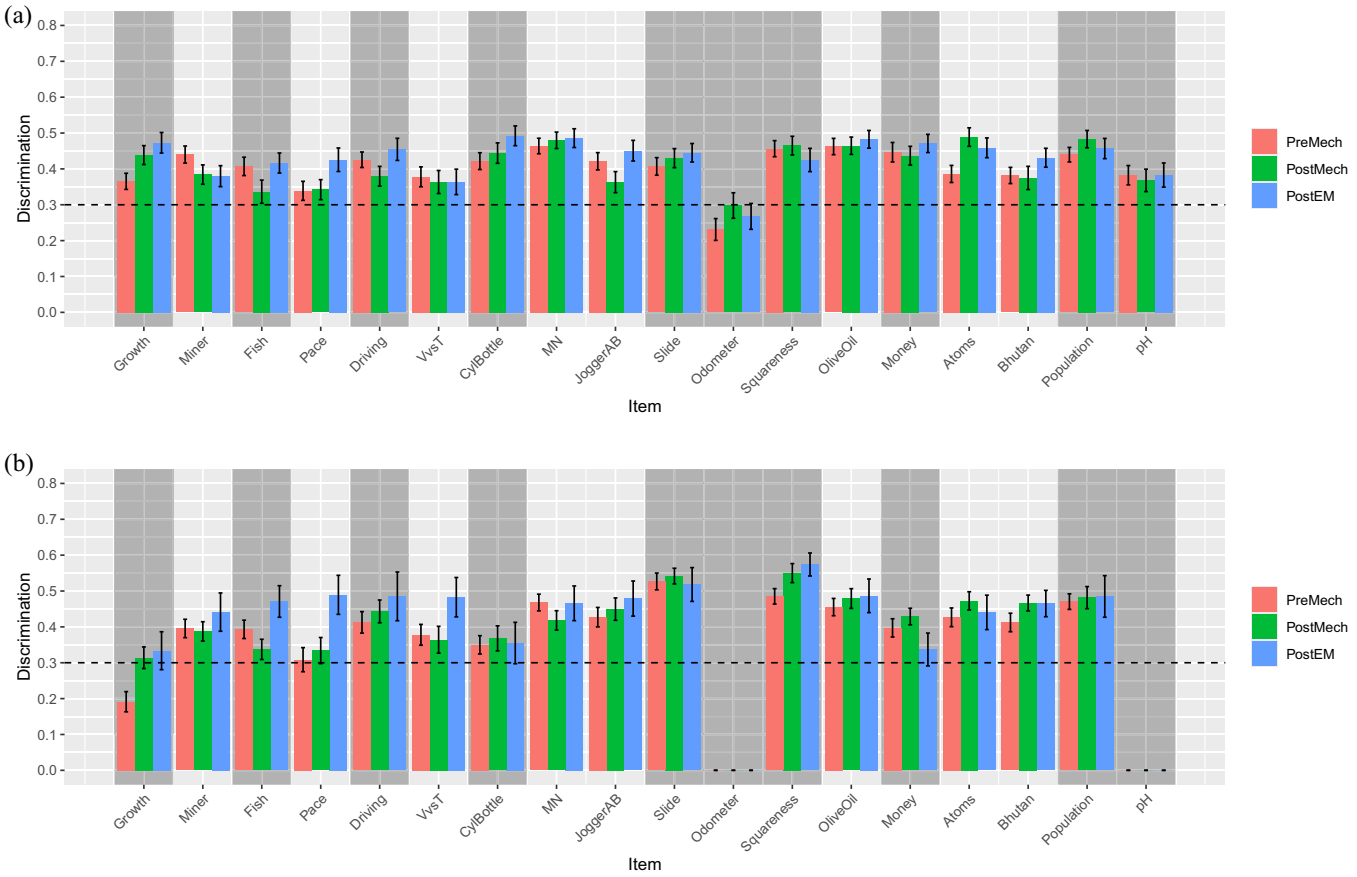


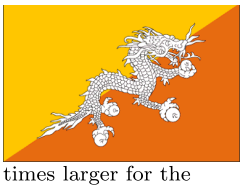
FIG. 6. Discrimination measures for each item on version (a) 1.6 and (b) 3.0 of the GERQN. The horizontal line represents the recommended minimum. These data were collected over six quarters at the main institution ($N = 2778$ for v1.6 and $N = 1612$ for v3.0). Gray highlight indicates items that were changed between versions. Odometer and pH were removed in v3.0.

