Components of the preparation gap for physics learning vary in two learner groups

Anita Delahay,[1] Marsha Lovett,[2,3] David Anderson,[4] and Surajit Sen[5]

[1]*College of Professional Studies, Northeastern University, Boston, Massachusetts 02115, USA*
[2]*Vice Provost for Teaching and Learning Innovation, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA*
[3]*Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA*
[4]*Department of Physics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA*
[5]*Department of Physics, State University of New York at Buffalo, Buffalo, New York 14260, USA*

(Received 9 December 2022; accepted 23 June 2023; published 31 August 2023)

The preparation gap [Salehi *et al.*, Phys. Rev. Phys. Educ. Res. **15**, 020114 (2019)] refers to gaps in students' prior knowledge that can negatively affect their learning as they engage in introductory physics courses. To better characterize the gap, the current study distinguished the impact of various prior knowledge components on learning gains. Measured components came from within the course domain (e.g., energy and force, angular kinematics) and outside it (e.g., algebra, vectors, calculus, and scientific reasoning). We conducted the study in two different institutional contexts: An algebra-based course offered at a Northeastern State University (NESU) and a calculus-based course offered at a Midwestern Private University (MWPU). Furthermore, we defined three levels of physics learning outcome measures with increasing difficulty. Multiple regression analysis was used to predict learning gains with the various prior knowledge components as predictor variables. The results indicate that greater prior knowledge from both within and outside the domain predicted higher learning gains and explained 30%–50% of the variance in outcome measures. Predictive, in-domain prior knowledge was the same for both groups—i.e., prior knowledge of energy and force, as measured by the Mechanics Baseline Test [Hestenes and Wells, Phys. Teach. **30**, 159 (1992)]. Predictive, outside-domain prior knowledge differed between the groups. Better scientific reasoning was highly predictive of learning in the NESU (algebra-based) group but did not predict learning in the MWPU (calculus-based) group. Math prior knowledge predicted learning in both groups, although different topics within the math domain. These results suggest that measuring distinguishable components of prior knowledge will better characterize the preparation gap in ways that can be informative to educators. Specifically, measuring multiple, distinct types of prior knowledge can indicate *which types* are leading to a preparation gap for some students, putting them at a disadvantage for learning, whereas measuring a single type of prior knowledge or measuring prior knowledge too coarsely (without distinguishing among types) cannot provide sufficient diagnostic power.

DOI: 10.1103/PhysRevPhysEducRes.19.020122

I. INTRODUCTION

College professors of physics may notice that students enter their courses with different types and amounts of prior knowledge, not only in the target domain of instruction (physics) but also in related knowledge areas, such as mathematics [1–4]. When prior knowledge is missing, researchers have termed this phenomenon a *preparation gap* [5] that can hinder success in introductory, undergraduate physics courses. The goal of this work is to further define and measure gap components that could be addressed by educators or the broader educational system to improve learning gains in these courses.

Prior knowledge has long been identified as a key predictor of learning and performance [6–9]. Studies suggest that prior knowledge within the target domain affects comprehension and retrieval [10–12], concept learning [13,14], category membership inferences [15], use of strategies related to learning and studying [16,17], and motivation [18]. Prior knowledge plays a special role in problem-solving, improving the perception of critical (vs surface) problem features [19] and allowing the solver to modify a familiar solution for use in the present problem [20]. These findings lead us to expect that domain prior knowledge (DPK)—i.e., prior knowledge in the domain of instruction—should be significantly, positively predictive of *learning*, defined as a positive change in knowledge over a specific period of time due to one or more learner-centered actions or behaviors, in introductory courses.

In practice, the benefit of DPK on learning has been variable when assessing novice learners in classroom-based studies. A recent meta-analysis [21] based on 493 studies found that the average effect of DPK on gain scores (i.e., learning) was slightly negative and not significantly different from zero. The researchers contrasted this result with the positive correlation they found between DPK and post-test performance. They attribute the positive correlation between pretest and post-test to the influence of past achievement on performance [8] and reiterate that this result does not show prior knowledge confers benefit to learning.

These findings echo Hake's [22] study of 62 course-level samples in physics that found a positive correlation ($r = 0.55$) between pretest and post-test scores, which he argued would occur in the absence of any instruction and was likely attributable to the stability of individual differences in performance. Indeed, when Hake measured the correlation of pretest scores to absolute gain (post-pre) or normalized gain (post-pre/100%-pre) scores, which reflect changes in performance, results ranged from negative to practically zero, suggesting no significant effect of DPK on learning. Both the meta-analytical study [21] and Hake's work point to the importance of correlating pretest scores with gains scores when determining the role of prior knowledge in learning, rather than simply finding the correlation of scores from pretest to post-test.

Coletta and Phillips [23] reexamined Hake's [22] data and found if they used (a) Hake's entire dataset [including traditional and interactive engagement (IE) classes], there was no significant correlation between pretest scores and normalized gain; (b) Hake's IE classes only, some correlation existed ($r = 0.25,\ p = 0.1$); and (c) Hake's IE classes combined with their own IE classes (more than doubling the number to 73 total), there was a larger correlation ($r = 0.39,\ p = 0.0006$). Coletta and Philips' results suggest that DPK may not predict learning in every classroom. Importantly, they also found that DPK was not the only, nor the largest, source of meaningful prior knowledge variation that impacted physics learning.

