

Assessing mathematical sensemaking in physics through calculation-concept crossover

Eric Kuo¹,[✉] Michael M. Hull,² Andrew Elby,³ and Ayush Gupta³

¹University of Illinois at Urbana-Champaign, Urbana, Illinois 61820, USA

²University of Vienna, 1010 Vienna, Austria

³University of Maryland, College Park, Maryland 20742, USA

 (Received 11 March 2019; revised 26 May 2020; accepted 1 July 2020; published 30 July 2020)

Professional problem-solving practice in physics and engineering relies on mathematical sense making—reasoning that leverages coherence between formal mathematics and conceptual understanding. A key question for physics education is how well current instructional approaches develop students’ mathematical sense making. We introduce an assessment paradigm that operationalizes a typically unmeasured dimension of mathematical sense making: use of calculations on qualitative problems and use of conceptual arguments on quantitative problems. Three assessment items embodying this calculation-concept crossover assessment paradigm illustrate how mathematical sense making can positively benefit students’ problem solving, leading to more efficient, insightful, and accurate solutions. These three assessment items were used to evaluate the efficacy of an instructional approach focused on developing students’ mathematical sense making skills. In a quasi-experimental study, three parallel lecture sections of first-semester, introductory physics were compared: two mathematical sense making sections, one with an experienced instructor and one with a novice instructor, and a traditionally taught section, as a control group. Compared to the control group, mathematical sense making groups used calculation-concept crossover approaches more often and gave more correct answers on the crossover assessment items, but they did not give more correct answers to associated standard problems. In addition, although students’ postcourse epistemological views on problem-solving coherence were associated with their crossover use, they did not fully account for the differences in crossover approach use between the mathematical sense making and control groups. These results demonstrate a new assessment paradigm for detecting a typically unmeasured dimension of mathematical sense making and provide evidence that a targeted instructional approach can enhance engagement with mathematical sense making in introductory physics.

DOI: [10.1103/PhysRevPhysEducRes.16.020109](https://doi.org/10.1103/PhysRevPhysEducRes.16.020109)

I. INTRODUCTION

When solving problems, the work of professional physicists and engineers relies on reasoning that leverages *coherence* between formal mathematics and conceptual understanding [1–6]—a form of reasoning which has been described as *mathematical sense making* [7–11]. For this reason, proficiency with mathematical calculations and conceptual reasoning in isolation, though necessary, may not be sufficient to prepare students for the complex, challenging problems they will face in their future work. To better specify the nature of mathematical sense making and its status in the physics curriculum, previous research has shed light on some aspects of students’ struggles and successes with mathematical sense making in specific instances [7,9,10,12–22] and demonstrated approaches to

assessing specific aspects of mathematical sense making [20–24]. Though this work has been illuminating, there is still much that is unknown about (i) how to assess different dimensions of mathematical sense making and (ii) how effectively different forms of physics instruction foster mathematical sense making.

In this paper, we propose a novel assessment paradigm of calculation-concept crossover that highlights one dimension of mathematical sense making identified in the literature and distinguishes it from common physics education research (PER) assessment approaches. We instantiate this paradigm through three assessment questions, each one probing this dimension of mathematical sense making in students’ problem-solving practice in a different way. Using these assessments, we compared the learning outcomes of two approaches to teaching introductory physics, a “traditional” approach and a “mathematical sense making” approach. The results show that an instructional approach designed to foster mathematical sense making—in conjunction with other PER-based active learning strategies—can produce problem-solving benefits detectable by targeted assessments.

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

Our primary aim is to demonstrate a new assessment paradigm that highlights a previously unmeasured dimension of mathematical sense making. Our secondary aim is to show that this dimension can be successfully nurtured through an instructional focus on mathematical sense making. To our knowledge, these assessment results are the first comparison of mathematical sense making outcomes for different, semester-long instructional approaches in large-lecture classroom settings.

II. THE CALCULATION-CONCEPT CROSSOVER ASSESSMENT PARADIGM AND HOW IT DIFFERS FROM EXISTING ASSESSMENT APPROACHES

A. How mathematical sense making crosses the boundaries of typical assessment assumptions in PER

One central assumption underlying assessment in PER is that students take different approaches to answering quantitative and qualitative questions. Quantitative questions are typically interpreted as probes of students' calculation skills (which include the conceptual analyses needed to set up the equations correctly and interpret the solution physically), while qualitative questions are commonly interpreted as uncovering "functional knowledge" [25] or "conceptual understanding." Comparing student performance on these two question types has been illuminating for PER, yielding the fundamental result that students can have greater difficulties in answering qualitative problems that instructors and researchers view as equivalent to or even simpler than analogous quantitative problems [25–27]. Along similar lines, Kim and Pak [28] found little relation between the number of quantitative problems solved in a physics course and performance on the Mechanics Baseline Test, which includes qualitative problems. On the whole, these findings can be interpreted as pointing to a lack of coherence between students' conceptual and calculation knowledge or skills, with students, on average, having weaker conceptual understanding and relatively stronger calculation skills.

Although students may commonly take different approaches to answering quantitative and qualitative questions, the notion of mathematical sense making suggests that this separation need not always hold. We hypothesize that coherence between formal mathematics and conceptual understanding can promote the more expansive use of conceptual and computational problem-solving approaches. In the rest of this subsection, we use examples from prior research to illustrate different ways in which mathematical sense making can expand the connections between conceptual and calculational approaches when answering qualitative or quantitative questions.

Sometimes quantitative questions can be answered with a conceptual argument, sidestepping a standard calculation. Consider the following task: write an expression for the acceleration of a falling ball experiencing air resistance.

This could be calculated by writing out Newton's 2nd law, $F_{\text{net}} = -mg + f(v) = ma$ [where acceleration upward is positive and $f(v)$ is the force of air resistance], and then solving for a . However, Sherin [12] found that 3rd-semester physics students could instead use their conceptual reasoning to generate equations without a first-principles derivation and calculation. These students immediately wrote down the equation $a(t) = -g + f(v)/m$, which expressed their conceptual idea that an "upward acceleration" from air resistance opposed a "downward acceleration" from gravity. Although there is only one mathematical step separating these two approaches, Sherin argued that these two solution methods use distinctly different reasoning. Specifically, Sherin described his students' reasoning as employing the *opposition* symbolic form, which combines the mathematical symbol template $\square - \square$ with the conceptual schema of two influences in opposition. By plugging in mathematical expressions representing the two influences, gravity and air resistance, students were able to express their conceptual idea of two accelerations in opposition. By this interpretation, symbolic forms facilitated mathematical sense making by tying the structure of equations (here, one term subtracted from another term) to an intuitive conceptual interpretation (here, one influence opposing another influence). This is one way in which mathematical sense making can be indicated when students answer quantitative (or symbolic) questions with conceptual arguments rather than standard calculations.

On the other hand, students can also use formal mathematics and calculations to inform and enrich their solutions to qualitative questions. For instance, Schwartz, Martin, and Pfaffman [29] investigated whether prompting a mathematical solution would help children aged 9–11 discover the factors determining whether a balance scale with objects on both sides would tip to the left or the right. Compared to a control group, prompting the use of math in their explanations helped the children better recognize the two key physical concepts, mass and distance from the pivot point, and, in some instances, even identify something like torque (mass \times distance) as the explanatory physical property. Here, the precision of mathematics led students to identify a more complete set of properties governing physical balance. Along similar lines, Sherin [13] found that two undergraduate physics students spontaneously used a calculation to resolve a conceptual tension between two competing intuitions on a qualitative question. The question presented a block traveling at an initial speed v_0 sliding on a rough surface before coming to rest; students were asked how the block's mass would affect the distance traveled. The students articulated two opposing conceptual effects: a greater mass would (i) decrease the sliding distance by increasing the force of friction and (ii) increase the sliding distance by increasing the block's inertia. To determine the result of these two competing effects, the students then used Newton's 2nd law to calculate the block's acceleration.

During their calculation of the final answer, $a = \mu g$, the m 's cancelled out. The students interpreted this mathematical cancellation as the two conceptual effects canceling out, making the block's acceleration—and, therefore, the distance traveled—independent of the mass. Here, both the final result and intermediate steps of a calculation helped students advance their conceptual understanding on a qualitative problem. The students' mathematical sense making here consisted of (i) leveraging the mathematical precision of a calculation to resolve a debate between two conceptual arguments and (ii) associating conceptual meaning with the “guts” of the intermediate equations and operations, specifically the cancellation of the m 's.

In an experimental demonstration of how calculations and conceptual arguments can cross the boundary between qualitative and quantitative problems, Singh [30] found that student performance on qualitative questions can be improved when an isomorphic quantitative problem is given beforehand. Written responses and one-on-one discussions indicated that the calculations used on the quantitative problems sharpened students' conceptual arguments on the subsequent qualitative problems.

Although students may typically reserve calculations for quantitative questions and conceptual arguments for qualitative questions, these examples indicate how mathematical sense making may be evidenced by a more expansive use of these two approaches. Next, we introduce an assessment paradigm that captures this kind of “crossover” reasoning.

B. The calculation-concept crossover assessment paradigm

The calculation-concept crossover assessment paradigm (Fig. 1) highlights problem-solving approaches that deviate from those typically assumed in qualitative and quantitative assessment. Calculation-concept crossover occurs when a calculation is used on a qualitative question and a conceptual argument is used on a quantitative question.

We intend for calculation-concept crossover to supplement, not replace, the typical assessment approaches in

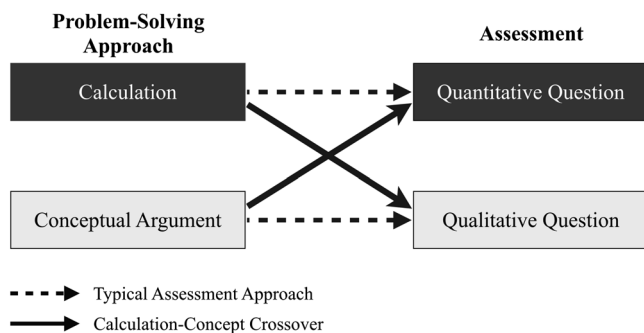


FIG. 1. Though typical assessment approaches link calculations to quantitative questions and conceptual arguments to qualitative questions, a paradigm for assessing mathematical sense making can search for calculation-concept crossover.

PER. Many quantitative questions in physics require a calculation to reach the answer, and many qualitative questions can be efficiently answered with a purely conceptual argument. These typical solution approaches are captured by the useful, existing assessment methods structured to (i) evaluate students' ability to solve quantitative problems with correct calculations (e.g., Ref. [31]), and (ii) assess conceptual understanding through qualitative questions (e.g., Refs. [32–34]). However, the calculation-concept crossover paradigm highlights the existence of productive, alternative solution approaches, which are typically not distinguished through these existing assessments methods. We contend that these crossover approaches represent mathematical sense making, because they demonstrate coherence between the realms of qualitative and quantitative problems rather than restricting certain problem-solving approaches to certain problem types.

In prior work on quantitative problem-solving assessments, we have distinguished the type of mathematical sense making embodied in calculation-concept crossover from those aspects already attended to by PER-based, quantitative problem-solving assessments [11]. Indeed, the existing research in quantitative problem solving already attends to connections between mathematical manipulations and conceptual reasoning, such as in emphasizing an initial conceptual analysis to guide the subsequent calculations [35–37] and final answer checks, which can include conceptual interpretations of the mathematical result (though this is not the only type of answer check) [20,21,23]. What calculation-concept crossover highlights is that conceptual arguments are not only useful for checking a calculated mathematical solution; they can also be used to reach a mathematical solution *in lieu* of a calculation. This possibility is not anticipated by problem-solving assessment rubrics that explicitly assess the success of an “equation manipulation step”¹ [11]. Therefore, although PER-based problem-solving instruction and assessments do combine conceptual reasoning with mathematical calculations, there still exists the standard expectation that quantitative problem solving *requires* a calculation. By presenting the possibility that conceptual arguments, instead of calculations, can be used to answer quantitative questions, the calculation-concept crossover paradigm expands the types of mathematical sense making attended to in quantitative assessments.

The calculation-concept crossover paradigm in Fig. 1 differs in two ways from typical approaches to studying and

¹To be fair, problem-solving rubrics are designed to avoid penalizing students for alternative solution paths or skipping steps if they were not needed by the student. However, these rubrics do present calculations as the default solution approach that one can skip and do not explicitly recognize conceptual arguments as a possible approach one can use instead of a calculation.

assessing mathematical sense making in the literature. First, it can operate with standard questions of the kind typically found on homework and exams. By contrast, most mathematical sense making studies, including our own, use specially crafted prompts—such as asking students to describe the meaning of a mathematical result [24], explain the different ways that a graph “makes sense” [22], check the validity of a quantitative answer [21,23], or construct novel equations from their understanding [12,38]. While certainly informative, these types of questions probe for mathematical sense making in ways that could be disconnected from the typical quantitative and qualitative prompts seen in a physics course and from the typical problem-solving practices students engage with in those courses. Second, our assessment of mathematical sense making relies solely on students’ written answers to exam questions. This differs from common analyses and assessments of mathematical sense making that rely on verbal discourse captured through observations of student discussions [9,10,17,18] or in one-on-one interviews [7,15]—which can include follow-up questions to probe student thinking further. Because students’ written answers on homework and especially on timed exams provide less rich information about students’ thinking, there is a danger that mathematical sense making might go undetected even when present. We present a study showing that students’ use (or not) of a crossover approach can be reliably coded from their written exam responses. In summary, we aim to detect mathematical sense making in students’ actual problem-solving practice in their physics course by analyzing their written exams solutions.

