Physics & Astronomy Faculty Works   Physics & Astronomy

7-1-2023

Development And Reliability Analysis Of A Split-Administration Test Of The Math Epistemic Games Survey

Stephen Hackler
Swarthmore College, shackle1@swarthmore.edu

E. Elliott

M. Eichenlaub

See next page for additional authors

Recommended Citation

Authors

Stephen Hackler, E. Elliott, M. Eichenlaub, and A. M. Sweeney

Development and reliability analysis of a split-administration test of the math epistemic games survey

Stephen Hackler,[1,2,*] Emily Elliott[3] Mark Eichenlaub,[4,5,*] and Alison M. Sweeney[1,6,*]

[1]*Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*
[2]*Department of Physics, Swarthmore University, Swarthmore, Pennsylvania 19081, USA*
[3]*Center for Teaching and Learning, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*
[4]*Physics Education Research Group, University of Maryland, College Park, Maryland 20742-4111, USA*
[5]*Art of Problem Solving, Lewisburg, Pennsylvania 17831, USA*
[6]*Departments of Physics and Ecology & Evolutionary Biology, Yale University, New Haven, Connecticut 06520-8106, USA*

The increasing and diversifying student enrollments in introductory physics courses make reliable, valid, and usable instruments for measuring student skills and gains ever more important. In introductory physics, in addition to teaching facts about mechanics, we also seek to teach our students the skills of "thinking like a physicist," or expertise in and intuition for physical problem solving. How and when these expert, intuitive problem-solving skills emerge during a STEM education, or what the most effective teaching methods might be, are not certain. A facile survey to measure students' "physics-thinking" skills in a pretest and post-test format is therefore desirable to measure and evaluate different pedagogical approaches. Prior investigators codified these skills as "epistemic games" (e.g., order-of-magnitude estimation, evaluating extreme cases) and developed and validated the math epistemic games survey (MEGS) to measure students' ability to employ these techniques. The original survey instrument is reliable and valid but has drawbacks in its length and in students' ability to recall questions between administrations. We employed factor analysis to split the MEGS into two mutually exclusive subtests and measured them to be equivalently reliable and valid as the full-length MEGS as originally formulated. The "split MEGS" is well suited for use as a pretest and post-test instrument to measure gains in expertise in problem solving in introductory physics courses.

I. INTRODUCTION

A. Epistemic games

Perhaps the most widely used physics education instrument is the Force Concept Inventory (FCI), which seeks to "help teachers probe and assess the commonsense beliefs of their students" [1] by asking a series of questions about various topics in Newtonian mechanics. Concepts within the FCI comprise six aspects of the Newtonian understanding of force, each of which is probed by a variety of questions. The FCI measures students' ability to recognize physics "facts" such as the third law and the definition of acceleration. Any physics teacher is all too aware that there

is a durable disconnect between students' ability to understand and recall physics facts (e.g., "acceleration is the rate of change in velocity with respect to time") and to solve physics problems (e.g., "find the time to collision of two trains of known velocity if one slams on the brakes to accelerate at $-1~\mathrm{m\,s^{-2}}$"). While the FCI and similar instruments are crucial for measuring conceptual gains in teaching and learning, there are comparatively few instruments that measure and quantify gains in physics problem-solving ability.

Previous research into student physics learning suggests that a critical element of gaining facility in problem solving is learning to effectively use a particular suite of analytical strategies and skills [2,3]. The language around these strategies varies between authors, as does the method by which they are established. Tuminaro and Redish documented the problem-solving strategies used by life science students in introductory physics courses and categorized their approaches into six "epistemic games." These games are (a) mapping meaning to mathematics, (b) mapping mathematics to meaning, (c) physical mechanism game,

---

(d) pictorial analysis, (e) recursive plug-and-chug, and (f) transliteration to mathematics [2]. Similarly, Black and Wittmann observed two intermediate mechanics students solving problems involving applying boundary conditions to find the solutions of differential equations and identified two different epistemic games that utilize similar underlying mathematical resources: finding a family of functions and fitting the physical situation [4]. Other research has described different epistemic games such as analytical derivation [5], answer making [6], and sensemaking [7]. Hu *et al.* examined the use of math in a professional physics lab and an industrial workplace and identified six professionally useful epistemic games: conceptual math modeling, analytical-numerical math modeling, design-oriented math modeling, fabrication, improving processes, and making meaning out of data games [8]. Not all research into problem-solving approaches in math and physics makes use of the phrase "epistemic games." Boudreaux *et al.* examined expert reasoning around ratio and proportion, identifying six "subskills": identify ratio as a useful measure where appropriate, interpret a ratio verbally, construct a ratio from measured values to characterize a physical process or system, apply a ratio to determine an unknown amount, translate between different ways of representing a proportional relationship, and scale a proportional relationship to analyze a physical process or system [3]. Brahmia *et al.* also used a combination of existing literature and data from introductory physics courses to identify three "facets" of student reasoning around the use and representation of mathematics in introductory physics courses: proportional reasoning, reasoning about signs, and covariational reasoning. They used these facets to design an instrument to measure quantitative literacy in introductory physics classes, the Physics Inventory of Quantitative Literacy (PIQL) [9].

The line of research on the process of inquiry has introduced the separate concepts of epistemic games (strategies used in building understanding) as well as epistemic forms (target structures that guide the inquiry process) [10], also using the related terms of subskills and facets [3,9] for similar but distinct concepts. In this work, we will use the term epistemic games to refer to strategies that one might use to solve a problem, and epistemic forms to refer to structures that organize information that must be either interpreted or completed to solve a problem. It is also important to note that the method of identifying particular games and strategies meaningfully impacts the strategies that were ultimately identified (as one might expect). It is not surprising that the epistemic games identified by Tuminaro and Redish that were observed in life sciences students taking introductory physics [2] differ from those identified by Hu *et al.* when studying the photonics industry [8]. The set of epistemic games one considers must be informed by the relevant population and the research question.