removal of items that were considered too difficult. For example, in version 2.0, the "Growth" item curves both intersected the origin (Fig. 3). The mathematics education researcher whom we interviewed noted that it would be a more valuable item if the curves had a nonzero intercept. The change in difficulty for this item between versions (see Fig. 5) suggests that students may have been answering this item correctly for the wrong reasons before we made the change. Other items were revised for language to clarify assumptions of the simple models presented; experts suggested that students may find simplicity unnecessarily distracting. Their suggestions were confirmed with student interviews. Finally, most of the experts we interviewed agreed that one item ("Odometer") was too hard, and it was removed; this is supported by student data, with fewer than 20% of students answering "Odometer" correctly on v1.6 (Fig. 5), and "Odometer" being the only item to consistently have a discrimination value below 0.3 (Fig. 6). Another item ("pH") was removed because it did not represent essential reasoning.

A subset of experts also noted that a second item, "Bhutan," would likely be too hard for their students [Fig. 7(a)]. This was confirmed with test statistics (Fig. 5). However, Bhutan is a multiple-choice-multiple-response (MCMR) item. Figure 5 only shows the fraction of completely correct responses for MCMR items; Fig. 7(b) suggests that many more students are choosing a subset of correct answers. In addition, among the experts who had concerns about Bhutan, the concern was that answer option (C) was too challenging. All the experts agreed it would be valuable to them if their students selected more correct answer choices and fewer incorrect answer choices, even if the students did not get the item completely correct. They also agreed that students' ability to reason about scaling in quantities that are not well described by a familiar equation (something that is measured by answer choice C) is a valued learning outcome of introductory physics for the target population. Therefore, we decided to keep this item in the inventory.

We chose to keep several items that have a higher than recommended fraction of students who choose the correct answer (in the language of classical test theory, difficulty >0.8). The students at the main institution typically have more access to prior mathematics instruction than the target population for the GERQN as a whole. We consider these items to be a positive feature of the test, and we expect the item statistics to be different for data collected in other educational settings.

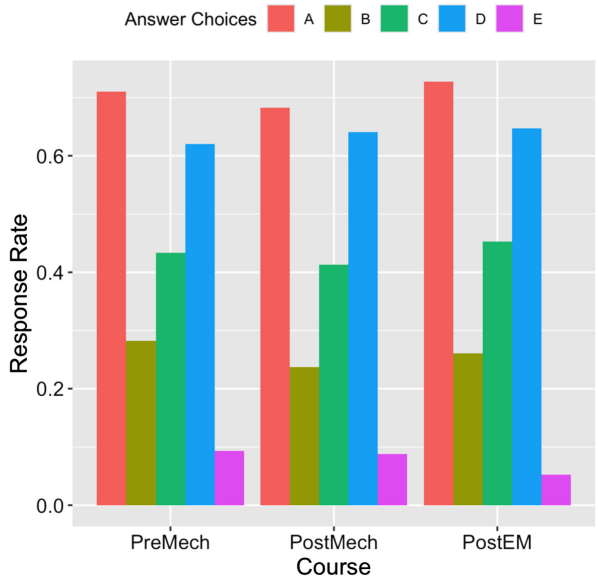
A person has a flag of Bhutan, shown at right. Another person has a larger flag of Bhutan. The larger flag is 2 times wider and 2 times taller than the smaller flag.



Which of the quantities below are 2 times larger for the larger flag compared to the smaller flag? *Choose all that apply.*

a. The distance around the edge of the flag.
b. The amount of cloth needed to make the flag.
c. The length of the curve forming the dragon's backbone.
d. The diagonal of the flag.
e. None of these quantities.

(a)



(b)

FIG. 7. (a) The Bhutan item, and (b) associated rates of student responses to each answer choice, treated independently. Bhutan did not change between versions; this plot includes data from both v1.6 and 3.0 ($N = 4231$).

VI. IMPLICATIONS FOR INSTRUCTION

One key goal in developing and validating the GERQN was to provide instructors with a practical tool to support PQL instruction. The GERQN can help identify meaningful learning objectives, track shifts in student performance over time, and highlight changes in specific skills through item-by-item analysis. Instructors can use it to better understand how their students reason by:

1. Monitoring average scores across cohorts or instructional periods.
2. Analyzing individual items to identify specific areas of difficulty.