C. Operationalizing “calculation” and “conceptual argument” approaches

Designing problems and evaluating the responses within the calculation-concept crossover paradigm requires an operationalization of two problem-solving approaches: calculations and conceptual arguments.

We take *calculation* approaches to yield solutions based upon formal mathematical manipulation rules and the numerical (or symbolic) answers they produce. As discussed above, calculations are commonly viewed as a necessary step for answering quantitative questions—the “execute the solution” step of standard problem-solving paradigms, known colloquially as the “chug” part of “plug and chug.” Again, as previous PER work in quantitative problem solving has shown, the calculations that arise in expert problem solving are not disjoint from conceptual understanding; conceptual understanding of the physical situation leads to the generation of the appropriate equations used in a calculation. The key operational feature we use to identify a calculation is whether explicit mathematical manipulations are used to produce the final answer.

By contrast, we take conceptual argument approaches to be those which reach a solution through physical concepts and or mathematical concepts. Use of physical concepts involves reasoning about physical entities and the processes between them. For example, a student could reason that pushing down on a piston adiabatically will increase the energy of the gas inside, because the piston does work on the gas by exerting a force over a distance. Use of mathematical concepts involves reasoning about physical quantities and functional relations between them, as represented in equations. For example, one could reason that if the voltage in a circuit were doubled and the total resistance were halved, then Ohm’s law, $I = V/R$, would predict that the current would quadruple. Again, although this answer is consistent with a mathematical calculation and draws on mathematical ideas of proportionality and multiplication, we operationalize this response as a conceptual argument to distinguish it from a calculation approach where one explicitly plugs in the values, performs the algebraic or arithmetic manipulations, and arrives at the answer that the current is quadrupled. Instead, in a conceptual argument, the mathematical expression indicates the proportional relationships that are used directly to make an argument for the final answer. Prior research has carefully distinguished between physical concepts and mathematical concepts for understanding physical systems [39]. Here, however, we group both together as “conceptual,” since both contrast with the mathematical manipulations used for calculations.

These operational definitions of calculation and conceptual argument approaches to solving a problem lead to three clarifying points. First, calculations may rely on conceptual knowledge and conceptual arguments may rely on calculations. How problem-solving approaches are coded relies ultimately on the primary warrant for the answer. For example, in determining the acceleration of an object, one might employ a conceptual analysis to identify the forces in this situation before executing an explicit calculation with Newton’s 2nd law. This would be coded as a calculation approach, not a conceptual argument, since the conceptual analysis is used in the service of the calculation that produces the final answer. By contrast, consider this hypothetical exam answer: “Since the forces acting to the left sum to $20\text{ N} + 40\text{ N} = 60\text{ N}$ and the forces acting to the right sum to $30\text{ N} + 30\text{ N} = 60\text{ N}$, these two forces cancel out and there is no acceleration in the horizontal direction.” Even though calculations of forces are involved, this approach would be coded as a conceptual argument, since the mathematical expressions are used in the service of a conceptual argument that is used to produce the final answer. A calculation approach to this problem would compute the net force and use Newton’s 2nd law to compute the acceleration explicitly. In summary, because physics equations are both representations of the conceptual relations between quantities and tools for mathematical

TABLE I. The three types of questions for assessing calculation-concept crossover used in this study, the type of crossover approach probed by each one, and the potential benefits of these crossover approaches.

Question type	Crossover type	Typical approach	Crossover approach	Potential benefit of using the crossover approach
Qualitative judgment	Using calculations on a qualitative question	Conceptual argument	Explicit calculation that describes qualitative behavior	Calculations may increase accuracy when conceptual understanding is weak or imprecise.
Isomorphic calculations	Using a conceptual argument on a quantitative question	Multiple calculations	Conceptual comparison of multiple scenarios	Conceptual arguments, specifically noticing a conceptual similarity across scenarios, efficiently sidesteps repeated calculations.
Cued symbolic evaluation	Using a conceptual argument on a quantitative question	Calculation	Inferring physical behavior from functional dependencies between variables	Conceptual arguments can detect errors in symbolic solutions derived by calculation.

manipulations [40], the presence or absence of mathematics alone cannot be used to categorize the approach. A classification of calculation or conceptual argument reflects how the mathematics is used in the overall problem-solving approach. Similarly, the presence or absence of conceptual reasoning alone cannot be used to categorize the approach. The crux of the categorization relies on whether the concepts were used to support a calculation or whether the concepts themselves were used to construct an argument for the final answer.

The second clarifying point is that incorrect approaches can also be classified as calculation or conceptual argument. Incorrect calculations that incorrectly model the physical system or contain errors in the mathematical manipulations are still calculation approaches. Incorrect conceptual arguments still embody a conceptual argument approach distinct from calculations.

The third clarifying point is that our coding scheme highlights new aspects of coherence seeking between qualitative and quantitative reasoning by backgrounding others. We do not distinguish between different routes to a calculation, such as generating equations from an initial conceptual analysis versus seeking an equation that includes the variables given in the problem. Although a complete assessment of problem-solving skills should make this distinction, previous research has already distinguished and studied these different approaches to calculation [41,42]. Our aim here is to explore a new dimension of mathematical sense making in physics assessment, the use of calculation-concept crossover approaches.

III. THREE ASSESSMENTS OF CALCULATION-CONCEPT CROSSOVER

Next, we describe the problem-solving benefits of calculation-concept crossover and detail three crossover

assessment items designed to illustrate these benefits (summarized in Table I). Informed by the prior literature, we will hypothesize what calculation and conceptual argument approaches students will take. These examples will illustrate how calculation-concept crossover is distinct from other assessment goals and why it is a worthwhile physics problem-solving behavior to study. We then present the results of an empirical classroom study showing how these assessments were used to measure the mathematical sense making outcomes of two different approaches to teaching introductory physics.

A. Using calculations on a qualitative question: Reliability and precision

1. The nature and benefits of this crossover reasoning

Consider a typical, qualitative circuit question with three identical bulbs, shown in Fig. 2:

What happens to the brightness of bulbs A and B when the switch is closed?

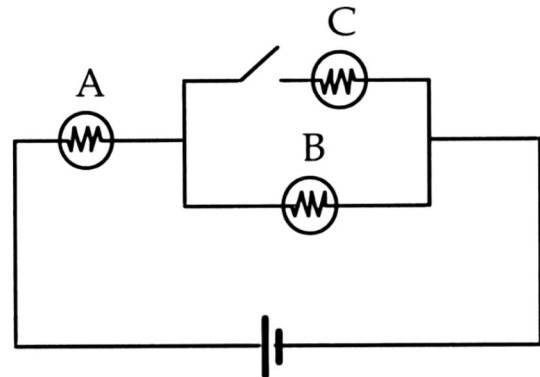


FIG. 2. A typical, qualitative circuit question (from Engelhardt and Beichner [43]).

TABLE II. The qualitative judgment question and associated standard questions.

Assessment type	Question text
	Two identical masses, each of mass $m = 50$ g, are fastened to each other with a bit of plastic explosive. We're going to launch it into the air and detonate the explosive at the highest point. (Ignore air resistance throughout this problem.)
Associated standard questions	<p>(a) Suppose we launch the pair of masses at an angle $\theta = 60^\circ$ above the horizontal, from a spring gun. The spring has a spring constant of $k = 1000$ N/m, and we compress it $x = 10$ cm ($\sin 60 = 0.87$; $\cos 60 = 0.5$). Find the maximum height of the pair of masses, taking its initial height to be 0.</p> <p>(b) At exactly that instant, when it's at the highest point, we detonate the explosive. And it so happens that the instant after the explosion, one mass (A) is not moving at all. Find the velocity of the other mass (B).</p> <p>(c) Find the distance between the masses A and B when they hit the ground.</p> <p>(e) Sketch a graph of the vertical and horizontal components of the velocity for mass B from the time of launch to the time it hits the ground. Explain your reasoning. You don't need to make precise calculations, just show the shape of the graph in your sketch.</p>
Qualitative judgment question	(d) During the explosion, mass B speeds up while mass A comes momentarily to rest. Does the overall mechanical energy of the two-mass system increase, decrease, or stay the same during that explosion? Explain.

Studies of student reasoning on similar qualitative circuit questions have identified common invalid conceptual arguments. Students may overgeneralize the rule that parallel branches in a circuit are independent, which could lead them here to incorrectly conclude that the brightness of bulb B will not change [44]. Additionally, students may use local, sequential reasoning to conclude, for instance, that the brightness of bulb A will not change because the current does not reach the switch until after it has passed through bulb A .

The mathematical machinery of calculations can provide precision not easily achieved through a conceptual argument. For example, when the switch is closed, bulb B receives a smaller fraction of the total current in the circuit—but the total current increases because the circuit's equivalent resistance decreases. Even if students identify these two effects of closing the switch, it may be difficult to work out which competing effect has a larger magnitude through a conceptual argument. Prior work has shown that, in cases with two competing effects that are underdetermined by conceptual arguments, students commonly claim that the effects exactly compensate, causing no overall change [45–52]. Here, calculating the power dissipated in each bulb directly is a more precise approach. Even though the question requires only a qualitative determination of how the brightness of the bulbs change, the mathematical machinery of the calculation determines which of the competing effects wins out. As with Sherin's students, taking a calculation approach, if well practiced and reliable, can help students avoid conceptual errors and complicated conceptual arguments.

2. Assessment item: qualitative judgment question

In our study, we used the following qualitative judgment question to investigate whether introductory physics

students will use calculations to precisely answer a qualitative question. The qualitative judgment question is embedded within a set of associated standard questions (Table II). For this qualitative judgment question (part d), the correct answer is that the overall mechanical energy of the system increases during the explosion.

Conceptual arguments (typical approach).—There are two different valid conceptual arguments. In terms of physical entities, one could reason about the energy transfer processes: the chemical potential energy of the explosive is released, doing mechanical work on the two masses and thereby increasing the mechanical energy of the system.² Another argument uses the mathematical concepts in the kinetic energy equation: because the mass of the system is effectively halved while the speed doubles, the kinetic energy will increase because the proportional dependence of KE on speed is greater than the dependence on mass ($KE \sim v^2$ vs $KE \sim m$). However, as with the circuit question described previously, students may also use invalid conceptual arguments. A student could overgeneralize a commonly stated rule: the law of conservation of energy states that energy is always conserved, so the mechanical energy stays the same. Also, a compensation argument could be used to incorrectly conclude that the total mechanical energy stays the same, because the amount

²This reasoning is correct for the class of problems where the explosion increases the speeds of both blocks in the center-of-mass frame without changing the speed of the center of mass in the rest frame, as in this problem. Without this condition, chemical potential energy can be used to decrease mechanical energy, such as when a rocket uses its engine to slow down. Because the assessment item only asks about this case and no other, we did not demand that students explicitly describe this special condition to be considered correct in their reasoning.

of kinetic energy lost by mass A could be exactly balanced out by the gain in KE of mass B .

Calculation (crossover approach).—A calculation can offer precision and safety from common conceptual argument errors. Here, because the qualitative judgment question is embedded in a series of quantitative problems, numerical results from previous parts can be used to explicitly calculate the pre-explosion and post-explosion kinetic energies. From part (b), the speed of masses A and B is 5 m/s immediately before the explosion and $v_A = 0$ and $v_B = 10$ m/s immediately after the explosion. Calculating the overall kinetic energy before and after the explosion shows that it increases from 1.25 to 2.5 J. Because the change in gravitational potential energy is negligible immediately before and after the explosion, the overall mechanical energy increases.

Either a conceptual argument or a calculation alone is sufficient to reach the correct qualitative answer. Although much of the problem invites calculations and supplies the quantitative values needed to calculate the change in energy, we predict that introductory physics students will tend to use conceptual arguments, because the question is phrased qualitatively. Here, the crossover approach is calculation, which we argue indicates mathematical sense making through an expanded domain of use for calculations. Because we do not tell students what approach to take, their spontaneous choices show both their knowledge and disposition for mathematical sense making during physics problem solving in their courses. Activation of formal calculations on qualitative problems is one of the benefits of mathematical sense making that we propose is not well attended to in typical instruction and assessment. We also predict that the use of conceptual arguments alone will be more prone to error than calculation, given the incorrect arguments predicted.