The set of four epistemic games that we consider here are those identified by Eichenlaub to be common games played by expert physicists in solving problems. These games are (a) considering extreme cases of a given problem, (b) dimensional or scaling analysis, (c) estimation of real-life quantities, and (d) mapping of variables in equations to physical concepts [11]. These games were employed in an instrument that measures students' use of sensemaking strategies, the math epistemic games survey (MEGS). In this work, we report on the results of MEGS administration from a large student population outside the original University of Maryland (UMD) context and also evaluate a strategy to use the original MEGS as a shorter, more tractable pretest and post-test instrument.

B. Development of the math epistemic games survey

Researchers in the Physics Education Research Group at the University of Maryland observed that students often struggle not with knowledge of appropriate concepts or details, but with identifying an appropriate strategy and identifying information pertinent to that strategy [11]. In short, students understood physics "facts" and could perform epistemic games, but in a more general context didn't know how to choose an epistemic game to solve a given problem. To the extent that these epistemic games are definable and quantifiable and that the ability to use them *ad hoc* is indicative of a student's overall progression toward expert-level physics problem solving [2], they represent a promising route for developing a survey instrument to quantify gains in problem-solving ability throughout a physics education. The four games considered by the MEGS were chosen because they are important parts of the analytical toolkit of an expert physicist but are rarely explicitly taught in introductory-level courses [11]. The MEGS is "a 30-question, multiple-choice concept inventory of mathematical questions set in the context of sensemaking, especially for physics for the life sciences" [11]. Life sciences students were chosen as the target sample due to the large size of introductory physics for life sciences students (both at UMD and many other institutions). In particular, the questions focused more on whether students could, upon encountering a problem, identify which epistemic game would be useful in arriving at a solution and correctly implement it. Therefore, questions did not prompt students to use a particular approach. For example, one question simply asks students "Approximately how many breaths does an average person take in their lifetime?" (measuring if students can estimate real-life quantities). Another asks students to consider how the number of nuclei in a slime mold scales with its radius (assessing dimensional or scaling analysis). The researchers administered the MEGS as a pretest and post-test to over 1500 students in introductory physics courses at three different institutions. See Fig. 1 for further examples of MEGS questions corresponding to each of the four epistemic games of interest.
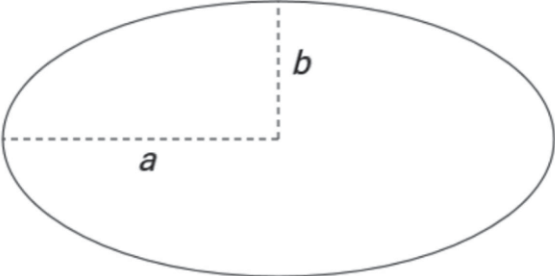
Estimation of Real-Life Quantities:	Dimensional/Scaling Analysis:
Q19. Which of these is closest to how fast the average person's hair grows? a. 5×10^{-11} cm/s b. 5×10^{-9} cm/s c. 5×10^{-7} cm/s d. 5×10^{-5} cm/s e. 5×10^{-3} cm/s	Q4. Which expression could represent the surface area of a solid object? Variables A, B, and C represent lengths, such as the length of the side of an object or the diameter of a circular object. a. $2(AB + A\sqrt{C^2 - A^2} + BC)$ b. $\sqrt{A^2 + B^2 + C^2}$ c. $\frac{\sqrt{2}}{2}A^2B$ d. $\frac{3AC}{2B}$ e. None of these could be a surface area
Considering Extreme Cases: Q20. Which of these is the formula for the area of the ellipse? a. πa^2 b. πb^2 c. $\pi(a + b)$ d. πab e. $\pi(\frac{a+b}{2})^2$	**Mapping Variables to Physical Concepts:** Q22. You buy 0.26 pints of olive oil for two dollars at the farmer's market. You plan next week to buy P pints of olive oil. Which expression gives how much this will cost? a. $ (2 * P) / 0.26$ b. $ P / 0.26$ c. $ (2 * 0.26) / P$ d. $ (P * 0.26) / 2$ e. $ 0.26 / (2 * P)$

FIG. 1.　Sample questions from the math epistemic games survey, representative of the four epistemic games of interest.

It is worth noting that the epistemic games tested in the MEGS are not the same set identified in previous research on a similar population by Tuminaro and Redish [2]. The earlier set of games was identified by interviewing students about their approach to solving physics homework problems. The students in this study were enrolled in an algebra-based introductory physics course, 50% of whom were biological science majors [2]. Critically, the students making *ad hoc* use of the games in this earlier study were not expert physicists, so the selection of games that they

used does not represent expert-level problem solving. The MEGS explicitly seeks to assess students' use of specific epistemic games that are central to "thinking like a physicist," and these games are not typically employed by novice nonmajors. The MEGS is designed to assess student gains in problem-solving expertise obtained during introductory physics courses, therefore it is sensible that it considers games played by expert physicists.

While the epistemic games considered by the MEGS were intended to capture expert physics problem solving, the difficulty and framing of the individual questions were designed with UMD's introductory physics for life sciences course in mind. This course contains predominantly third-year students (though included some second and fourth years as well), all of whom had taken at least one semester of chemistry and two semesters of biology, but no previous college physics [11]. The MEGS intentionally uses a biological frame for its questions to increase student familiarity and reduce the unhelpful complexity of additional background information. If the MEGS approach for assessing gains in problem solving by life-sciences majors is successful, similar instruments could be designed to measure the use of the same four important games in different student populations (upper-level physics majors, chemistry majors, engineering majors, etc.).