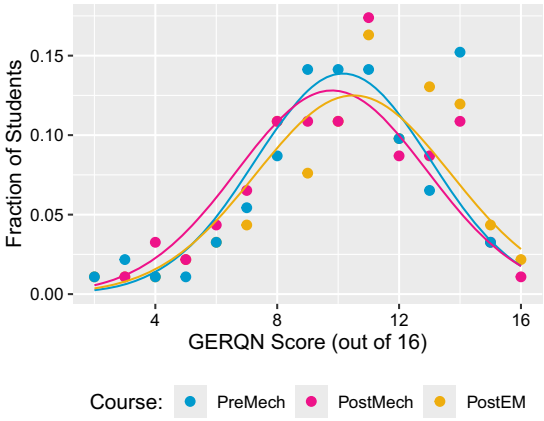


FIG. 8. GERQN score distribution over the yearlong introductory algebra course at the main institution. Only students who spent longer than 10 min and completed the test in each course are shown; these data are from v3.0. Data are fitted with a normal distribution ($N = 92$).

Initial findings at the main institution echo results from the PIQL: PQL does not significantly improve through traditional physics instruction alone (Fig. 8). The GERQN thus offers a valuable tool to help instructors address these persistent challenges within their own classrooms.

Aligning early PQL development with students' recent math coursework can boost student confidence and sense of agency in applying mathematical reasoning in future courses. This is especially important for students from lower socioeconomic status districts, where access to physics and calculus in high school is often limited [60]. Algebra-based physics is especially common as a first physics course in high school. Precollege teachers—experienced in designing creative, student-centered learning—could use the GERQN to develop engaging approaches to PQL instruction.

During validation, several early-career college faculty expressed strong interest in supporting the reasoning assessed by the GERQN but also highlighted the need for professional development to deepen their own understanding of PQL. Precollege teachers would also benefit from this type of dedicated professional development. The GERQN has the potential to serve as a foundation for such efforts. Initially validated with algebra-based physics students at an R1 institution, it is now being tested across a range of postsecondary contexts, including two-year colleges, minority-serving institutions, and rural campuses.

Future efforts will focus on validating the GERQN in precollege classrooms, creating professional development resources, and designing replicable workshops for both college and K-12 educators. Of course, no single assessment can fully address the longstanding challenges students face in developing PQL. We view the GERQN as a catalyst for informed instruction that can help move the needle on this way of reasoning.

ACKNOWLEDGMENTS

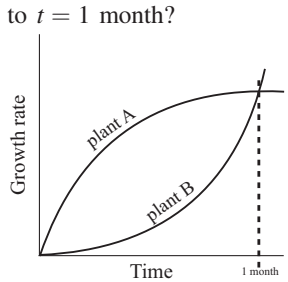
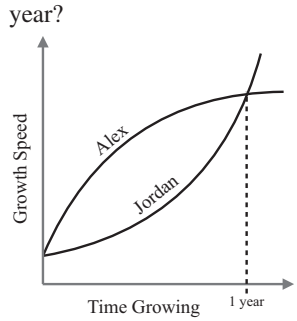
The authors thank Andrew Boudreaux and Stephen Kanim for their intellectual contributions to this work developing the framework and assessment items, Michael Loverude for contributions to our investigations of covariational reasoning, and those who participated in interviews for their thoughtful feedback. We would also like to thank our project advisory board members: Mehri Fadavi, Cameron Byerly, and Elizabeth Schoene. We are grateful to the contributions and support of the University of Washington Physics Education Group and the leadership of the introductory physics courses—David Smith, Nikolai Tolich, Kazumi Tolich, and Peter Shaffer in particular for making data collection possible over several years. This work was supported by the National Science Foundation under Grants No. DUE-2214283 and No. DGE-1762114.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available upon publication because it is not technically feasible and/or the cost of preparing, depositing, and hosting the data would be prohibitive within the terms of this research project. The data are available from the authors upon reasonable request.

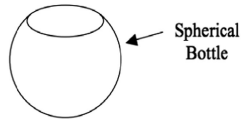
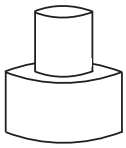
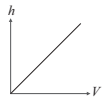
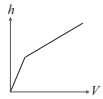
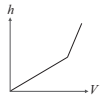
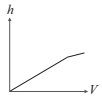
APPENDIX

Here, we provide a comparison of a sample of items in the PIQL (available on PhysPort) and GERQN v3.0 for the interested reader.