Although our assessment goal is to highlight crossover approaches, the deeper reason for valuing these approaches is that coherence between multiple approaches provides reliable problem solving through redundancy. Although we predict that calculation approaches will be more accurate on this problem, checking for coherence between calculations and conceptual arguments will provide the best safety net against errors. Mathematical sense making allows problem solvers to check their own solutions by answering the question in multiple ways and making sure that calculations and conceptual arguments cohere.

B. Using conceptual arguments on quantitative problems: Finding efficient insights by leveraging conceptual structure

1. Nature and benefits of this crossover reasoning

Wertheimer [53] asked 6th-grade students to solve arithmetic problems of this type: $(283 + 283 + 283 + 283 + 283)/5 = ?$ Although students could solve the problem correctly because they had learned addition and

long division, some used an explicit mathematical calculation, even though a more efficient conceptual argument invoking an understanding of addition and division can be used here. Similarly, some physics problems that invite quantitative calculation may be answered more efficiently and effectively with a conceptual argument. Kuo, Hull, Gupta, and Elby [15] illustrated this same phenomenon in introductory physics through the following problem:

Suppose you are standing with two tennis balls on the balcony of a fourth-floor apartment. You throw one ball down with an initial speed of 2 meters per second; at the same moment, you just let go of the other ball, i.e., just let it fall. What is the difference in the speeds of the two balls after 5 seconds—is it less than, more than, or equal to 2 meters per second? (use $g = 10 \text{ m/s}^2$ and neglect air resistance)³

In contrast to calculating the speeds of each ball after 5 s to find that the difference in speeds is 2 m/s, some students found an alternative argument: the difference in speeds after 5 s will be the same as the initial difference, 2 m/s, because both objects gain the same amount of speed over 5 s. Even though the formal calculation will yield the correct result, this conceptual argument provides an elegant, insightful answer that bypasses the need for a calculation.

Another example of conceptual approaches leading to efficiency and insight comes when solving a series of isomorphic questions. Consider the following pair of problems:

Linear momentum collision: A block of mass $1.4 M$ is initially traveling at a speed v_0 when it collides with another block of mass $3.7 M$ which is initially at rest. After the collision, the two blocks stick together, traveling at the same speed. What is the final speed of the two-block system?

Angular momentum collision: A solid disk of mass $1.4 M$ and radius R is initially rotating at an angular speed of ω_0 when it collides coaxially with another solid disk of mass $3.7 M$ and radius R , which is initially at rest. After the collision, the two disks stick together, rotating at the same speed. What is the final angular speed of the two-disk system?

A common conceptual structure exists for these two problems. Initially, the (angular) momentum is all in the object of mass $1.4M$. After the totally inelastic collision, the final (angular) speed of the two-object system can be calculated. For the first problem, the final speed of the two-block system can be calculated to be $0.275v_0$. Although a

³Although the final question is actually phrased qualitatively (“is the answer less than, more than, or equal to 2 meters per second?”), the conceptual argument approach here still illustrates the effective insights that can be afforded through conceptual argument. Although conveying the benefits of conceptual argument, we recognize that this would only be an exact example of calculation-concept crossover if the question was solely phrased to require a precise, numerical answer (“What is the difference in the speed of the two balls after 5 seconds?”).

similar calculation can be done to find the final angular speed of the two-disk system, by noticing the common conceptual and numerical structure of these two problems, the solution of the linear momentum problem can be directly mapped onto the angular momentum one without additional calculation, giving the answer $0.275\omega_0$. As in the case of Wertheimer’s arithmetic problem and Kuo *et al.*’s kinematics problem, a conceptual argument pointing out the isomorphism between problems here provides an efficient, elegant way to avoid explicit calculations on a quantitative problem.

2. Assessment item: isomorphic calculation questions

In our study, we investigate whether students make conceptual arguments on quantitative problems with isomorphic calculation questions about a block on a ramp (Table III). For the isomorphic calculation questions (parts b and c), the correct answer is the same as the answer to part a: $mg \sin \theta$.

Calculation (typical approach).—For all three questions, the correct calculation uses Newton’s 2nd law to calculate the forces parallel to the surface the ramp. The force F on the block directed up the ramp is due to the tension in the string or static friction. The force down the ramp is the component

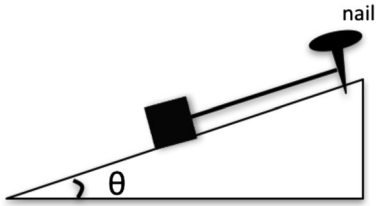
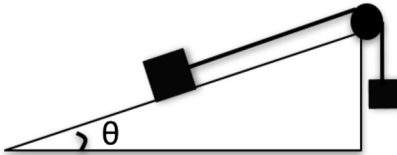
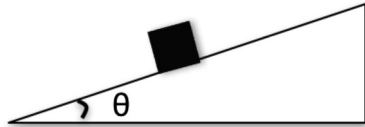
of gravity on the block parallel to the ramp, $mg \sin \theta$. Since the block stays in place, its acceleration is zero. Newton’s 2nd law, $\Sigma F = ma$, yields $F = mg \sin \theta$.

Conceptual argument (crossover approach).—After part (a), subsequent questions can be answered by pointing out the conceptual isomorphism between the questions: in all cases, the component of the block’s weight down the ramp is balanced by a force pointing up the ramp. Therefore, if the answer for the up-the-ramp force in part (a) is $F = mg \sin \theta$, then the answer to subsequent parts must also be $mg \sin \theta$. Instead of performing an identical calculation repeatedly, this conceptual argument leverages the isomorphism between problems to efficiently obtain the solution.

We predict that introductory physics students will tend to use calculations here. When questions are asked one at a time, students can focus on answering each one separately rather than seeking coherence across questions [54]. However, the calculations here are standard enough that we predict that conceptual arguments, though more insightful, will not be more accurate than calculations.

Again, the deeper reason for valuing these crossover approaches is that access to multiple approaches provides more reliable problem solving through redundancy. The conceptual argument here provides a direct check of the

TABLE III. The isomorphic calculation questions and associated standard question.

Assessment type	Question text
Associated standard question	<p>A block of mass M sits on a ramp of angle θ.</p> <p>(a) First, suppose the block is frictionless and is held in place by a light string extended parallel to the surface of the ramp, as shown here. Write an expression for the magnitude of the tension in the string.</p> 
Isomorphic calculation Questions	<p>(b) Instead of a peg, we have the cord connect over a pulley to another block. The second block is just the right mass so that the first block remains at rest. Write an expression for the magnitude of the tension in the string.</p>  <p>(c) Now, suppose there’s no string, but the block stays in place because of friction (with coefficient of static friction μ) between the block and the ramp. Write an expression for magnitude of the friction force by the ramp on the block.</p> 

calculations, and vice versa. While we focus on the cross-over approaches because they demonstrate an insight and because we expect them to be less common in students' thinking, the ultimate goal is for students to be able to provide self-checks through multiple convergent problem-solving approaches.

C. Using conceptual arguments to detect errors in quantitative solutions

1. Nature and benefits of this crossover reasoning

A third benefit of coherence between formal mathematics and conceptual reasoning is the ability to detect errors in quantitative solutions. Physicists often view symbolic answers as superior to numerical answers, because a symbolic answer explicitly represents relationships between the quantities. Similarly, students are able to “[map] mathematics to meaning” in physics problem solving [18]. This is one way that students can check their answers in quantitative problem solving. Some common checks compare the units, signs, and magnitude of the mathematical answer to one’s conceptual understanding of the system. With directed instruction, students can also engage in analyzing the limiting or special cases of a symbolic expression in terms of the expected physical behavior [20,23].

Here, we investigate a particular conceptual argument: evaluating the functional dependencies of a symbolic expression against the expected physical behavior. One way to assess this conceptual argument is to see if students spontaneously perform this comparison when obtaining a

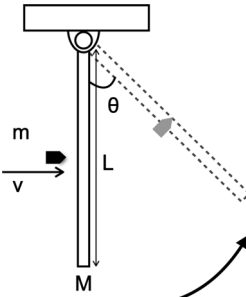
symbolic answer to a quantitative problem. However, a student’s behavior on classroom tasks may depend largely on time constraints, their expectations about “what the professor wants,” and other such factors. Partly for this reason, here we aim to directly assess calculation-concept crossover *skill* instead of proclivity + skill by engineering a problem that explicitly cues students toward the crossover approach, rather than looking for its spontaneous use.

2. Assessment item: cued symbolic evaluation question

Instead of first asking students to calculate a symbolic solution to a quantitative problem, we ask them to evaluate the solution to a quantitative problem without performing the relevant calculation (Table IV). Rather than searching for a spontaneous evaluation of symbolic expressions, this cued symbolic evaluation question directly asks for it. By disallowing calculations and explicitly requesting an evaluation of an expression first, this item tests students’ skill at using conceptual reasoning to debunk an incorrect, proposed solution. We label this as a crossover approach, because we do not allow students to use a calculation to evaluate a quantitative solution.

This question is not a novel assessment; it is common to ask students to check whether a symbolic answer is plausible. We include it here (i) to frame this existing prompt as being aligned with the calculation-concept crossover paradigm and (ii) to include an assessment in this study that is more continuous with existing assessments of mathematical sense making.

TABLE IV. The cued symbolic evaluation question and associated standard problem.

Assessment type	Question text
	
	<p>A uniform rod of length L ($=1.00$ m) and mass M ($=1.80$ kg) is hanging vertically from a frictionless pivot at its top end. A bullet of mass m ($=400$ g) strikes the rod at the center of the rod and gets embedded in it (See figure). Right at the instant before the bullet hits the rod, the velocity of the bullet was entirely horizontal (and perpendicular to the rod) and the magnitude of the bullet’s velocity was v ($=100$ m/s). You can imagine that the rod with the embedded bullet would rotate about the pivot. The moment of inertia of the rod about the pivot is $\frac{1}{3}ML^2$.</p>
Cued symbolic evaluation question	(a) Solving for the angular speed of the rod with the embedded bullet immediately after the collision, a student comes up with this answer: $\omega = 3Mv/2mL$. Is that a plausible answer? Explain your reasoning to her without solving for that angular speed yourself.
Associated standard question	(b) Now solve for the angular speed of the rod with the embedded bullet immediately after the collision.

Conceptual argument (crossover approach).—Here, two valid conceptual arguments (in our dataset) correctly lead to rejection of the proposed equation for angular speed. Both involve comparing the mathematical expression to expected physical outcomes. One pathway is to reject the mathematical dependencies as not reflecting the physical dependencies of the system. For example, a larger rod mass should resist the motion more, yet the proposed mathematical equation says that angular speed ω increases as the rod’s mass M increases. Similarly, a larger bullet mass m will cause the rod to move faster after the collision, but the proposed equation says that the angular speed decreases as the bullet’s mass increases.

The other pathway is to compare speeds before and after the collision. Because $v_f = \omega L/2 = 3M/4mv$ and $3M/4m > 1$, this equation implies that the speed of the bullet increases after the collision, which violates both conservation of momentum and common sense. As a distinguishing example of our coding scheme, this comparison involves mathematical manipulations, but it is not coded as a calculation. It is coded as a conceptual argument, because the symbolic result is used in service of a conceptual argument that determines the final answer: “no, this answer is not plausible.”

Although dimensional analysis is a commonly prescribed method for checking the consistency of a symbolic solution, dimensional analysis cannot detect the errors in this expression since it has the correct units.

IV. MATHEMATICAL SENSEMAKING INSTRUCTION TO FOSTER COHERENCE BETWEEN CALCULATIONS AND CONCEPTS

A. A mathematical sense making instructional approach

We hypothesized two impedances to crossover approach use in introductory physics. First, developing the knowledge and skills to leverage the coherence between calculations and concepts across qualitative and quantitative problems is difficult. Second, typical instructional approaches may not emphasize this coherence, leading to formation of epistemological views that calculations and conceptual reasoning are distinct. In introductory physics courses, this epistemological messaging could arise in several ways: quantitative and qualitative questions can be seen as separate kinds of problems; quantitative calculations can be perceived to be “real” physics, time constraints caused by content coverage demands restrict the time needed for students to engage in deep coherence seeking between calculations and conceptual reasoning, and so on. Indeed, as Hammer [55] documented, students immersed in these instructional environments can view physics as consisting of disconnected pieces and problem solving as formula selection and manipulation, views that

likely impede the development of calculation-concept crossover and other flexible problem-solving approaches.