C. Initial trials of the MEGS

Initial trials of the MEGS found that, in general, students score relatively poorly and do not, on average, improve after a semester of physics [11]. Similar to the FCI [1], there is a mismatch between instructors' sense of the difficulty of MEGS questions (instructors expect that their students will perform well) and student performance (students generally do not know how to approach the questions productively). Average scores are around 60%, before and after a year of physics. Given that the MEGS assesses strategies important to experts, this disconnect is perhaps unsurprising.

Researchers also identified challenges with administering the MEGS as a pretest and post-test. The full-length MEGS is too long (>30 min of active test-taking time) to allow ready use during synchronous class time. Its utility as a post-test is limited by students' ability to recall questions from a prior administration [11]. The MEGS is also a relatively difficult diagnostic instrument, and our data suggest that fatigue may decrease scores on the full-length test. During MEGS development, the idea of developing a shorter form was discussed to address some of these challenges [11].

D. Splitting the MEGS

In this work, we explore the division of the MEGS into two nonoverlapping subtests. We thereby halve the time needed to administer the exam (thus reducing student fatigue and class time cost) and eliminate the possibility of pretest question recollection on the post-test. This same approach has been applied to the FCI, with promising

results [12]. Our goals are to make the MEGS more tractable to administer and more conducive to generating reliable data about students' degree and acquisition of physics problem-solving expertise.

Diagnostic tools require assessments of reliability (are measurements internally consistent and stable across trials?) and validity (does the tool measure what it purports to measure?) [13]. So, dividing any instrument into halves requires that both resulting tests correctly assess the same metrics as the original test and that they do so with equivalent reliability to each other. There are a variety of standard ways to measure an instrument's performance, many of which were part of the original design of the MEGS, which included statistical analysis of initial student data to evaluate question difficulty, degree of discrimination, and internal consistency between survey questions [11]. Importantly, it is not critical that the original instrument and the resultant subtests be perfectly co-reliable. Indeed, we expect that student effort and motivation (both of which may impact the reliability of an instrument) are influenced by the length of the survey. What is important is whether the two subtests are co-reliable with respect to each other, which is essential for an instrument to effectively act as a pretest and post-test. We find that our proposed division of the MEGS is reliable and ameliorates student fatigue and pretest vs post-test recall, making the "split MEGS" an improved metric relative to the original formulation.

II. METHODS

A. Administration

We formulated the MEGS as a Qualtrics project, with the question order and wording exactly matching the most recently published version [11]. The test was formatted with one question per page (except in the case of multipart questions, for which all parts were on the same page) to allow recording of the time spent on each question.

With the lead instructor's permission, the full MEGS v1.1 was administered as published to students in a subset of introductory physics sections (PHYS101 and PHYS150) at the University of Pennsylvania. PHYS101 is predominantly taken by prehealth students who have extensive biology background and little or no college physics or calculus experience. PHYS150 is predominantly taken by engineering majors with more experience applying mathematics compared to PHYS101. Both courses cover approximately the same traditional Newtonian mechanics content and differ in the mathematical sophistication of the lectures and assignments. One might suspect that different levels of introductory physics would perform noticeably differently on the MEGS and that an optimal division of questions would vary between different courses. To devise two subtests that could be generally applicable to any introductory physics course, the data from both courses were aggregated. Administration took place in person, usually during the normal class period for the course,

and students had 45 min to complete the test. Depending on the instructor, 10% of students had the incentive of a dropped homework grade to complete the test, and 90% of students had no incentive. No identifying information about the participants was recorded or stored, and results from each section were aggregated separately.

In the subsequent semester, the two subtests that resulted from our analysis of the full-test results were administered in a similar manner to all students enrolled in introductory physics laboratory sections at the University of Pennsylvania (PHYS101, PHYS 102, PHYS 150, PHYS 151, and PHYS171). PHYS102 and PHYS152 are traditional electromagnetism courses that follow PHYS101 and PHYS151, respectively. PHYS171 is taken predominantly by physics majors and covers calculus-based Newtonian mechanics, emphasizing calculus to prepare students for upper-level physics courses. An equal number of students received test A as a pretest and test B as a post-test and the reverse. The pretest was taken by students ~3 weeks into the semester, minimizing the participation of students who did not complete the course, and the post-test was taken one week before the last day of classes. Students could choose to opt out of this study, and all identifying information was removed before subsequent analyses.

B. Factor analysis

The data from all sections were aggregated. This was done so that the eventual subtests would be broadly relevant to many different levels of introductory physics, rather than being specific to a particular population. These aggregate data were transformed into a binary array of incorrect or correct responses to each question for each student. Any question with greater than 95% or less than 5% accuracy across all students was discarded. We also filtered out any individual test that answered the decoy question incorrectly, which selected "I don't know" for more than 40% of questions, and/or that left more than 40% of questions blank. The number of responses for all administrations before and after filtering is shown in Table I. Two tests were performed to judge the data's fitness for factor analysis. First, Bartlett's test of sphericity compared the correlation matrix to the identity matrix, computing a measure of how related or unrelated the data are [14]. Second, a Kaiser-Meyer-Olkin measure of sampling adequacy was computed, evaluating what proportion of the data's variance

could be due to underlying factors [15]. Both measures supported the suitability of our dataset for factor analysis.