Item and summary	PIQL [6]	GERQN v3.0
Growth Context was adjusted, and "rate" was changed to "speed." The functions were also moved off the origin.	The graph shown represents the *growth rate* vs *time* for two plants. Which of the following statements best describes the growth of the two plants from $t = 0$ to $t = 1$ month?  a. Plants A and B have the same amount of growth. b. Plant A has experienced more growth than plant B. c. Plant B has experienced more growth than plant A. d. The graph does not provide enough information to compare the growth of the two plants.	The graph at right represents how fast two children are growing vs time. The children are named Alex and Jordan, and their growth is measured starting on their 10th birthday. Which of the following choices best describes how much the children have grown in one year?  a. Alex and Jordan have grown the same amount. b. Alex has grown more than Jordan. c. Jordan has grown more than Alex. d. The graph does not provide enough information to compare how much the two children have grown.

(Table continued)

(Continued)

Item and summary	PIQL [6]	GERQN v3.0
Bottle The spherical bottle was changed to a cylindrical one with two sections of different diameter, effectively linearizing the problem.	Assume that water is poured into a spherical bottle at a constant rate. Which of the following graphs best represents the height of the water, h, in the spherical bottle as a function of the amount of water in the bottle, V?  a. b. c. d. e. 	Water is poured into an empty bottle until it is full. The bottle is shaped like two cylinders, as shown at right. Which graph best represents the height of the water in the bottle, h, as a function of the amount of water in the bottle, V?  a.  b.  c.  d.  e.
Fish Function adjusted to be at the appropriate mathematical level. Question stem includes more description to account for the discrete nature of the new function.	The Wildlife Game Commission released 500 fish into a lake. The function $N(t)$ defined by $N(t) = \frac{600t+500}{0.5t+1}$ represents the approximate number of fish in the lake as a function of time (in years). Which one of the following best describes how the number of fish in the lake changes over time?	500 fish are released into a lake. The number of fish in the lake is recorded at the end of each year. In the first year, the number of fish gets smaller. After the first year, $N(t)$ represents the number of fish recorded, and t represents the number of years since the 500 fish were originally released into the lake. The equation for $N(t)$ is $N(t) = 500 - (30/t)$, starting at $t = 1$. Which of the following choices best describes the trend in the fish population after the first year, over a period of many years?

(Table continued)

(Continued)

Item and summary	PIQL [6]	GERQN v3.0
	a. The number of fish gets larger each year but does not exceed 500. b. The number of fish gets larger each year but does not exceed 1200. c. The number of fish gets smaller each year but does not get smaller than 500. d. The number of fish gets larger each year but does not exceed 600. e. The number of fish gets smaller each year but does not get smaller than 1200.	a. The number of fish keeps getting smaller until the fish are gone. b. The number of fish keeps getting smaller but does not drop below 470. c. The number of fish eventually grows to nearly 500 again. d. The number of fish eventually grows to a number greater than 500.
Inverse G/pace Replaced a changing rate of change with a constant rate of change.	Near the surface of Earth, the acceleration due to gravity is 9.8 $\mathrm{m/s^2}$ (note: $\frac{\mathrm{m}}{\mathrm{s^2}} = \frac{\mathrm{m/s}}{\mathrm{s}}$). For objects in vertical free fall near the surface of Earth, the number 9.8 provides the following information: *the speed of the object will change by* 9.8 m/s *during each second of its motion.* Consider the reciprocal of this number, 0.1 $\mathrm{s^2/m}$ (note: $\frac{\mathrm{s^2}}{\mathrm{m}} = \frac{\mathrm{s}}{\mathrm{m/s}}$). For objects in vertical free fall near the surface of Earth, what specific information does this quantity (0.1 $\mathrm{s^2/m}$) convey? Select the single best choice below.	A person runs at 3 m/s, meaning that the runner moves 3 m each second they are running. The reciprocal of 3 m/s is 0.33 s/m. Which of the following choices best describes the specific information that 0.33 s/m tells us about the motion of the person?
	a. The speed of the object will change by 0.1 m/s during each second. b. The motion of the object will change by 0.1 $\mathrm{s^2}$ in each meter. c. It takes 0.1 s for each 1 m/s change in the object's speed. d. It takes 0.1 $\mathrm{s^2}$ for the object to fall each meter. e. None of these make sense; the number 0.1 does not have a valid interpretation in this context.	a. It takes the runner 0.33 s to move 1 m. b. It takes the runner 1 s to move 0.33 m. c. It takes the runner 0.33 s to move 0.33 m. d. The runner's speed is 0.33 s/m. e. None of these make sense; the quantity 0.33 s/m does not have a specific meaning.
Jogger Numerals and language simplified; reasoning statements removed from the answers.	Joggers A and B start running at the same time from the same location. Jogger A is slower than jogger B (0.6 times the speed of jogger B) but runs for a longer time (1.5 times the amount of time that jogger B runs). How does the distance traveled by A compare to the distance traveled by B? Select the answer with the best reasoning.	Joggers A and B start running at the same time from the same location. Jogger A is slower than jogger B (0.6 times the speed of jogger B) but runs for twice as much time as jogger B. How does the distance traveled by A compare to the distance traveled by B?