This study presents results from a mathematical sense making (MS) instructional approach, developed to help combat the content-based and epistemological challenges to seeking coherence during physics problem solving. We hypothesized that the MS instruction was a better instructional approach for fostering the coherence-seeking and problem-solving flexibility that mark mathematical sense making. To investigate this hypothesis, we conducted a quasi-experimental classroom study using the three calculation-concept crossover assessments described earlier to compare learning outcomes of the mathematical sense making instruction to those of a more traditionally taught course. In doing so, we had two goals. Our primary goal was to demonstrate a novel assessment approach for highlighting underexamined and typically unmeasured aspects of physics learning. In particular, we investigated the usefulness of the calculation-concept crossover assessment paradigm for characterizing dimensions of mathematical sense making in introductory physics. Our secondary goal was to characterize student thinking in the mathematical sense making courses, to see if this approach could produce measurable benefits when compared to traditional instructional methods. We see these dual goals as mirroring those of other PER-based assessments. For example, the introduction of the Force Concept Inventory highlighted the, at the time, unmeasured conceptual learning goals for Newtonian physics [32]. At the same time, the Force Concept Inventory was used to demonstrate the benefits of active-engagement pedagogy over traditional approaches [56].

The mathematical sense making instruction draws on common pedagogical techniques from educational research for conceptual and epistemological development. For example, the large lecture incorporates peer-instruction-style clicker questions and peer discussion [26,57]. Instructors use the clicker questions to generate student discussion of ideas. Importantly, for problems where responses do not converge to the correct answer, the instructor will elicit explanations for the two or three most popular choices. The classroom motto is to figure out not only why the right reasoning is right, but also why the wrong explanations are wrong. This aims to give students experience with resolving inconsistencies in the service of developing more coherent understandings. Additionally, the instruction focuses on developing students’ epistemologies for coherence seeking between calculation and conceptual reasoning; the instruction in the mathematical sense making course contains explicit epistemological messaging along these lines, as described in more detail by Redish and Hammer [58]. To provide a feel for this curriculum in action, descriptions of two course elements used to foster mathematical sense making are available in Ref. [59].

B. Predictions of mathematical sense making students' performance on crossover assessments

In this study, we pose our three crossover assessment items to students in a control (CTRL) lecture section and mathematical sense making lecture sections of the same physics course. The control class emphasized conceptual understanding and quantitative problem solving of the type typically emphasized in end-of-chapter textbook problems, with class discussions in lecture arising from student questions. There was no clear indication of PER-based instructional methods being used in the control class. Overall, our prediction is that the mathematical sense making instruction fosters coherence between calculations and concepts, coherence that does not automatically develop through standard instructional approaches, and therefore the mathematical sense making students will use more crossover approaches than the control students do.

Moreover, we interpret calculation-concept crossover approaches as indicating both knowledge and epistemological stances supporting mathematical sense making. On the one hand, calculation-concept crossover indicates that students have developed knowledge and skills that support fluency with each reasoning approach and the flexibility to apply them for different reasoning tasks. On the other hand, the flexibility illustrated by calculation-concept crossover reveals epistemological stances that support coherence and integration. Therefore, we predict that the mathematical sense making instruction will support greater crossover approach use, in part, through epistemological sophistication, so students in the mathematical sense making course will express epistemological views that more strongly favor coherence between mathematics and concepts in problem solving (i.e., a stronger MS epistemology).

If these two predictions are correct, we will also test whether surveyed epistemologies can completely explain the greater use of crossover approaches in the mathematical sense making sections. One reason for this test is to see whether epistemological surveys can completely capture the benefits of the mathematical sense making instruction. For the proposed crossover assessments to provide an instrumental benefit, it is important that surveyed epistemologies are not sufficient for explaining effects measured by crossover assessments. If they contain no additional information beyond that obtained in epistemological surveys, crossover assessments, though theoretically interesting, do not add any additional power for resolving differences between course outcomes over (easier to score) epistemological surveys.

For each of the three crossover problems, we also investigate specific predictions about whether use of crossover approaches increases the correctness of students' answers:

Qualitative judgment question (exploding blocks).—In light of the common conceptual reasoning errors possible, we find it reasonable to predict that, compared to CTRL

instruction, the MS instruction will increase correctness on this tricky qualitative problem by increasing calculation use (the crossover approach) on this problem. As with Sherin's students, the calculation may bring increased precision.

Isomorphic calculation questions (ramps).—Because these types of problems can be solved with simple calculations, the predicted increase in conceptual arguments (the crossover approach) used by the MS students will demonstrate insight and efficiency, but perhaps not increased accuracy.

Cued symbolic evaluation question (ballistic pendulum).—Even though the item tells students not to use calculations, we predict that, compared to CTRL students, MS students will use conceptual arguments for evaluating mathematical expressions more often, leading to more correct evaluations.

V. METHOD

A. Participants

Participants were undergraduate students enrolled in a first-semester calculus-based introductory physics course, taken mostly by engineering majors, at a large, public, research university. Over 15 weeks, the weekly class time consisted of 2.5 h of lecture in a large lecture hall led by an instructor and a 50-min discussion section led by a teaching assistant (TA). 347 students across three course sections consented to have their data used in this study. Consent rates were relatively low for the CTRL section (56%) as compared to two mathematical sense making sections taught by physics education researchers (94%).

B. Design

Because of large enrollment, students at this university were split between three different sections of the course, each with a separate lecture instructor, discussion sections, homework assignments, and midterm exams. The control class was taught by a theoretical physicist. One class using the mathematical sense making curriculum was taught by a senior physics education researcher (MS). Both of these instructors had taught in this department for at least ten years and were regarded as excellent instructors. Another class using the MS curriculum was taught by a junior physics education researcher, who was teaching a large lecture course for the first time (MS-nov), though they had taught smaller courses using research-informed instructional methods. The two mathematical sense making classes used the same lecture materials and homework assignments. The instruction in the CTRL course was not affected by this study. The CTRL instructor taught as they normally would. (Note: we are using "they" as the gender-neutral pronoun for all instructors).

The primary comparison of interest is between CTRL and MS groups, demonstrating what can be achieved by two experienced instructors, each using their respective

teaching approaches. A comparison between the MS-nov and CTRL groups is of secondary interest, to investigate possibilities for first-time, large-lecture instructors to accomplish the novel goals of the mathematical sense making curriculum.

The key assessments occurred at two points. A set of crossover assessments were included on a common final exam, co-designed by the three instructors. Each of the three free-response problems included one crossover assessment item—qualitative judgment, isomorphic calculation, or cued symbolic evaluation—as a subpart, attached to associated standard problems. The test also contained 10 multiple-choice items, which are not included in our analysis. Students in all three instructional groups took the 2-h final exam simultaneously. As described below, we separately coded students’ responses to the crossover items and to the associated standard problems.

In addition, a modified expectations survey (MPEX2) [60] was given during the last week of class. This survey contained 29 items, most from the MPEX2 and some created to target the mathematical sense making curriculum’s explicit goals. We took 15 items from this survey related to math-concept coherence and seeking coherence during problem solving to construct an MS epistemology score. Students completed the survey online, outside of class. There was no systematic presurvey given. See Supplemental Material at [61] for the 15 items used to construct the MS epistemology score.

C. Coding scheme for crossover assessments and the associated standard problems

For the associated standard problems, calculation problems were coded only on whether students used the correct approach. Approaches that correctly plugged problem-specific values into appropriate equations were coded as correct, even if arithmetic errors led to incorrect final answers. Because some subparts of each problem were related, answers that correctly utilized incorrect values calculated in previous subparts were coded as correct, to avoid multiple penalties for initial errors. For the problem requiring graphs, the graphs were coded as correct if they correctly represented the qualitative behavior.

For each of the three crossover assessments, we coded for (i) the approaches taken and (ii) correctness. The next subsections describe the specific coding scheme for each problem. If a student used multiple approaches, as least one approach correctly leading to the final answer was sufficient to be coded as correct. See Supplemental Material at [61] for a more detailed discussion of the coding scheme, examples of coded student work, and details on how disagreements between coders were resolved.

1. Qualitative judgment question (exploding blocks)

For the qualitative judgment problem, there were approach codes for both calculation and conceptual

argument. Solutions that were coded as calculation (the crossover approach here) included plugging numerical or symbolic values into a mathematical expression and then performing mathematical manipulations to compute values that determined the answer. The conceptual argument code indicated justifications for a final answer that did not use an explicit calculation, either by reasoning about physical quantities or reasoning about mathematical dependencies using relevant equations. Since calculation and concept use were independent, a student could be coded as attempting both or attempting neither.

Solutions were coded as *correct* if they indicated that the mechanical energy increased and gave a correct justification. The correct calculation involved correctly calculating the kinetic energies before and after the explosion. Correct conceptual arguments argued that (i) the chemical energy in the explosive was converted to mechanical energy, (ii) the explosion did work on the masses, increasing their kinetic energy, or (iii) halving the mass and doubling the speed would lead to an increase in mechanical energy since kinetic energy depends more strongly on speed than mass (since $KE \sim m$ and $KE \sim v^2$).

2. Isomorphic calculation questions (ramps)

For the isomorphic calculation problem, we again coded for calculations and conceptual arguments. Calculation required producing a mathematical expression and performing manipulations to determine the final answer. Simply writing a mathematical expression or describing a calculation in words was not sufficient to be coded as calculation. The conceptual argument (crossover approach) stated that the situation was isomorphic to a previous problem, so the answer should be the same as before. As there were two isomorphic calculation questions, the crossover code on this quantitative question was given when students used a conceptual argument for at least one part.

Solutions were coded as correct if the correct expression for the relevant force was given, $mg \sin \theta$, and if either the calculation or conceptual argument was correct. Students’ correctness score was the sum of their correctness on the two isomorphic calculation questions.

3. Cued symbolic evaluation question (ballistic pendulum)

Because the cued symbolic evaluation question explicitly directed students not to perform a calculation, we were stricter in when we gave a conceptual argument code than we were in the qualitative judgment question (exploding blocks). We coded an approach as conceptual argument (crossover approach) only when the student evaluated the given expression against the expected physical behavior. More specifically, an approach was conceptual when (i) the direct and/or inverse proportional dependences in the given expression were tested against the expected physical behavior, or (ii) the speeds before and after collision were

compared and tested against the expected physical behavior. Again, the conceptual code is the crossover code here, since the standard version from which this question is adapted is quantitative (i.e., “calculate the symbolic expression”).

Solutions were coded as correct if the given mathematical expression for angular speed was deemed implausible and a correct approach was taken. A correct use of proportional dependence would reject the given expression because the proportional relations between the rod’s mass or the bullet’s mass and final angular speed are incorrect. A correct speed comparison would reject this expression because it says that the (linear or angular) speed increased after collision, a physical impossibility. These two conceptual approaches were the only two approaches found to correctly debunk the proposed expression.

4. Interrater reliability

The coding scheme was initially generated by three of the authors by examining a small subset of student responses. Then, after initially using another small set of student responses to calibrate their coding, the first and second authors coded 45% of students’ responses to all three crossover problems and the associated standard problems, distributed proportionally across the data collected from the three instructional groups. After this initial round, a second round of coding broke down the approach and accuracy of the crossover problems in greater depth, as reported in the coding scheme. In this second round, the first and second authors recoded a subset of student responses (20% of the total data corpus⁴). After each round of coding, the authors discussed disagreements and modified the coding scheme to resolve those disagreements. After these two rounds of coding and discussion, the first author then coded all remaining responses. For all results presented, the coders reached an average of 95% agreement on all codes before discussion (average $\kappa = 0.85$; lowest code agreement = 88%, lowest code $\kappa = 0.75$).

VI. RESULTS

The subsequent analysis excludes the 23 students (5 CTRL students, 15 MS-nov students, and 3 MS students) who did not attempt all three crossover problems, leaving 324 students (CTRL $n = 72$, MS $n = 134$, MS-nov $n = 118$). The exclusion of these 23 students did not change the overall patterns of significance in the results.

⁴For the calculation attempt codes on the isomorphic calculation problems, only 30 responses were coded, because this was determined to be a simple code to apply. Agreement on this code was in-line with the average agreement for all other codes.

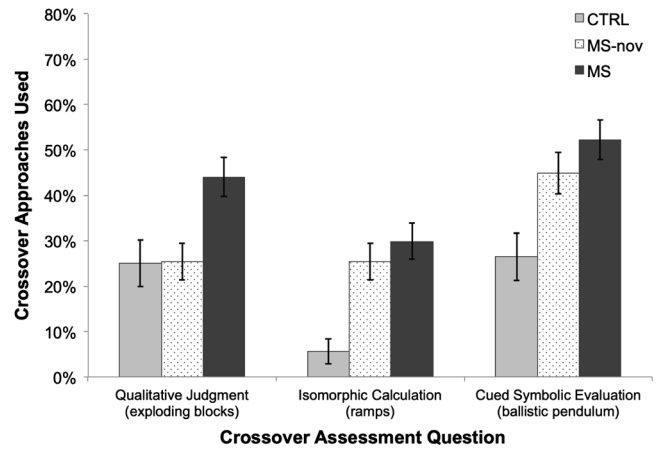


FIG. 3. Percentage of calculation-concept crossover approaches used in the three instructional groups on the crossover assessments. Error bars represent ± 1 SEM, calculated from the binomial distribution.