An exploratory factor analysis was then performed using the Python FactorAnalyzer package [16], using a promax rotation and principal factor extraction. MEGS questions were sorted by the factor that corresponded to their highest loading, considering only those factors with an eigenvalue greater than 1.0 [17]. This had the effect of grouping MEGS questions into distinct categories in which student performance on questions within the same category was well correlated. We used these categories identified in factor analysis to create two mutually exclusive subtests by dividing the questions sorted into a given factor as evenly as possible between the two subtests. The aim was to create two mutually exclusive tests on which a given student would score similarly. Investigator oversight was needed in the case of factors that contained an odd number of questions or in the case of multipart questions that were sorted into the same factor. To avoid any potential loss of validity, multipart questions from the original MEGS were kept together and placed on the same subtest. Finally, a few questions were sorted to maintain equivalence of subtest length, mean score, score variance, and time to complete between the two subtests. The decoy question was included in both subtests to maintain an indicator of student attention. Finally, question ordering was preserved between the original MEGS and the two subtests (i.e., if question A preceded question B on the full MEGS and both questions ended up on the same subtest, question A would precede question B on the subtest as well). The factor analysis of the full-MEGS data suggests that after this optimization procedure, the resulting subtests will assess a near-equivalent set of cognitive skills.

The data from the full-MEGS administration were then re-analyzed with respect to the two optimized subtests. This procedure resulted in seven sets of student data for reliability analysis: full MEGS (Full), test A questions from the full MEGS (Full-A), test B questions from the full MEGS (Full-B), test A used as a pretest (G1A), test B used as a pretest (G2B), test A used as a post-test (G2A), and test B used as a post-test (G2B). Aggregate statistics and reported metadata (self-reported effort, self-assessed accuracy, timing, etc.) were compared between the seven sets.

Cronbach's alpha, a measure of an instrument's internal consistency that is often used to assess reliability, was also

TABLE I. Test administration participation and response.

	Fall 2019 full administration	Spring 2020 subtest A as pretest	Spring 2020 subtest B as pretest	Spring 2020 subtest A as post-test	Spring 2020 subtest B as post-test
Before filtering	182	286	286	293	293
After filtering	177	261	260	260	272
Percent usable	97.3%	91.3%	90.9%	88.7%	92.8%

calculated for each of the seven datasets [18]. Kolmogorov-Smirnov goodness-of-fit tests (KS tests) were also performed between all six subtest datasets. KS tests are used to compare two different sets of values and determine how likely they were to have been sampled from the same underlying probability distribution [19]. Though it can be used to compare a measured dataset to a known distribution, we used KS tests exclusively to determine whether two subtest datasets were sampled from the same distribution as each other, rather than to determine the nature of the underlying distribution. KS tests have four relevant parameters: a significance level (α), a critical value (D_{crit}), a KS statistic (D), and a p value (p). KS statistics correspond to the chosen significance level (i.e., if the datasets are likely sampled from the same distribution, the KS statistic will be lower than the critical value). The expression for the critical value is

$$D_{\mathrm{crit}} = C(\alpha)\sqrt{\frac{N_A + N_B}{N_A N_B}}, \qquad (1)$$

where $C(\alpha) = 1.36$ for the significance level of $\alpha = 0.05$. The p value for the KS test is the likelihood that the two datasets are sampled from the same distribution. If the KS statistic is greater than the critical value or if the p value is less than the significance level, the datasets are likely sampled from different distributions.

III. RESULTS

The full administration of the MEGS produced 177 usable responses from 182 test submissions. The data were well suited for factor analysis: Barlett's test for the full MEGS administration yielded a p value of 6×10^{-25} and a KMO score of 0.699. The filtering for questions of excessively high or low accuracy removed question no. 16 (accuracy $= 3.39\%$), and question no. 25, which was the decoy question. After this filtering, the factor analysis
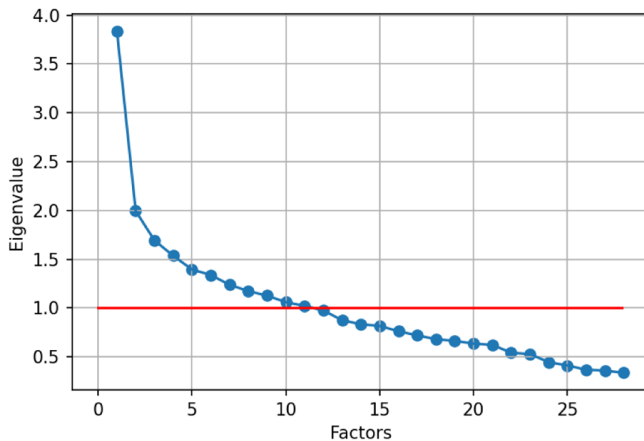

FIG. 2. Scree plot for the full MEGS administration. The red line indicates the chosen eigenvalue cutoff of 1.0.

revealed 10 factors with an eigenvalue greater than 1.0 (Fig. 2). These 10 factors accounted for 61.5% of the variance in the data. The 10 factors are shown in Table II, along with the question lists for the final two subtests.

TABLE II. Results of exploratory factor analysis and the final division of subtests A and B. Question numbers refer to the original full MEGS and the arrows indicate which subtest received which questions. The left and right columns show the two subtests in chronological order.

Subtest A		Subtest B
Q2[a]	Factor 1	
	$\Leftarrow$ Q17	Q1
	Q20[b] $\Rightarrow$	
	$\Leftarrow$ Q23[c]	
	$\Leftarrow$ Q24[c]	
Q3[a]	Factor 2	Q8
	$\Leftarrow$ Q4	
Q4	Q12 $\Rightarrow$	Q9
	$\Leftarrow$ Q27	
	Q28 $\Rightarrow$	
Q5	Factor 3	Q11
	Q13[d] $\Rightarrow$	
Q6	Q15[d] $\Rightarrow$	Q12
	Factor 4	
Q7	$\Leftarrow$ Q7	Q13[d]
	Q9 $\Rightarrow$	
Q10	Q14[d] $\Rightarrow$	Q14[d]
	$\Leftarrow$ Q29	
	Factor 5	
Q17	Q1 $\Rightarrow$	Q15[d]
	Factor 6	
Q22	$\Leftarrow$ Q6	
	Q19 $\Rightarrow$	Q18
	$\Leftarrow$ Q26	
Q23[c]	Factor 7	Q19
	$\Leftarrow$ Q2[a]	
Q24[c]	$\Leftarrow$ Q5	Q20[b]
	Q8 $\Rightarrow$	
Q25[e]	Q18 $\Rightarrow$	Q21[b]
	Factor 8	
Q26	$\Leftarrow$ Q10	
	Q30 $\Rightarrow$	Q25[e]
	Factor 9	
Q27	$\Leftarrow$ Q3[a]	Q28
	$\Leftarrow$ Q22	
Q29	Factor 10	
	Q11 $\Rightarrow$	Q30
	Q21[b] $\Rightarrow$	