(Table continued)

(Continued)

Item and summary	PIQL [6]	GERQN v3.0
	a. The distance traveled by A is greater than B because A runs for more time. b. The distance traveled by B is greater than A because B runs faster. c. They both run the same distance because although A runs for more time, B runs faster and it balances out. d. The distance traveled by A is greater because although B runs faster, A runs long enough that he passes B and keeps going once B has stopped. e. The distance traveled by B is greater because although A runs for more time, A doesn't run long enough to travel as much distance as B traveled before she stopped.	a. The distance traveled by A is greater than B. b. The distance traveled by B is greater than A. c. They both run the same distance. d. There's no way to tell without knowing their speeds.
Internal energy/money Context changed to money, variables are more explicitly defined, and fewer answer options are provided.	The internal energy of a system can be increased by doing positive work on the system or by heating it, and it can be decreased by cooling the system or if the system does work. Which of the following equations represent(s) this relationship (U is the internal energy of the system, Q is positive when energy flows into the system, and$W$ is positive when positive work is done on the system)? *Choose all that apply.* a. $\Delta U = Q - W$ b. $\Delta U = -Q + W$ c. $\Delta U = Q + W$ d. $-\Delta U = Q + W$ e. $-\Delta U = Q - W$ f. $-\Delta U = -Q + W$	ΔM represents the change in the amount of money that is in your wallet: • The value of ΔM is greater than zero when you receive money. • The value of ΔM is less than zero when you spend money. J represents the money paid for doing a job: • The value of J is greater than zero if you are paid to do a job. • The value of J is less than zero if you pay for someone else to do a job. G represents the money traded for goods: • The value of G is greater than zero if you sell something. • The value of G is less than zero when you buy something. Which of the following equations fully represent(s) the relationship ΔM, J, G if there is one job done and one trade made? *Choose all that apply.* a. $\Delta M = J + G$ b . $\Delta M = J - G$ c. $\Delta M = -J + G$ d. $\Delta M = -J - G$

[1] O. Uhden, R. Karam, M. Pietrocola, and G. Pospiech, Modelling mathematical reasoning in physics education, Sci. Educ. **21**, 485 (2012).

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

[3] J. D. Gifford and N. D. Finkelstein, Categorical framework for mathematical sense making in physics, Phys. Rev. Phys. Educ. Res. **16**, 020121 (2020).

[4] J. A. Czocher and H. L. Hardison, Attending to quantities through the modelling space, in *Mathematical Modelling Education in East and West*, edited by F. K. S. Leung, G. A. Stillman, G. Kaiser, and K. L. Wong (Springer International Publishing, Cham, 2021), pp. 263–272.

[5] P. W. Thompson, Quantitative reasoning and mathematical modeling, in *New Perspectives and Directions for Collaborative Research in Mathematical Education*, edited by L. L. Hatfield, S. Chamberlain, and S. Belbase (University of Wyoming, Laramie, WY, 2011), Vol. 1, pp. 33–57.

[6] S. White Brahmia, A. Olsho, T. I. Smith, A. Boudreaux, P. Eaton, and C. Zimmerman, Physics inventory of quantitative literacy: A tool for assessing mathematical reasoning in introductory physics, Phys. Rev. Phys. Educ. Res. **17**, 020129 (2021).