A. Mathematical sense making supports use of calculation-concept crossover approaches

Figure 3 shows the percentage of calculation-concept crossover approaches used by the three instructional groups. Our primary comparison of interest is between the MS and CTRL groups. The MS group used more crossover approaches than the CTRL instructional group on all three crossover assessments: qualitative judgment (exploding blocks), $\chi^2(1, N = 206) = 7.25$, $p = 0.007$ isomorphic calculation (ramps), $\chi^2(1, N = 206) = 16.5$, $p < 0.001$, and cued symbolic evaluation (ballistic pendulum) $\chi^2(1, N = 206) = 12.8$, $p < 0.001$. This confirmed our main prediction for all three crossover assessments: MS instruction better supported the calculation-concept crossover when solving physics problems compared to CTRL instruction.

The MS-nov group partially matched our predictions for the mathematical sense making curriculum, using more crossover approaches than CTRL students on two of the three crossover assessments: isomorphic calculation (ramps), $\chi^2(1, N = 190) = 12.0$, $p < 0.001$, and cued symbolic evaluation (ballistic pendulum), $\chi^2(1, N = 190) = 6.52$, $p = 0.01$. The MS-nov group’s use of crossover approaches did not differ from the CTRL group’s on qualitative judgment (exploding blocks), $\chi^2(1, N = 190) < 0.01$, $p > 0.90$.

B. Differences in correctness on crossover assessments matched differences in crossover approach use for predicted problems

Turning to the correctness on each of the three crossover assessments (Fig. 4), mathematical sense making students generally outperformed CTRL students on the two predicted crossover assessments: qualitative judgment (exploding blocks) and cued symbolic evaluation (ballistic pendulum). The MS group outperformed the CTRL group on both of these problems: qualitative judgment

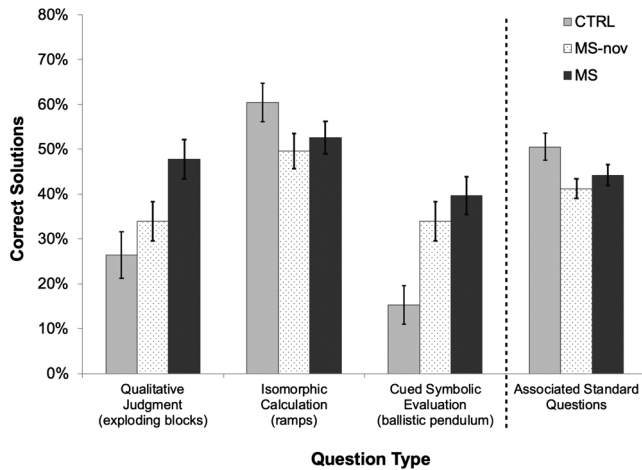


FIG. 4. Percentage of correct solutions of the three instructional groups on the three crossover assessments and the associated standard questions. Error bars represent ± 1 SEM, calculated from the binomial distribution for binary measures.

(exploding blocks), $\chi^2(1, N = 206) = 8.89, p = 0.003$, and cued symbolic evaluation (ballistic pendulum), $\chi^2(1, N = 206) = 12.9, p < 0.001$. The MS-nov group outperformed the CTRL group on cued symbolic evaluation (ballistic pendulum), $\chi^2(1, N = 190) = 7.90, p = 0.005$, but not qualitative judgment (exploding blocks), $\chi^2(1, N = 190) = 1.18, p = 0.28$.

For isomorphic calculation (ramps), we predicted that there would be no significant difference in correctness between groups, since crossover approach use (conceptual argument recognizing the isomorphism between problems) would not help students be more accurate; the calculation is relatively straightforward. Results matched this prediction: On these questions, the CTRL group trended to be more correct than either mathematical sense making group, but this difference was not significant when comparing to either the MS group, $t(172.1) = 1.35, p = 0.18$, or the MS-nov group, $t(188) = 1.82, p = 0.07$.

One possibility for why the mathematical sense making groups outperform the CTRL group on the crossover items might be generally better physics problem-solving skill in the MS and MS-nov groups. To explore this possible explanation, we compared the three groups on their mean performance on the 6 associated standard problems. There was a difference between the three groups on the total associated standard problem score, $F(2, 321) = 3.03, p < 0.05$. To correct for multiple comparisons, we made pairwise group comparisons with the Games-Howell procedure. The only significant pairwise difference was that the CTRL group scored higher on associated standard problems than the MS-nov students, $p = 0.04, d = 0.38$. The MS-nov group’s worse performance on the standard problems makes their better performance on the cued symbolic evaluation question even more notable. Similarly, even though there was no significant performance difference between the MS and CTRL groups

on the standard problems, the MS students were more accurate on the crossover assessments, as predicted. These results indicate that the benefits of the mathematical sense making instruction were not detected as “general” problem solving skills. Rather, there is a particular benefit that is captured by the calculation-crossover assessments.

Given this conclusion, one follow-up question is whether MS and MS-nov students outperformed CTRL students on the crossover questions *because* they used crossover approaches more frequently on those items. Notably, the patterns of significant differences between groups in correctness in Fig. 4 match the patterns of differences for crossover approaches used in Fig. 3. On the cued symbolic evaluation question, the connection is an obvious one. The results of the coding revealed that only the two coded crossover approaches, comparing proportional dependencies to physical behavior or comparing initial and final speeds, yielded a correct judgment. Therefore, success on this question is by definition connected to the coded crossover approach use.

However, the connection between crossover approaches and correct answers on the qualitative judgment problem bears closer analysis. In principle, both conceptual argument and calculation (the crossover approach here) can yield the correct answer. Figure 5 breaks out the approach categories coded (calculation only, conceptual argument only, both calculation and conceptual argument) for the three instructional groups and indicates what percentage of solutions taking each approach was correct. The crossover approach percentage shown in Fig. 3 is the sum of “calculation only” and “calculation and conceptual argument” approaches. On this problem, only 19% of conceptual argument approaches were correct. In comparison, approaches that included a calculation were much more successful.

A breakdown of the common conceptual argument errors shows that this low correctness rate comes from

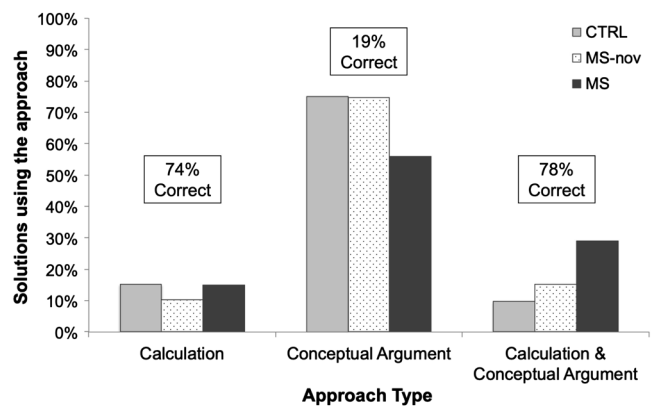


FIG. 5. Approaches used on the qualitative judgment question (exploding blocks)—calculation, conceptual argument, and both calculation and conceptual argument—along with the percentage of correct answers produced by each approach.

misapplications of common explanations in introductory physics. 49% of students that only gave conceptual arguments concluded that the mechanical energy will remain the same. The two most common justifications were (i) the general principle “energy is always conserved” and (ii) a compensation argument—the energy lost by the stopped block would be exactly gained by the accelerated block, leaving the overall mechanical energy the same. 19% of students that only gave conceptual arguments concluded that energy would decrease, commonly citing nonmechanical energy released by the system in the explosion (e.g., heat, light, sound, deformation, etc.). The remaining errors were incorrect justifications of the correct final answer or solutions that left the final answer ambiguous.

This breakdown suggests that the MS group’s increased use of crossover approaches—here, calculations—and the higher accuracy of approaches that incorporated a calculation explain why the MS group was more correct on the qualitative judgment problem than the CTRL group. To test this mediation, we performed a $3 \times 2 \times 2$ log-linear analysis, using instruction (CTRL, MS-nov, or MS), crossover approach use (did or did not include a calculation), and correctness (correct or incorrect) as the three factors. Log-linear analysis tests for relationships between multiple categorical variables. Our analysis used log-linear model selection, which starts with a completely saturated model, including all one-way, two-way, and three-way relationships and removes the highest-order relationships that do not significantly contribute to the fit of the model, one at a time until all remaining terms contribute significantly to the model. The final model contains only the highest-order, significant relationships between factors. In the first step of model selection, the three-way association was deemed to not significantly contribute to the fit of the model, $\chi^2_{\text{change}}(2, N = 324) = 1.63$, $p = 0.44$, and was removed. This indicated that the percentage of correct answers produced by each approach did not differ by instructional group. In the second step, the relationship between instruction and correctness was removed, $\chi^2_{\text{change}}(2, N = 324) = 326$, $p = 0.20$. This indicated that there was no direct association between instructional group (CTRL vs MS vs MS-nov) and correctness. In the final model, instruction was associated with approach, $\chi^2_{\text{change}}(2, N = 324) = 12.4$, $p = 0.002$, and approach was associated with correctness, $\chi^2_{\text{change}}(1, N = 324) = 104$, $p < 0.001$, confirming that the link between MS group and correct answers is mediated by crossover approach use. The final overall model fit did not significantly deviate from the data, $\chi^2(4, N = 324) = 4.90$, $p = 0.30$.

C. Mathematical sense making supports explicit coherence seeking

One reason for valuing the different calculation and conceptual argument approaches to a problem is that

TABLE V. The percentage of explicit coherence-seeking approaches (both calculation and conceptual reasoning) used on the qualitative judgment question and the percentage of students who gave at least one explicit coherence-seeking solution on the isomorphic calculation questions. * indicates percentage is greater than CTRL percentage, $p < 0.01$.

	Explicit coherence approaches	
	Qualitative judgment question	Isomorphic calculation questions
CTRL	9.7%	2.8%
MS-nov	15.3%	16.1%*
MS	29.1%*	14.9%*

multiple approaches provide a method for self-checking during problem solving. Demanding that multiple approaches must converge on the same answer can warn a problem solver of errors in any one approach. In addition to looking at crossover approach use, we can look for such explicit demonstrations of coherence-seeking through solutions giving both a calculation and a conceptual argument for an answer. Table V shows the percentage of approaches that demonstrated this kind of explicit coherence seeking, omitting the cued symbolic evaluation question because it prompted students not to use a calculation. The explicit coherence-seeking approaches on the qualitative judgment question are just a renaming of the “calculation & conceptual argument” approach shown in Fig. 5.

In sum, the patterns of explicit coherence approaches mirror the patterns for crossover approach use. MS students gave more explicit coherence-seeking responses than CTRL students on the qualitative judgment question, $\chi^2(1, N = 206) = 10.1$, $p = 0.001$, and the isomorphic calculation question, $\chi^2(1, N = 206) = 7.25$, $p = 0.007$. The MS-nov group did not display explicit coherence seeking more than the CTRL group on the qualitative judgment question, $\chi^2(1, N = 190) = 1.20$, $p = 0.27$, but they did on the isomorphic calculation questions, $\chi^2(1, N = 190) = 8.07$, $p = 0.004$. This illustrates the success of the mathematical sense making instruction for having students explicitly demonstrate coherence between calculations and conceptual arguments in their solutions.

Although standard problem-solving paradigms often include a “check your answer” step at the end, they differ from our focus on explicit coherence by making calculations primary and other approaches a secondary check of that calculation. Our focus on coherence seeking between calculations and conceptual reasoning places the emphasis on the coherence rather than the primacy of one approach over another. This is more descriptive of a wider range of problem-solving approaches, as an initial conceptual argument may precede the calculation rather than follow it.

D. Associations with problem-solving epistemologies: MPEX2 results

On top of completing the crossover assessments, 240 students (CTRL: $n = 47$; MS-nov: $n = 94$; MS: $n = 99$) also completed the modified version of the MPEX2 (leaving no more than 2 out of 32 items blank) in the final week of the course. Before analyzing the results, we selected 15 MPEX items that were tied to the mathematical sense making instructional goals of fostering coherence-seeking and problem-solving flexibility. Favorable responses to these items had a high reliability ($\alpha = 0.82$) and were combined into an MS epistemology score (percentage of favorable responses, ranging from 0 to 100%). The averages for each instructional group are shown in Fig. 6.

There was a significant difference between MS epistemology scores by instruction, $F(2, 237) = 19.6$, $p < 0.001$. *Post hoc* comparisons using the Games-Howell test reveal that MS-nov students scored higher than CTRL students, $p < 0.001$, $d = 0.70$, and MS students scored higher than MS-nov students, $p = 0.007$, $d = 0.44$ (Implying, of course, that MS students scored higher than CTRL students, $p < 0.001$, $d = 1.18$). Notably, no CTRL students had an MS Epistemology score above 60% whereas 22% of MS-nov students and 43% of MS students did. This matched our expectation that the mathematical sense making curriculum would favorably impact students' epistemologies.