[a]Multipart question about dye diffusion.
[b]Multipart question about an ellipse.
[c]Multipart question about student-professor ratios.
[d]Multipart question about precursor concentration.
[e]Decoy question.

TABLE III. Division of questions corresponding to four epistemic games between subtest A and subtest B.

	Subtest A	Subtest B
Extreme cases	Q5	Q12, Q18, Q20, Q28, Q30
Dimensional or scaling analysis	Q4, Q7	Q1, Q8, Q9, Q29
Estimation of real-life quantities	Q6, Q26	Q11, Q19
Mapping variables to physical concepts	Q2, Q3, Q10, Q17, Q22, Q23, Q24, Q27	Q13, Q14, Q15, Q21

Table III shows how the questions corresponding to each of the four epistemic games of interest were divided between the two subtests.

Examination of these factors showed that, in general, questions sorted into the same factor required the same problem-solving approach. For example, factor 6 in Table II contains the question "Estimate the thickness of a page in a typical textbook.", "Which of these is closest to how fast an average person's hair grows?", and "[Given a chart of some items and their monetary value per kilogram] Where would US \$100 bills fit on this chart?". All three of those questions require a student to approximate, to the nearest order of magnitude, some initially unknown quantity of something they have experienced. This skill (estimation of real-life quantities) was one of the four epistemic games that the MEGS was initially designed to assess. However, the fact that more than four factors were identified as significant suggests the existence of other aspects of student responses that are distinct from the four epistemic games incorporated in the design of the MEGS.

Take for example factor 2, which contains the following questions: "[Given a list of expressions in terms of lengths A and B] Which expression could represent the surface area of a solid object?"; "[Given an expression for the surface area of a cylinder] Which of these [provided expressions] would be the best approximation to the surface area of a long thin cylinder?"; "[Given a list of expressions in terms of your running speed and a moving sidewalk's speed] How fast would an observer standing on the ground next to the sidewalk see you moving?"; and "[Given the expression for the reduced mass of a two-body system, μ] If m_1 represents the mass of the earth and m_2 represents the mass of a small satellite, which of these [provided expressions] would be the best approximation for μ?". The last three questions can all be solved by considering the extreme cases of the system (one of the epistemic games around which the MEGS was designed). The first question is solved by unit analysis (only one of the possible answer choices has units of area), which closely aligns with the dimensional or scaling analysis epistemic game. However, all four of these questions involve some degree of correctly analyzing and simplifying abstract mathematical expressions, suggesting that perhaps the mathematical literacy needed to perform that task is significant in understanding student performance on the MEGS. We will discuss additional possible explanations for the formation of these additional factors in the following section of this paper.

The four tests from the split administration produced on average a 90% usable response rate (261/286 of G1A, 260/293 of G2A, 260/286 of G1B, and 272/293 of G2B were usable). Of data taken from the full administration, questions from subtest A had a mean score of 57.4%, and questions from subtest B had a mean score of 56.5%, both approximately six percentage points lower than the scores from the same questions answered during the split administration counterparts. Both groups taking subtests showed score gains of 1.3 percentage points between the pretest and post-test: G1A had a mean score of 62.5% as compared to G1B's mean score of 63.8% and G2B had a mean score of 62.9% as compared to G2A's mean score of 64.2%.

To compare the two subtests, student responses from the full administration were split corresponding to questions on each subtest. Their distributions, as well as the score distribution for the full MEGS, are shown in Fig. 3.

The distributions for each subtest were also compared via the Kolmogorov-Smirnov test (KS-test) to evaluate whether student performance on each subtest was sampled from the same overall distribution. The critical
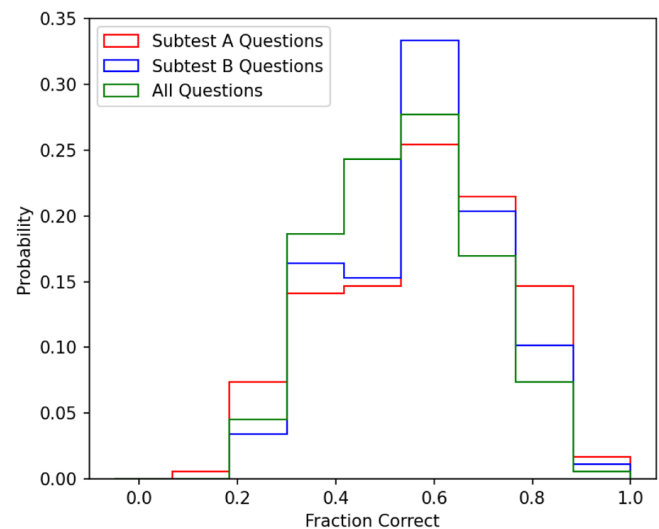