[7] B. Ojose, Mathematics literacy: Are we able to put the mathematics we learn into everyday use?, J. Math. Educ. **4**, 89 (2011), https://journalofmathed.scholasticahq.com/article/90403-mathematics-literacy-are-we-able-to-put-the-mathematics-we-learn-into-everyday-use.

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

[9] D. Hu and N. S. Rebello, Using conceptual blending to describe how students use mathematical integrals in physics, Phys. Rev. ST Phys. Educ. Res. **9**, 020118 (2013).

[10] 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).

[11] 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).

[12] J. Vlassis, The role of mathematical symbols in the development of number conceptualization: The case of the minus sign, Philos. Psychol. **21**, 555 (2008).

[13] R. R. Bajracharya, V. L. Sealey, and J. R. Thompson, Student understanding of the sign of negative definite integrals in mathematics and physics, Int. J. Res. Undergrad. Math. Educ. **9**, 62 (2023).

[14] 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).

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

[16] AP Physcs 1: Algebra-Based: Course and Exam Description (College Board, New York, NY, 2024), https://apstudents.collegeboard.org/courses/ap-physics-1-algebra-based/assessment.

[17] E. Mazur, *Principles & Practice of Physics* (Pearson, Boston, 2015).

[18] R. D. Knight, B. Jones, and S. Field, *College Physics: A Strategic Approach*, 4th ed. (Pearson, Boston, 2019).

[19] 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).

[20] P. W. Thompson and M. P. Carlson, Variation, covariation, and functions: Foundational ways of thinking mathematically, in *Compendium for Research in Mathematics Education*, edited by J. Cai (National Council of Teachers of Mathematics, Reston, VA, 2017), pp. 421–456.

[21] S. R. Jones, Multivariation and students' multivariational reasoning, J. Math. Behav. **67**, 100991 (2022).

[22] A. B. Arons, Cultivating the capacity for formal reasoning: Objectives and procedures in an introductory physical science course, Am. J. Phys. **44**, 834 (1976).

[23] D. E. Trowbridge and L. C. McDermott, Investigation of student understanding of the concept of acceleration in one dimension, Am. J. Phys. **49**, 242 (1981).

[24] J. J. Bissell, A. Ali, and B. J. Postle, Illustrating dimensionless scaling with Hooke's law, Phys. Educ. **57**, 023008 (2022).

[25] C. Zimmerman, A. Olsho, M. Loverude, and S. White Brahmia, Expert covariational reasoning resources in physics graphing tasks, arXiv:2306.00921.

[26] P. Emigh, R. R. Siegel, J. W. Alfson, and E. Gire, Student reasoning about multivariable covariation in thermodynamics, in *Proceedings of the Physics Education Research Conference 2019, Provo, UT* (American Association of Physics Teachers, 2019).

[27] S. Van den Eynde, P. van Kampen, W. Van Dooren, and M. De Cock, Translating between graphs and equations: The influence of context, direction of translation, and function type, Phys. Rev. Phys. Educ. Res. **15**, 020113 (2019).

[28] N. Altindis, K. A. Bowe, B. Couch, C. F. Bauer, and M. L. Aikens, Exploring the role of disciplinary knowledge in students' covariational reasoning during graphical interpretation, Int. J. STEM Educ. **11**, 32 (2024).

[29] A. Olsho, C. Zimmerman, A. Boudreaux, T. I. Smith, P. Eaton, and S. White Brahmia, Characterizing covariational reasoning in physics modeling, in *Proceedings of the Physics Education Research Conference 2022, Grand Rapids, MI* (American Association of Physics Teachers, 2022), pp. 335–340.

[30] A. H. Akatugba and J. Wallace, Mathematical dimensions of students' use of proportional reasoning in high school physics, Sch. Sci. Math. **99**, 31 (1999).

[31] D. P. Maloney, C. Hieggelke, and S. Kanim, nTIPERs: Tasks to help students "unpack" aspects of newtonian mechanics, in *Proceedings of the Physics Education Research Conference 2010, PER Conference Invited Paper, Portland, OR* (American Association of Physics Teachers, 2010), Vol. 1289, pp. 33–36.

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

[33] A. Boudreaux, S. 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 2020, Virtual Conference* (American Association of Physics Teachers, 2020), pp. 45–50.

[34] E. F. Redish and E. Kuo, Language of physics, language of math: Disciplinary culture and dynamic epistemology, Sci. Educ. **24**, 561 (2015).