These results suggest one possibility: the increased use of crossover approaches by the MS and MS-nov students is explained by their MS epistemology score. That is, the mathematical sense making instruction impacts both students' problem-solving approaches and their espoused views on how calculations and concepts should be used together in problem solving. Therefore, the crossover assessments, though indicating differences in mathematical sense making outcomes between instructional groups, may

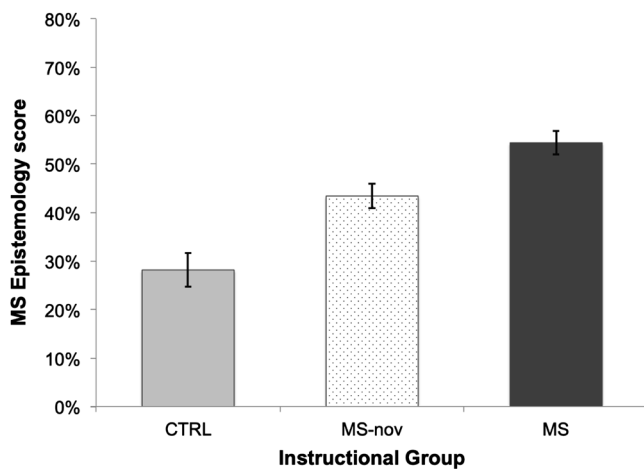


FIG. 6. Average MS epistemology score (% of favorable responses) for each instructional group. Error bars represent ± 1 standard error of the mean.

not measure anything distinguishable from students' surveyed epistemologies. To see whether this was true, we tested a model using instructional group to predict crossover approach use while controlling for MS epistemology score. We summed the number of crossover approaches used by each student, ranging from 0–3. Because the distribution of crossover approaches used was skewed toward zero, we used a Poisson-distributed general linear model with the number of crossover approaches as the dependent variable and instructional group and MS epistemology score as the independent variables. Each student's crossover approach total is modeled through

$$y = \exp[C + b_{MS-nov}x_{MS-nov} + b_{MS}x_{MS} + b_{MS-epistemology}x_{MS-epistemology}],$$

where y is the number of crossover approaches a student used, x_{MS-nov} is 1 for students in the MS-nov group and 0 otherwise, x_{MS} is 1 for students in the MS group and 0 otherwise, $x_{MS-epistemology}$ is a student's MS-epistemology score (ranging from 0 to 100), and the b 's are the associated model coefficients for the x 's.

The model is plotted in Fig. 7 and the model coefficients are shown in Table VI. The model fit for the CTRL group is only shown for MS epistemology scores from 0% of 60%, because no CTRL students score outside of this range. The coefficient for MS epistemology score is significant and positive. This supported the prediction that epistemological views favoring coherence between calculations and concepts are associated with crossover approach use. In addition, compared to CTRL students, MS students used significantly more crossover approaches even after controlling for MS epistemology score. The difference between MS-nov and CTRL groups was not significant. This indicated that even for CTRL and MS students who

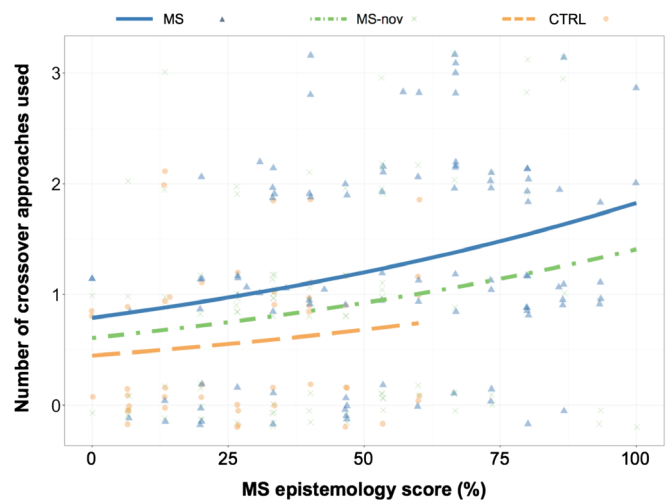


FIG. 7. Scatterplot of number of crossover approaches used vs MS epistemology score, with the model fit plotted for each instructional group.

TABLE VI. The coefficients of the general linear model with number of crossover approaches used as the dependent variable and instructional group and MS epistemology score as the independent variables.

Factor	b	SE_b	Z	p
Constant	-0.80	0.21	3.83	<0.001
CTRL
MS-nov	0.31	0.23	1.35	0.18
MS	0.57	0.23	2.51	0.01
MS epistemology score	0.0084	0.0027	3.11	0.002

received the same MS epistemology score, MS students used more crossover approaches. Said another way, MS epistemology score alone does not explain the increased use of crossover approaches by MS students. The pattern of significant results also holds when excluding all students with MS epistemology scores greater than 60%, testing only the region of overlap between all three instructional groups. A model which included an interaction between MS epistemology score and instruction group was tested, and the interaction was found to be nonsignificant.

VII. DISCUSSION

The study presented here aims to accomplish two goals. Our primary goal was to introduce calculation-concept crossover as an assessment paradigm that highlights one dimension of mathematical sense making: more expansive and coherent uses of calculations and conceptual arguments. At the heart of our paradigm is the rejection of an assumed dichotomy underlying the design of standard physics assessments: quantitative questions test calculation skill and qualitative questions test conceptual understanding. Unlike previous investigations of mathematical sense making, which relied on nonstandard problems or qualitative analysis of student discourse, the crossover assessments resemble typical physics problems and can be coded solely on the basis of students' written solutions. Importantly, the crossover assessment measurement was shown to be distinguishable from standard physics problem solving and an existing epistemological survey, indicating an instrumental benefit of calculation-crossover assessment.

The secondary goal of this study was to illustrate how different pedagogical approaches can lead to different mathematical sense making outcomes. The mathematical sense making curriculum investigated here focused on developing coherence between calculations and conceptual reasoning, as well as on developing epistemological views that support such coherence seeking. This two-pronged focus on coherence changed how students approach problem solving, increasing the use of calculation-concept crossover approaches. Overall, these crossover approaches were more accurate, efficient, and insightful than the common alternatives. Mathematical sense making instruction

also increased explicit demonstrations of coherence, providing additional evidence for the coherence-seeking outcomes of this instructional approach. Compared to the CTRL group, the predicted effects all bore out exactly for the MS group and partially for the MS-nov group, showing that even an instructor teaching a large lecture course for the first time can achieve some of the positive benefits of the mathematical sense making curriculum.

A. Considering alternative interpretations of the instructional comparison results

1. Interpreting impacts of low consent rates in the control classroom

Compared to the mathematical sense making classes, the control class had a lower consent rate (56%). Informed consent was collected from students in the control class at the beginning and end of the semester, concurrent with pre-(in-class) and post-MPEX2 (online, at-home) survey administration. How does this potential sampling bias impact the results found? We find no easy answers to this question and instead share some of our thinking on this issue. A methodological study [62] found that higher course grades predicted increased completion rates for a low-stakes assessment. Such a sampling bias could lead to under- or overestimates of the differences between the mathematical sense making and control groups (though these estimates could just as well accurately represent population differences). However, we have competing intuitions about the direction of the likely bias. On the one hand, we might expect higher-performing students to have developed a greater fluency with physics ideas, and this fluency supports seeing and adopting crossover approaches. On the other hand, success in a traditionally taught course emphasizing standard physics problems may not be associated with the inclination to use calculation-concept crossover, which is distinct from standard quantitative problem-solving methods. Responsibly interpreting the impact of this low consent rate would require us to have a better model of how course performance, course expectations, standard problem-solving skills, epistemological views, and other relevant factors interact with students' propensity for and skill with crossover approaches. The potential impact of a sampling bias is another reason to value future replications of this initial study.

2. The role of students' physics classroom expectations

One might ask how much of the differences between the control and mathematical sense making groups stems from "classroom expectation scaffolding." In the mathematical sense making classes, students learn crossover reasoning is rewarded; in class discussions and in comments on graded work, the instructors valued student use of multiple problem-solving approaches and seeking coherence between them. The grading on midterm exams

reflected this, as students could receive some credit for articulating multiple approaches, seeking coherence (or noting unresolved incoherence) between these approaches, and articulating intuitive insights, even if the final answer or approach was incorrect. On the final exam, which contained the crossover and standard problem-solving questions, we find it plausible that students in the mathematical sense making courses, on average, had a greater expectation that crossover approaches—and seeking coherence between calculations and conceptual reasoning more generally—would be rewarded.

For these reasons, we do not discount the role of classroom expectations in our results. Instead, we interpret our findings as indicating what students practice and expect to do in the mathematical sense making and control courses. That is, the findings indicate that mathematical sense making students are more likely to engage in crossover approaches *within their courses*, and these crossover approaches are more likely to yield correct answers on two of the crossover assessments—the qualitative judgment question (exploding blocks) and the cued symbolic evaluation question (ballistic pendulum). Classroom expectations aligned with the mathematical sense making curriculum’s explicit focus on (a) providing opportunities to practice constructing coherence between calculations and conceptual arguments and (b) fostering physics epistemologies that support crossover may play a role in supporting crossover use. Differences between classroom expectations in the different courses is not a confounding factor, but rather, an inextricable part of what we are measuring when we measure students’ classroom-based use of crossover approaches.

From our situated cognition perspective [63,64], the relevant question for future research is not “how much do these results indicate differences in classroom expectation rather than differences in conceptual-mathematical learning,” but rather “how does what students practice in the control or mathematical sense making classrooms inform their problem-solving practices in future educational and professional contexts?” Investigating the impacts of the control and mathematical sense making instruction on future problem-solving practice is a difficult empirical task that will, among other challenges, require an improved understanding of how student thinking emerges from complex interactions between one’s prior educational experiences and the contextual features of one’s present situation.

B. Implications for assessment in PER

1. Calculation-concept crossover: Articulating a mathematical sense making assessment goal

Better theoretical models for mathematical sense making assessment can bring instructional focus toward mathematical sense making. We present the calculation-concept crossover as a paradigm that articulates one previously underexplored dimension of mathematical sense making: expansive use of calculations and conceptual arguments

across different problem types. This crossover demonstrates several benefits of mathematical sense making. First, crossover approaches can lead to greater problem-solving accuracy when one approach is prone to error—for example, on the qualitative judgment question (exploding blocks), students who used a calculation did not fall prey to common invalid conceptual arguments. Second, crossover approaches can evidence greater efficiency and insight by breaking from typical approaches seen in physics class. However, the larger point of calculation-concept crossover is not that students should use one problem-solving approach or another. Rather, the broader benefit signaled by the paradigm is the existence of multiple, coherent problem-solving pathways. Just as engineers employ redundant systems to avert catastrophic failures, the coherence between multiple problem-solving approaches provides a safety net to catch mistakes that might be made through any one approach.

Calculation-concept crossover may be a productive lens for expanding teachers’ in-the-moment assessments of and responses to student reasoning. When students use invalid problem-solving approaches, instruction is clearly needed to attend to these errors. However, calculation-concept crossover brings to light how instruction can support mathematical sense making, even when students execute correct problem-solving approaches. When students solve quantitative problems with correct calculations, instructors could aim to introduce alternative conceptual arguments, when productive; when qualitative problems are answered correctly with conceptual arguments, instructors could similarly demonstrate how a calculation could produce similar answers. The paradigm of calculation-concept crossover can help attune teachers’ attention to the coherence between multiple approaches as a mathematical sense making-based goal of physics instruction.

The calculation-concept crossover paradigm also has implications for the design of physics problems. Many standard physics problems are best suited to one particular type of problem-solving approach. On the other hand, crossover assessments must invite productive calculations and conceptual arguments. Although we expect that some standard problems will invite students to seek coherence between different approaches and different components of their knowledge, many will not, so assessment of student sense making will require the careful design of new assessment questions.

Although we believe that it can inform the design of future assessments, it is important to note that the calculation-concept crossover paradigm is neither necessary nor sufficient for designing such problems. The instructors and researchers who designed the crossover assessments used in this study did so by drawing on their prior instructional experience and expertise with mathematical sense making, not the explicit articulation of calculation-concept crossover that was only developed later. Therefore, we do not

expect that exposing assessment developers to the calculation-concept crossover in the form of Fig. 1 alone, without that prior experience and expertise, will be sufficient to guide the development of additional assessment items. A more direct contribution of the crossover paradigm is to highlight the mathematical sense making evident in how students can approach quantitative and qualitative problems. We also believe that calculation-concept crossover could help describe and synthesize the efforts of researchers and instructors already attending to mathematical sense making.