FIG. 3. Score distribution for the full MEGS administration and for the subtest questions coming from the same full administration.
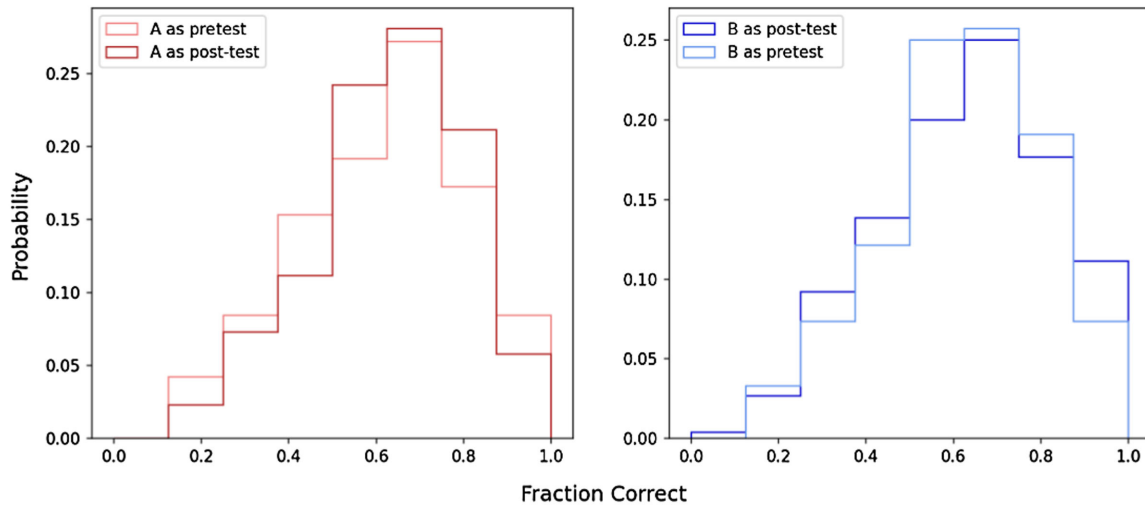
FIG. 4. Score distributions for the four tests administered during the split administration. The left plot compares student performance on test A between the pretest for group 1 and the post-test for group 2. The right plot compares student performance on test B between the pretest for group 2 and the post-test for group 1).

value for this test was 0.145, the KS statistic was 0.096, and the p value was 0.389, all of which indicate that the distributions of scores on the two subtests were likely sampled from the same overall distribution.

The four tests from the split administration were analyzed the same way, with their distributions shown in Fig. 4 and the results of their KS tests shown in Table IV. In all cases, the KS statistic is less than the critical value and the p value is greater than the significance level.

Data for the corresponding subsets of the full administration and the pre- and post-subtests from the split administration were also compared via KS test, shown in Table V. Unlike our findings in Table IV, we found that the responses to subtest A from the full-test administration were not sampled from the same distribution as the responses from either subtest A administration in the split administrations. The same was true for subtest B. For all cases, the KS statistic exceeded the critical value and the p value was less than the significance level.

We also measured test completion time. In general, subtest A takes students slightly longer than subtest B,

and pretests take longer than post-tests. During the full administration, subtest A had a median completion time (MCT) of 12 min and 7 s, and subtest B had an MCT of 10 min and 18 s. From the split administration, pretest A (G1A) had an MCT of 13 min and 9 s and pretest B (G2B) had a median completion time of 10 min and 24 s, post-test A (G2A) had a median completion time of 12 min and 32 s, while post-test B (G1B) had a median completion time of 9 min and 33 s.

We also characterized Cronbach's alpha, a measure of the internal consistency or reliability of our tests; a parameter value greater than 0.7 generally indicates that a measurement is internally consistent (Table VI). The full administration and all four split-administration tests had Cronbach's alpha greater than 0.7, but the value was somewhat less than 0.7 (0.661 and 0.541) when comparing the subtests within the full administration.

TABLE V. KS-test results for comparisons between the full-administration subtests and the corresponding split-administration tests. Bolded KS statistic values indicate that they are above the critical KS coefficient. Bolded p values indicate that they are below the chosen significance threshold.

	G1A (PreA)	G2A (PostA)
Full TestA	$D_{\mathrm{crit}} = 0.132$	0.133
	$\boldsymbol{D = 0.150}$	**0.182**
	$\boldsymbol{p = 0.0150}$	**0.0016**

	G1B (PostB)	G2B (PreB)
Full TestB	$D_{\mathrm{crit}} = 0.133$	0.131
	$\boldsymbol{D = 0.260}$	**0.207**
	$\boldsymbol{p = 9.42 \times 10^{-7}}$	$\boldsymbol{1.64 \times 10^{-4}}$

TABLE IV. KS-test results for comparisons of split-administration score distributions. The listed values are critical value, KS statistic, and p value, respectively.

	G1A (PreA)	G2B (PreB)
G2A (PostA)	$D_{\mathrm{crit}} = 0.119$	0.118
	$D = 0.072$	0.053
	$p = 0.471$	0.826
G1B (PostB)	0.119	0.118
	0.054	0.058
	0.810	0.826

TABLE VI. Cronbach's alpha values for all MEGS administrations. Italicized values are below the threshold of 0.7.

	Cronbach's α
Full MEGS	0.754
Full TestA	*0.661*
Full TestB	*0.541*
G1A (preA)	0.750
G2A (postA)	0.705
G1B (postB)	0.764
G2B (preB)	0.721

IV. DISCUSSION

The scree plot in Fig. 2 suggests that there are around ten underlying factors that capture the variance seen in student performance on the full MEGS. This may be somewhat surprising, as the MEGS was designed to test four epistemic games. Indeed, we performed an exploratory factor analysis (rather than a confirmatory one) to allow for the possibility that student performance on MEGS questions is measurably influenced by more than just the particular epistemic game being queried. Work on other similar instruments has also found more factors influencing student performance than just the skill being directly assessed. For example, the PIQL was designed to measure three facets of quantitative literacy. However, exploratory factor analysis similarly revealed a clear and important substructure to student performance apart from the three factors being examined [9]. In the PIQL case, the researchers selected one question from each of the identified factors and discarded the others to produce a more efficient instrument [9]. This study similarly improves the efficiency of the MEGS, but instead of selecting a single question from each underlying factor, we split each factor in two.