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

[36] A. Olsho, S. White Brahmia, A. Boudreaux, and T. Smith, The physics inventory of quantitative reasoning: Assessing student reasoning about sign, in *Proceedings of the 22nd Annual Conference on Research in Undergraduate Mathematics Education*, edited by A. Weinberg, D. Moore-Russo, H.Soto, and M. Wawro (The Special Interest Group of the Mathematical Association of America (SIGMAA) for Research in Undergraduate Mathematics Education, Oklahoma City, OK, 2019).

[37] S. White Brahmia, Mathematization in introductory physics, Ph.D. thesis, Rutgers, The State University of New Jersey, 2014.

[38] S. White Brahmia, A. Olsho, A. Boudreaux, T. I. Smith, and C. Zimmerman, A conceptual blend of physics quantitative literacy reasoning inventory items, in *Proceedings of the 23rd Annual Conference on Research in Undergraduate Mathematics Education*, edited by S. Smith Karunakaran, Z. Reed, and A. Higgins (The Special Interest Group of the Mathematical Association of America (SIGMAA) for Research in Undergraduate Mathematics Education, Boston, MA, 2020), pp. 862–867.

[39] 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 (2011).

[40] H. Zhang, A. Perry, and I. Lee, Developing and validating the artificial intelligence literacy concept inventory: An instrument to assess artificial intelligence literacy among middle school students, Int. J. Artif. Intell. Educ. **35**, 398 (2024).

[41] P. H. Santoso, E. Istiyono, Haryanto, and H. Retnawati, Validating light phenomena conceptual assessment through the lens of classical test theory and item response theory frameworks, Phys. Educ. **59**, 025002 (2024).

[42] E. A. Holt, J. Duke, R. Dunk, and K. Hinerman, Development of the inventory of biotic climate literacy (IBCL), Environ. Educ. Res. **30**, 2210 (2024).

[43] D. McKee and G. Orlov, The economic statistics skills assessment (ESSA), Int. Rev. Econ. Educ. **44**, 100272 (2023).

[44] A. J. Cetnar, A. Besemer, V. Bry, C. R. Buckey, J. Burmeister, A. Rodrigues, L. Schubert, M. Speidel, S. Sutlief, and A. S. Yu, Introduction to concept inventories for medical physics education, J. Appl. Clin. Med. Phys. **24**, e14130 (2023).

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

[46] R. L. Doran, *Basic Measurement and Evaluation of Science Instruction* (National Science Teachers Association, Washington, DC, 1980).

[47] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 20103 (2009).

[48] T. A. Brown, *Confirmatory Factor Analysis for Applied Research* (Guilford Publications, New York, NY, 2014).

[49] J. C. Hayton, D. G. Allen, and V. Scarpello, Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis, Organ. Res. Meth. **7**, 191 (2004), https://journals.sagepub.com/doi/10.1177/1094428104263675.

[50] Otter.ai, https://otter.ai/ (2019).

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

[52] L. M. Hartley, J. Momsen, A. Maskiewicz, and C. D'Avanzo, Energy and matter: Differences in discourse in physical and biological sciences can be confusing for introductory biology students, BioScience **62**, 488 (2012).

[53] The same analysis was conducted with version 1.6, with the same result. Version 1.6 data shown in this paper are also filtered for students who took more than 10 min and left 5 items or fewer blank.

[54] B. Terluin, D. L. Knol, C. B. Terwee, and H. C. de Vet, Understanding Ferguson's δ: Time to say good-bye?, Health Qual. Life Outcomes **7**, 38 (2009).

[55] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 7th ed. (Pearson, Harlow, 2014).

[56] P. Eaton, B. Frank, K. Johnson, and S. Willoughby, Comparing exploratory factor models of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **15**, 020133 (2019).

[57] *Handbook of Structural Equation Modeling*, edited by R. H. Hoyle (The Guilford Press, New York, NY, 2023).

[58] T. A. Brown, Confirmatory factor analysis for applied research, in *Methodology in the Social Sciences* 2nd ed. (The Guilford Press, New York, London, 2015).

[59] N. K. Bowen and S. Guo, *Structural Equation Modeling, Pocket Guides to Social Work Research Methods* (Oxford University Press, New York, 2012).

[60] S. White Brahmia and G. L. Cochran, Underprepared for physics: Reframing the narrative on readiness and instruction in calculus-based, introductory physics courses, Phys. Today (to be published).