2. The mistake in labeling qualitative questions as “conceptual questions” and quantitative questions as “calculation questions”

This study expands the discussion around how typical types of quantitative and qualitative physics assessment questions should be interpreted and designed. In PER, it is still standard (and productive) to treat quantitative questions as assessments of calculation skill and qualitative questions as assessments of conceptual knowledge. Quantitative problem-solving rubrics are often designed to assess the quality of the calculation used and are not well suited to assess purely conceptual approaches. On the other side, banks of qualitative questions are often explicitly labeled as conceptual questions, such as in the Force Concept Inventory or the Conceptual Survey of Electricity and Magnetism. In many cases, these classifications are wholly accurate. On many typical quantitative questions requiring a precise numerical or symbolic result, calculations are necessary (e.g., finding the final velocity of a block sliding down a ramp while experiencing friction, given the relevant ramp parameters). Similarly, many qualitative questions require conceptual understanding (e.g., naming the forces acting on a block sliding down a ramp while experiencing friction).

Yet, we have shown that these standard interpretations would not accurately capture students’ reasoning on our crossover assessments. On the qualitative judgment problem, the calculation approaches were more likely to yield correct answers. In these cases, it would be an error to interpret the correct answers as indicating only conceptual understanding. Similarly, on the isomorphic calculation problems, it would be an error to interpret correct answers as only indicating calculation skill, since many students reached the correct answer through a conceptual argument. These interpretations are consequential, because they lead to different instructional implications. For example, interpreting poor performance on qualitative questions as weak conceptual understanding suggests that conceptually focused instruction is needed. However, in the paradigm of calculation-concept crossover, performance can be improved—at least in some cases—by helping students see the usefulness of calculations for qualitative problems.

3. Testing calculation-concept crossover and standard problem solving: An example of multidimensional assessment

Both the experienced and novice mathematical sense making instructors’ students used more crossover approaches, gave more correct answers on crossover questions, and espoused stronger coherence-favoring epistemologies than the CTRL students. However, the MS-nov students performed significantly worse on standard quantitative problems than the CTRL students, whereas the MS students did not. One obvious question is: for the novice instructor, did the gains in mathematical sense making come at the expense of standard problem-solving skills?

This question cannot be fully addressed empirically in this study. One reason is the lack of a CTRL-nov group in our study, which would allow for an estimate of the degree to which lower instructor experience impacts standard problem-solving skills, independent of the instructional approach. Yet, even with such a group, there are many uncontrolled differences across the courses in this study that are not captured by the labels CTRL, MS-nov, and MS. Teaching is a complex practice [65] and underspecified differences between the courses on a variety of instructional dimensions—such as classroom management strategies, relationships between students and instructors, or how the instructors’ teaching styles embody personal values and beliefs—could all contribute to the differences found between course sections. Replication studies are needed to see if this pattern in learning outcomes is robust, given the possible instructional variations allowed between CTRL and MS approaches.

Although this study does not settle the question of whether the MS-nov course represents an instructional trade-off between different learning outcomes, we note that assessments investigating multiple dimensions of learning are what allow this question to be asked in the first place. Methodologically, we argue that this study embodies an approach to multidimensional assessment that is necessary for evaluating the success of instructional approaches on multiple learning goals. Even when an instructional approach shows significant gains for one learning goal, it is important to know its effect on other learning goals, whether it is positive, negative, or neutral.

Historically, PER has usually emphasized the ways in which new instructional methods can improve learning for one learning goal, such as conceptual understanding, without compromising other goals, such as quantitative problem-solving skill. For this reason, the idea of trade-offs between multiple learning goals has not been addressed. For example, in the case of active learning versus traditional lecture, meta-analytic studies have shown that active learning leads to greater learning gains along multiple course objectives, including better exam scores, better scores on concept inventories, and lower class failure rates [66]. In this case, the data suggest that active learning

approaches are strictly better than traditional lecture on a variety of educational goals. However, little is known about the comparative benefits of different active learning environments, and as PER investigates these finer-grained instructional differences, the possibility exists that promoting learning in one direction leads to trade-offs in another. At some point, instructional decisions may need to rest on decisions of value: what outcomes do I value more and what outcomes do I value less. In the case of the two types of assessments in our study, some instructors may be willing to risk sacrificing some levels of basic problem-solving competence for increased mathematical sense making; others may not. Research assessing multiple dimensions of learning can illuminate these potential debates.

Additionally, having multiple types of assessments can allow for testing of as-of-yet untested empirical questions that could inform those value judgments. For example, how do the two dimensions of problem solving investigated in this study support learning and success in future STEM courses and careers? How do these problem-solving skills (along with others) at the introductory level seed students' trajectories toward expertise? Another key question is about the effects of these different learning outcomes on retention and persistence—in physics specifically and STEM fields more generally. Students wanting to engage in key disciplinary sense making practices, such as coherence seeking and mathematical sense making, may lose interest in STEM domains if they are primarily training on more routine problem-solving competencies [67]. In this way, early exposure to mathematical sense making may prove to be more valuable in the long run to students' educational trajectories. Longitudinal hypotheses like these often go untested, because they require methodological power, serious time investment, and, as we argue, multiple types of assessment.

Although any one assessment highlights a dimension of learning, relying too heavily on any single assessment may obscure potential learning along other dimensions. In addition to assessing the potential learning benefits of an instructional approach, researchers should also seek to accurately assess potential trade-offs. This will demand a nuanced look at instructional comparisons and require multidimensional assessments. The potential payoff will be better methodological tools and empirical data for making informed decisions about instructional goals.

C. Mathematical sense making and adaptive expertise

At a broader level, this work takes a step towards understanding adaptive expertise in physics education. Hatano and Inagaki [68] distinguished routine and adaptive expertise: while routine expertise involves using standard approaches in familiar situations, adaptive expertise allows people to find new solutions to new problems.

This adaptation can involve modification of known procedures or invention of novel approaches. We contend that calculation-concept crossover marks adaptive expertise. On our crossover assessments, students could break from standard approaches (i.e., calculations on quantitative problems or conceptual reasoning on qualitative problems) to find more efficient, effective, and insightful solutions. Although our crossover assessments are not so far from the typical problem space of introductory physics, we believe that students using crossover approaches here demonstrate adaptability and flexibility that could forecast success in adapting to new problems in the future.

Schwartz, Bransford, and Sears [69] broke adaptive expertise into two components: efficiency and innovation (Fig. 8). Importantly, while both routine and adaptive experts can behave efficiently in familiar settings, it is innovation that differentiates these two courses of expertise. Considering these two dimensions together, Schwartz, Bransford, and Sears argue that both efficiency and innovation should proceed together in the development of adaptive expertise, and they argue that a focus on just one or the other will not be as successful at helping students transfer their knowledge to new situations. They hypothesize an optimal adaptability corridor that balances both efficiency and innovation in instruction.

We can use this efficiency-and-innovation framework to interpret our study in two ways. In terms of instruction, we propose that the promotion of student sense making in the lecture provided important innovation experiences in mathematical sense making classrooms (see Ref. [59] for an example). These opportunities to invent and be innovative are a key part of the mathematical sense making curriculum and aim to foster the skills, experience, and dispositions students

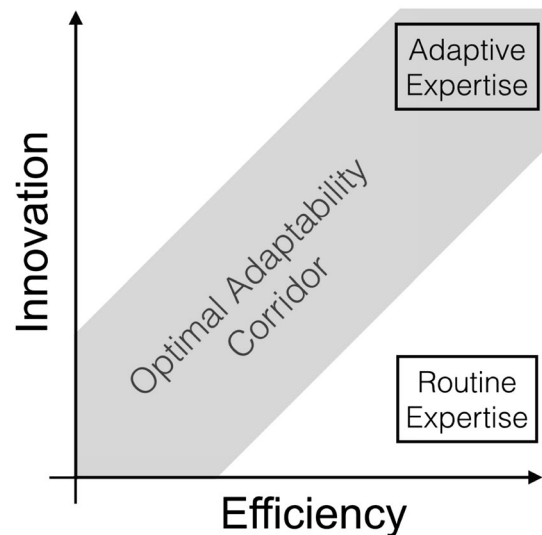


FIG. 8. Two courses of expertise plotted on a 2D space of innovation vs efficiency (adapted from Schwartz, Bransford, and Sears [69]).

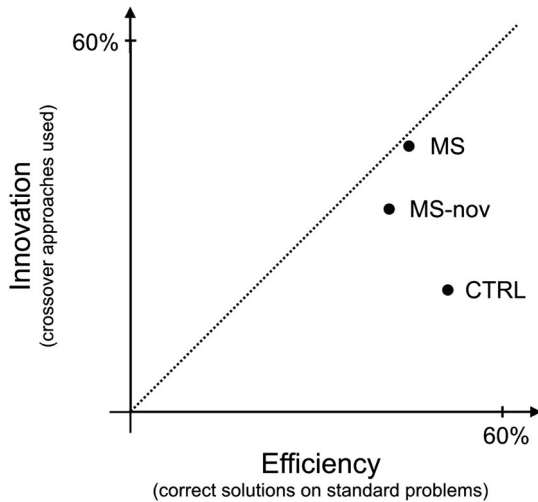


FIG. 9. Plot of the study results by innovation vs efficiency. The dotted line represents a balance of equal levels of efficiency and innovation. The MS student average is closest to this line whereas the CTRL student average is furthest from this line.

will need to be innovative in future settings. While standard, lecture-based instructional approaches seem to focus on efficiency, many active-engagement instructional approaches balance content learning with opportunities for student innovation and invention. This suggests that many existing PER-based instructional methods may offer opportunities to develop adaptive expertise, and it would be an interesting direction for future research to investigate how these existing instructional approaches might help foster that expertise.

In terms of the learning outcomes of this study, we can map standard problem-solving accuracy as being a measure of routine efficiency and crossover approach use as being a measure of innovation (i.e., breaking from more standard approaches and choosing to use a productive crossover approach). Figure 9 plots this mapping for the three instructional groups in our study, where the dotted line represents equal success on efficiency and innovation measures. Although this mapping should not be taken too seriously since the exact percentages depend as much on the comparative difficulty of the standard and crossover assessment problems as they do on students' problem-solving skill, this cartoon provides a starting point for thinking about educational possibilities in introductory physics.

The results of our study may serve as a counterexample to the common idea that training for adaptive expertise can only occur after a sufficient amount of routine expertise is developed. This trajectory is embodied by introductory teaching that focuses on basic skill development, testing for their efficient use on familiar problem types, and saving adaptivity and innovation for future courses that can leverage this earlier instruction. In this trajectory, the introductory courses would teach the basic skills, and the

upper-division courses would later provide opportunities to evolve those basic skills into the adaptive skills that constitute “thinking like a physicist.” Results from the control classroom in our study illustrate the expected outcomes of this “efficiency-before-innovation” model. By contrast, the results of the mathematical sense making instruction may indicate that, even in introductory courses, aiming to balance efficiency and innovation can be fruitful. The MS group showed higher levels of crossover approach use than the CTRL group while demonstrating a comparable level of performance on the standard problems. Here, the mathematical sense making instruction illustrates the possibility of effectively developing “efficiency alongside innovation.”

It might also be hypothesized that the ability to develop efficiency alongside innovation relies on instructor experience, but the performance of the MS-nov students serves as a possible counterexample; namely, the MS-nov group may indicate that even first-time large-lecture instructors may be able to balance efficiency with innovation to some degree. This is indicated by the shorter distance of the MS-nov group from the dotted line in Fig. 9, representing equal levels of efficiency and innovation, compared to the CTRL group.

In this way, mathematical sense making may be able to provide additional insight into the development of adaptive expertise. The results from this study might be interpreted to suggest that, when aiming to develop adaptive expertise, training for efficiency before innovation may be taking the long way around. Efficiency before innovation may miss the opportunity for invention skill to develop at the same time as skills for efficiency. Focusing on efficiency with familiar routines may also result in learners relying on these routines even in the face of new opportunities for innovation and novel exploration [70], as they did for students who did not leverage the isomorphism between problems on the isomorphic calculation questions. Attempts to help students learn the standard procedures may even cue students to avoid nonstandard methods and, inadvertently, suppress the search for new, more efficient and elegant approaches [71], raising potential barriers to future innovation. Linking mathematical sense making with adaptive expertise in future research could provide additional data to clarify the risks and opportunities associated with traversing different trajectories toward physics expertise.

VIII. CONCLUSION

One goal for physics instruction is to foster the mathematical sense making that students will need in their educational and professional futures. Success with this goal requires both innovative instructional approaches for teaching mathematical sense making and innovative assessments for measuring the success of that instruction.

We offer calculation-concept crossover as one assessment candidate that might be used in assessing and further developing mathematical sense making-focused instructional approaches, such as the one examined in this study.

ACKNOWLEDGMENTS

The research project described here was supported by NSF EEC-0835880.