To make sense of the additional factors identified by our analysis, consider the other skills needed to correctly respond to a particular question. Prior research suggests that the process of inquiry involves not only the playing of some epistemic game but also the navigation of some epistemic form [10], a "structure that guides the inquiry process" [20]. In the case of MEGS questions, the storage or encoding of information into tables, graphs, equations, and case studies asks students to navigate a variety of different epistemic forms in order to arrive at correct answers. We speculate that the students in our original sample, many of whom come from other STEM fields, would have had different degrees of the facility with these different epistemic forms, and this difference in the facility with epistemic forms created additional factors in our analysis.

This explanation is consistent with our data for factor 3 (Table II). This factor includes two parts from a three-part question that requires comprehension of cellular processes, comprehension of a case study, interpretation of algebraic expressions, and consideration of scaling and proportionality

within these expressions. While other questions on the MEGS required the same epistemic game as these three, no other question additionally required the navigation of these epistemic forms. By using the factors identified from an exploratory analysis, we can capture a great deal more of the variance in the initial dataset. To allow only four contributing factors in our analysis is to assume the four games that the MEGS was designed to test are each mastered orthogonally by students. Our results suggest that mastery of each of the epistemic games in question is necessary but not sufficient for a high score on the MEGS. There are other important influences on student performance, such as the ability to simultaneously navigate epistemic forms. Given that the MEGS ideally captures something about real-world problem-solving skills, including realistic epistemic forms is an important aspect of the test. However, it would be an unhelpful oversimplification to ignore the interaction of mastery of epistemic games with familiarity with epistemic forms.

Our approach to creating two MEGS subtests resulted in an asymmetric split of questions representing each of the four epistemic games tested (Table III). As there has yet been no research that shows meaningful gains on the MEGS, we do not know for certain if proficiency with one epistemic game precedes proficiency with another. However, the data we do have about the factors that describe student performance actually suggest the opposite. Our results show that mastery of the four games is correlated with each other and also with mastery of epistemic forms. It is irrelevant if a student who expertly examines the extreme cases of a problem but maps variables to physical concepts at a novice level would perform disproportionately well on subtest B and poorly on subtest A because such a student has never been observed.

Our analysis begs the question of whether the epistemic forms framing the MEGS questions should be eliminated so that the results better correlate with the ability to perform a particular epistemic game. Indeed, prior research into student error-checking suggests that, if provided with a framework prompting the use of specific strategies, students do gain proficiency with and increase utilization of those strategies. However, such direct prompting had the consequence of students assuming a "script-like" approach to error-checking "in contrast to more fluid application that we might expect of expert physicists, and hope to develop in our students" [21]. In developing the MEGS, the choice was made to include these contextualizing frames specifically so that the instrument would test students' ability to identify a productive epistemic game to play, as well as their ability to successfully play it [11]. Ultimately, the most reliable way to determine how successfully students will answer a given question is to administer the test and analyze the responses, as we have done in this project. Expert understanding and instructor intention around the epistemic games, forms, and frames needed to correctly solve a problem are inevitably idealizations that miss the

nuances of student cognition that influence performance. They should therefore be thought of as a coarse-grained framework to guide discussions of physics mastery, rather than an inviolable designation for how to understand and predict student performance on any given question.

These many confounding influences (facility with a particular game, familiarity with a particular frame, and comfort with a particular form) make it extremely challenging to assess whether two instruments are co-reliable through expert examination of the questions. We therefore employ statistical metrics to compare student performance on various versions of the MEGS. A KS test showed that student scores on each subtest and the full MEGS were likely drawn from the same distribution, suggesting the subtests are well suited for use as a pretest and post-test. Each subtest required a roughly equal amount of time to complete, another positive feature of a split administration. So, by the metrics of average score and completion time, factor analysis was successful in dividing the full MEGS into two equivalent halves.

Cronbach's alpha indicated an adequate degree of internal consistency for the full administration and both subtests when administered individually. Results were less internally consistent when subsets of the full-administration data were considered. Given that these scores were also marginally lower, we predict that the lower internal consistency within the full test is due to student fatigue during the full administration. For this reason, the shorter subtests we developed here may be a better measure of student skill than the full-length test. On the other hand, Cronbach's alpha has been shown to sometimes be very large for multiscale instruments (like the MEGS, which is based on four distinct approaches) [22].

When considering the split administration tests, it is important to consider the existing context of MEGS scores and gains across a semester. Eichenlaub found that, across a variety of courses and administrations, "the mean score on the MEGS is 17.7" (equating to 59% accuracy), and gains in MEGS performance between a pretest and a post-test were "often negative, and usually small" [11]. In our sample, student performance on the pretest administration has a very similar score distribution to that of the post-test (Fig. 4). The KS-test results (Table IV) similarly show that all four split test administrations are likely sampled from the same distribution. While this implies that students had negligible gains in skills measured by the MEGS after a semester, it also indicates that the proposed division and administration have produced tests that are co-reliable. While we did not specifically measure test-retest reliability or split-halves reliability, our data suggest that our proposed division of the full MEGS does form two tests on which students reliably score similarly, independent of which test they are given first and independent of when during a single semester the test is administered. These are all traits that are desired from a pretest-post-test administration of a diagnostic instrument.

Though grade distributions look similar by eye between the full- and split-test administrations, the KS-test results show that, compared to both pretests and post-tests, student performance on the full administration obeyed a different distribution, with lower average scores on the full-test administration. During the original development of the MEGS, Eichenlaub observed that in interviews, almost all students were able to solve almost all problems in a one-on-one setting with little substantive input (beyond so-called "metacognitive prompts") from the instructor [11]. This suggests that student performance on the MEGS is less about knowledge of physics facts (though that is obviously essential) and more about students' ability to employ metacognitive skills while solving problems. The self-similarity of all other comparisons of test administrations is therefore further evidence that the lower scores on the full administration are due to fatigue and not any difference in student ability. Therefore, the higher average score on the split administrations suggests that the split format meaningfully increases student engagement and test validity while decreasing both student time and class time needed to apply the instrument.

Our study speaks most directly to the reliability of the split administration of the MEGS. The other major aspect of instrument quality is validity or how well the instrument measures what it seeks to measure. Given that student performance improves on the split test relative to the full administration, likely due to lessened fatigue during the shorter instrument, we expect that this split administration is at least as valid as the full MEGS for measuring students' ability to use epistemic games.

The number of students involved in the full administration was relatively small, such that random fluctuations in student performance could have had a bigger impact on aggregate results. The relative numbers of students from different levels of introductory physics were also not consistent between the two semesters (PHYS150 students in the course primarily for pre-engineers vs PHYS101 for premeds, were overrepresented in the full-administration data), complicating some comparisons between the administrations. Future work is needed to examine how sensitive our method for subtest generation is to small changes in the composition of the initial population used to perform the factor analysis.

Potentially, the most significant differences between the two administrations were the differences in the testing environment between the full administration and the subtest administrations. For the full administration, participation was entirely voluntary, with no incentive offered to 90% of students and a small extra credit given to the remaining 10%, and all testing took place in a synchronous classroom. By contrast, the split administration was treated as part of a course, where completion of each test was mandatory, the tests were completed remotely with no time limit, and communication between students and investigators took place largely over official course websites. In order to separate

the influence of those factors on student performance, more controlled administration approaches would be necessary. However, all comparisons between subtests here are reasonably free of confounding factors of administration and we can therefore evaluate their mutual reliability and validity.

Our results suggest that splitting the full MEGS test into the proposed subtests results in two mutually exclusive tests that can be administered at different points across the semester and be mutually reliable. Performance on the split tests is similar to performance on the full MEGS, and some of our data suggest that the shorter test may improve reliability and validity by reducing fatigue. Since the tests' questions are mutually exclusive, our split MEGS facilitates use as a pretest and post-test without student recall of prior questions, a previously identified weakness of the MEGS. We believe we have developed an instrument that is well suited for use as a pretest-post-test diagnostic of

students' ability to use effectively epistemic games. The originally published MEGS instrument and the two subtests developed in this work can be found in the Supplemental Material [23].

ACKNOWLEDGMENTS

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[2] J. Tuminaro and E. F. Redish, Elements of a cognitive model of physics problem solving: Epistemic games, Phys. Rev. ST Phys. Educ. Res. **3**, 020101 (2007).

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

[4] K. E. Black and M. C. Wittmann, Epistemic games in integration: Modeling resource choice, in *Proceedings of the Physics Education Research Conference 2007, Greensboro, NC* (AIP, New York, 2007).

[5] R. R. Bajracharya and J. R. Thompson, Analytical derivation: An epistemic game for solving mathematically based physics problems, Phys. Rev. Phys. Educ. Res. **12**, 010124 (2016).

[6] Y. Chen, P. W. Irving, and E. C. Sayre, Epistemic game for answer making in learning about hydrostatics, Phys. Rev. ST Phys. Educ. Res. **9**, 010108 (2013).

[7] T. O. B. Odden and R. S. Russ, Sensemaking epistemic game: A model of student sensemaking processes in introductory physics, Phys. Rev. Phys. Educ. Res. **14**, 020122 (2018).

[8] D. Hu, K. Chen, A. E. Leak, N. T. Young, B. Santangelo, B. M. Zwickl, and K. N. Martin, Characterizing mathematical problem solving in physics-related workplaces using epistemic games, Phys. Rev. Phys. Educ. Res. **15**, 020131 (2019).

[9] S. W. 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).

[10] A. Collins and W. Ferguson, Epistemic forms and epistemic games: Structures and strategies to guide inquiry, Educ. Psychol. **28**, 25 (1993).

[11] M. Eichenlaub, Mathematical sensemaking via epistemic games, doctoral dissertation, University of Maryland, College Park, 2018.

[12] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the Force Concept Inventory into two equivalent half-length tests, Phys. Rev. ST Phys. Educ. Res. **11**, 010112 (2015).

[13] L. Cohen, L. Manion, and K. Morrison, Validity and reliability, in *Research Methods in Education* (Routledge, London, 2017), pp. 245–284.

[14] M. S. Bartlett, Tests of significance in factor analysis, Br. J. Math. Stat. Psychol. **3**, 77 (1954).

[15] H. F. Kaiser and J. Rice, Little Jiffy, Mark IV, Educ. Psychol. Meas. **34**, 111 (1974).

[16] I. Persson and J. Khojasteh, Python packages for exploratory factor analysis, Struct. Equation Model. **28**, 983 (2021).

[17] H. F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. **20**, 141 (1960).

[18] L. J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika **16**, 297 (1951).

[19] A. Kolmogoroff, Confidence limits for an unknown distribution function, Ann. Math. Stat. **12**, 461 (1941).

[20] L. Sherry and M. Trigg, Epistemic forms and epistemic games, Educ. Technol. **36**, 38 (1996).

[21] T. R. Sikorski, G. D. White, and J. Landay, Uptake of solution checks by undergraduate physics students, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.087.

[22] K. S. Taber, The use of Cronbach's alpha when developing and reporting research instruments in science education, Res. Sci. Educ. **48**, 1273 (2018).

[23] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.19.020152 for the complete test of the originally published MEGS and the two subtests as described in the present work.