-
- [1] J. Clement, Creative model construction in scientists and students: The role of imagery, in *Analogy, and Mental Simulation* (Springer Netherlands, 2008).
- [2] J. Gainsburg, The mathematical modeling of structural engineers, *Math. Think. Learn.* **8**, 3 (2006).
- [3] N. J. Nersessian, *How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science in Cognitive Models of Science*, edited by R. Giere (University of Minnesota Press, Minneapolis, MN, 1992), pp. 3–44.
- [4] O. Bueno, Dirac and the dispensability of mathematics, *Stud. History Philosophy Sci. Part B* **36**, 465 (2005).
- [5] M. B. Kustusch, D. Roundy, T. Dray, and C. A. Manogue, Partial derivative games in thermodynamics: A cognitive task analysis, *Phys. Rev. ST Phys. Educ. Res.* **10**, 010101 (2014).
- [6] L. Branchetti, A. Cattabriga, and O. Levrini, Interplay between mathematics and physics to catch the nature of a scientific breakthrough: The case of the blackbody, *Phys. Rev. Phys. Educ. Res.* **15**, 020130 (2019).
- [7] A. Gupta and A. Elby, Beyond epistemological deficits: Dynamic explanations of engineering students' difficulties with mathematical sense-making, *Int. J. Sci. Educ.* **33**, 2463 (2011).
- [8] S. Kapon and M. Schwartz, Nurturing sense making of, through, and with a mathematical model, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC*, edited by A. Traxler, Y. Cao, and S. Wolf (AIP, New York, 2018).
- [9] B. W. Dreyfus, A. Elby, A. Gupta, and E. R. Sohr, Mathematical sense-making in quantum mechanics: An initial peek, *Phys. Rev. Phys. Educ. Res.* **13**, 020141 (2017).
- [10] J. Gifford and N. Finkelstein, Categorizing mathematical sense making and an example of how physics understanding can support mathematical understanding, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT*, edited by Y. Cao, S. Wolf, and M. B. Bennett (AIP, New York, 2019).
- [11] M. M. Hull, E. Kuo, A. Gupta, and A. Elby, Problem-solving rubrics revisited: Attending to the blending of informal conceptual and formal mathematical reasoning, *Phys. Rev. ST Phys. Educ. Res.* **9**, 010105 (2013).
- [12] B. L. Sherin, How students understand physics equations, *Cognit. Instr.* **19**, 479 (2001).
- [13] B. L. Sherin, Common sense clarified: The role of intuitive knowledge in physics problem solving, *J. Res. Sci. Teach.* **43**, 535 (2006).
- [14] E. F. Redish and E. Kuo, Language of physics, language of math: Disciplinary culture and dynamic epistemology, *Sci. Educ.* **24**, 561 (2015).
- [15] E. Kuo, M. M. Hull, A. Gupta, and A. Elby, How students blend conceptual and formal mathematical reasoning in solving physics problems, *Sci. Educ.* **97**, 32 (2013).
- [16] M. Eichenlaub and E. F. Redish, *Blending Physical Knowledge with Mathematical Form in Physics Problem Solving in Mathematics in Physics Education*, edited by G. Pospiech, M. Michelini, and B.-S. Eylon (Springer International Publishing, Cham, 2019), pp. 127–151.
- [17] T. J. Bing and E. F. Redish, Analyzing problem solving using math in physics: Epistemological framing via warrants, *Phys. Rev. ST Phys. Educ. Res.* **5**, 020108 (2009).
- [18] 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).
- [19] S. Brahmia, A. Boudreaux, and S. E. Kanim, Obstacles to mathematization in introductory physics, [arXiv:1601.01235](https://arxiv.org/abs/1601.01235).
- [20] K. Hahn, P. Emigh, M. Lenz, and E. Gire, Student sense-making on homework in a sophomore mechanics course, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH*, edited by L. Ding, A. Traxler, and Y. Cao (AIP, New York, 2017).
- [21] M. Lenz, P. Emigh, and E. Gire, Surprise! students don't do special-case analysis when unaware of it, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC*, edited by A. Traxler, Y. Cao, and S. Wolf (AIP, New York, 2018).
- [22] P. Emigh, J. Alfson, and E. Gire, Student sense making about equipotential graphs, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC*, edited by A. Traxler, Y. Cao, and S. Wolf (AIP, New York, 2018).
- [23] T. Sikorski, G. White, and J. Landay, Uptake of solution checks by undergraduate physics students, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH*, edited by L. Ding, A. Traxler, and Y. Cao (AIP, New York, 2017).
- [24] S. W. Brahmia, A. Olsho, T. I. Smith, and A. Boudreaux, A framework for the natures of negativity in introductory physics, [arXiv:1903.03806](https://arxiv.org/abs/1903.03806).
- [25] L. C. McDermott, Orsted medal lecture 2001: "Physics Education Research—the Key to Student Learning," *Am. J. Phys.* **69**, 1127 (2001).
- [26] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).
- [27] B. Thacker, E. Kim, K. Trefz, and S. M. Lea, Comparing problem solving performance of physics students in inquiry-based and traditional introductory physics courses, *Am. J. Phys.* **62**, 627 (1998).

- [28] E. Kim and S. J. Pak, Students do not overcome conceptual difficulties after solving 1000 traditional problems, *Am. J. Phys.* **70**, 759 (2002).
- [29] D. L. Schwartz, T. Martin, and J. Pfaffman, How mathematics propels the development of physical knowledge, *J. Cognit. Dev.* **6**, 65 (2005).
- [30] C. Singh, Assessing student expertise in introductory physics with isomorphic problems. II. Effect of some potential factors on problem solving and transfer, *Phys. Rev. ST Phys. Educ. Res.* **4**, 010105 (2008).
- [31] J. L. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. A. Jackson, A. Mason, Q. X. Ryan, and J. Yang, Assessing student written problem solutions: A problem-solving rubric with application to introductory physics, *Phys. Rev. Phys. Educ. Res.* **12**, 010130 (2016).
- [32] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, *Phys. Teach.* **30**, 141 (1992).
- [33] D. P. Maloney, T. L. O’Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students’ conceptual knowledge of electricity and magnetism, *Am. J. Phys.* **69**, S12 (2001).
- [34] R. K. Thornton and D. R. Sokoloff, Assessing student learning of newton’s laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, *Am. J. Phys.* **66**, 338 (1998).
- [35] J. I. Heller and F. Reif, Prescribing effective human problem-solving processes: Problem description in physics, *Cognit. Instr.* **1**, 177 (1984).
- [36] R. J. Dufresne, W. J. Gerace, P. T. Hardiman, and J. P. Mestre, Constraining novices to perform expertlike problem analyses: Effects on schema acquisition, *J. Learn. Sci.* **2**, 307 (1992).
- [37] R. Mualem and B. S. Eylon, Junior high school physics: Using a qualitative strategy for successful problem solving, *J. Res. Sci. Teach.* **47**, 1094 (2010).
- [38] T. J. Bing and E. F. Redish, Epistemic complexity and the journeyman-expert transition, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010105 (2012).
- [39] A. A. diSessa, A “theory bite” on the meaning of scientific inquiry: A companion to Kuhn and Pease, *Cognit. Instr.* **26**, 560 (2008).
- [40] O. Uhden, R. Karam, M. Pietrocola, and G. Pospiech, Modelling mathematical reasoning in physics education, *Sci. Educ.* **21**, 485 (2012).
- [41] J. H. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Expert and novice performance in solving physics problems, *Science* **208**, 1335 (1980).
- [42] L. N. Walsh, R. G. Howard, and B. Bowe, Phenomenographic study of students’ problem solving approaches in physics, *Phys. Rev. ST Phys. Educ. Res.* **3**, 020108 (2007).
- [43] P. V. Engelhardt and R. J. Beichner, Students’ understanding of direct current resistive electrical circuits, *Am. J. Phys.* **72**, 98 (2004).
- [44] L. C. McDermott and P. S. Shaffer, Research as a guide for curriculum development: An example from introductory electricity. Part I: Investigation of student understanding, *Am. J. Phys.* **60**, 994 (1992).
- [45] T. O. Pride, S. Vokos, and L. C. McDermott, The challenge of matching learning assessments to teaching goals: An example from the work-energy and impulse-momentum theorems, *Am. J. Phys.* **66**, 147 (1998).
- [46] B. W. Frank, S. E. Kanim, and L. S. Gomez, Accounting for variability in student responses to motion questions, *Phys. Rev. ST Phys. Educ. Res.* **4**, 020102 (2008).
- [47] M. E. Loverude, C. H. Kautz, and P. R. Heron, Helping students develop an understanding of archimedes’ principle. I. Research on student understanding, *Am. J. Phys.* **71**, 1178 (2003).
- [48] C. H. Kautz, P. R. L. Heron, P. S. Shaffer, and L. C. McDermott, Student understanding of the ideal gas law, Part II: A microscopic perspective, *Am. J. Phys.* **73**, 1064 (2005).
- [49] B. A. Lindsey, P. R. L. Heron, and P. S. Shaffer, Student ability to apply the concepts of work and energy to extended systems, *Am. J. Phys.* **77**, 999 (2009).
- [50] M. Kryjevskaja, M. R. Stetzer, and N. Grosz, Answer first: Applying the heuristic-analytic theory of reasoning to examine student intuitive thinking in the context of physics, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020109 (2014).
- [51] N. Weinlader, E. Kuo, B. Rottman, and T. Nokes-Malach, A new approach for uncovering student resources with multiple-choice questions, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT*, edited by Y. Cao, S. Wolf, and M. B. Bennett (AIP, New York, 2019).
- [52] R. A. Lawson and L. C. McDermott, Student understanding of the work-energy and impulse-momentum theorems, *Am. J. Phys.* **55**, 811 (1987).
- [53] M. Wertheimer, *Productive Thinking* (Harper and Row, New York, 1959).
- [54] J. T. Shemwell, C. C. Chase, and D. L. Schwartz, Seeking the general explanation: A test of inductive activities for learning and transfer, *J. Res. Sci. Teach.* **52**, 58 (2015).
- [55] D. Hammer, Epistemological beliefs in introductory physics, *Cognit. Instr.* **12**, 151 (1994).
- [56] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.* **66**, 64 (1998).
- [57] J. E. Caldwell, Clickers in the large classroom: Current research and best-practice tips, *CBE Life Sci. Educ.* **6**, 9 (2007).
- [58] E. F. Redish and D. Hammer, Reinventing college physics for biologists: Explicating an epistemological curriculum, *Am. J. Phys.* **77**, 629 (2009).
- [59] Two Elements of a Mathematical Sensemaking Approach to Teaching Introductory Physics, <http://hdl.handle.net/2142/107789>.
- [60] T. L. McCaskey, Comparing and contrasting different methods for probing student epistemology and epistemological development in introductory physics, Ph.D. thesis, University of Maryland, 2009, <https://drum.lib.umd.edu/handle/1903/9824>.
- [61] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.16.020109> for the 15 survey items used to construct the MS epistemology score and a more detailed discussion of the crossover assessment coding scheme, including examples of student work and details on how disagreements between coders were resolved.
- [62] J. M. Nissen, M. Jariwala, E. C. Close, and B. Van Dusen, Participation and performance on paper- and computer-based low-stakes assessments, *Int. J. STEM Educ.* **5**, 28 (2018).

- [63] J. S. Brown, A. Collins, and P. Duguid, Situated cognition and the culture of learning, *Educ. Res.* **18**, 32 (1989).
- [64] J. G. Greeno, A. Collins, and L. B. Resnick, *Cognition and Learning in Handbook of Educational Psychology*, edited by D. C. Berliner and R. C. Calfee (Simon & Schuster Macmillan, New York, 1996), pp. 15–46.
- [65] L. Shulman, Knowledge and teaching: Foundations of the new reform, *Harv. Educ. Rev.* **57**, 1 (1987).
- [66] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8410 (2014).
- [67] B. A. Danielak, A. Gupta, and A. Elby, Marginalized identities of sense-makers: Reframing engineering student retention, *J. Eng. Educ.* **103**, 8 (2014).
- [68] G. Hatano and K. Inagaki, *Two Courses of Expertise in Child Development and Education in Japan*, edited by H. Stevenson, H. Azuma, and K. Hakuta (Freeman, New York, 1986), pp. 262–272.
- [69] D. L. Schwartz, J. D. Bransford, and D. Sears, *Efficiency and Innovation in Transfer in Transfer of Learning from a Modern Multidisciplinary Perspective*, edited by J. P. Mestre (Information Age Publishing, Greenwich, CT, 2005), pp. 1–51.
- [70] D. L. Schwartz, C. C. Chase, and J. D. Bransford, Resisting overzealous transfer: Coordinating previously successful routines with needs for new learning, *Educ. Psychol.* **47**, 204 (2012).
- [71] E. Kuo, N. R. Hallinen, and L. D. Conlin, When procedures discourage insight: epistemological consequences of prompting novice physics students to construct force diagrams, *Int. J. Sci. Educ.* **39**, 814 (2017).