Coletta and Phillips [23] measured both in-domain, physics knowledge and *outside-domain* prior knowledge of scientific reasoning and found that variation in the outside-domain prior knowledge was an especially potent predictor of learning. Other researchers have also measured DPK and outside-domain prior knowledge of algebra [2,3]. Similar to Coletta and Phillips' results when they used a scientific reasoning measure, algebra predicted physics learning to a greater degree than starting knowledge of physics. Consistently, outside-domain prior knowledge has been highly predictive of physics learning in novices.

We refer to this outside-domain prior knowledge as *ancillary prior knowledge* (APK), which is prior knowledge of concepts and skills that are outside the domain of instruction and yet regularly applied when learning concepts and skills within the domain [24]. For example, knowledge of algebra and vector arithmetic (from the domain of mathematics) would be APK when students are learning to solve problems using Newton's laws of motion (in the target domain of physics) because knowledge of algebra and vectors is required for solving these types of problems. The word *ancillary* denotes necessary support to the primary activity—i.e., learning in the target domain—while falling outside the lesson's domain.

The key finding that motivated this study is that gaps in ancillary knowledge can hinder performance in domain-related learning. This finding has been widespread. The APK construct has been applied in studies occurring in a broad range of learning domains. In each of these studies, weaker APK related to less learning, e.g., in chemistry [25–27], engineering [28], biology [29], economics [30], psychology [24], and geosciences [31].

Gaps in APK are perhaps more concerning than gaps in physics knowledge for students in introductory physics courses, because domain knowledge is the deliberate focus of instruction, whereas APK often is not. Some APK (e.g., a lesson on vector math) may be included in the instruction, but such ancillary knowledge is rarely systematically measured or taught. Therefore, variability in the untaught APK areas (i.e., knowledge that is often not addressed in the lesson) is an *unaddressed* source of variance in task performance and a potential source of task failure.

Ancillary knowledge that is deemed necessary for domain learning could be identified prior to domain instruction and provide valuable information to educators. Instructors may find it hard to recognize students' missing ancillary knowledge during normal domain instruction because it can be hard to distinguish the sources of task failure: missing APK vs missing DPK vs both. After identifying such APK gaps, however, remediation has proven an effective instructional strategy, e.g., remediating gaps in mathematics knowledge has proven effective in the context of physics and engineering instruction [28,32,33].

The prior studies providing evidence that variation in APK predicts learning gains usually contrast only one type of APK with DPK vs multiple types of APK, cf. Hudson and Liberman [34]. Moreover, APK and DPK are typically measured monolithically using a single pretest, instead of measuring components of PK that may draw on multiple pretest measures. By contrast, our approach is to measure variability in students' starting knowledge of all relevant sources of DPK (e.g., energy, force, and angular kinematics) and APK (e.g., algebra, vectors, calculus, and scientific reasoning). One goal of measuring and statistically controlling for multiple, instruction-relevant APK and DPK precursors of domain learning is to reduce the likelihood of biased or spurious correlations [35], i.e., cases where variation in one type of prior knowledge appears related to learning gains but is also related to an unmeasured type of prior knowledge that is the actual predictor of learning.

Furthermore, only a few prior studies have considered how DPK and APK predict learning when considering differences in groups' characteristic prior knowledge. Researchers have contrasted groups based on institution selectivity [5,23], science, technology, engineering, and mathematics (STEM) vs non-STEM majors [36] or general vs advanced course level [3]. Such studies in differently characterized samples are limited and, therefore, cross-sample conclusions are still emerging.

Moreover, studies comparing differently characterized groups of learners have included different mixes of PK and/or prior achievement (PA) variables, so a direct comparison is difficult. It is important to note that PA differs from PK in several ways. PA measures (e.g., SAT, ACT, IQ, and GPA), by design, are coarse-grained measures of participants' accumulated learning and may reflect the influence of additional constructs (e.g., motivation, socioeconomic status, and intelligence) leading to confounding among these constructs. PA is often measured by college readiness assessments, aptitude assessments, or past performance and is highly predictive of learning and performance ($d = 0.67$) [8]. PK measures, on the other hand, generally align more closely with the target tasks or lessons and therefore have the potential to be much finer-grained measures of knowledge (and, often, more recent).

Despite the small number of studies that have used differently characterized samples, differences in the factors that predict learning have been found between samples. Coletta and Phillips [23] administered a physics concept inventory (CI) at pretest and correlated these results to normalized gains on the same CI at post-test at four institutions: one highly selective and three less selective. A subset of students at one less selective institution also took the Classroom Test of Scientific Reasoning (CTSR) [37], a measure of APK scientific reasoning. The researchers found a correlation between individual students' DPK and their normalized CI gains—and an even stronger relationship between APK scientific reasoning and normalized CI gains for the subset that took the CTSR—at the three less selective universities, showing a role for both DPK and APK in learning. However, there was no relationship of DPK to normalized gains at the highly selective university (APK scientific reasoning was not measured in this group).

The authors suggested that DPK was not correlated with CI gains for students at the highly selective university due to the measure not being sensitive enough (i.e., it did not pick up variability within the sample). This highlights the importance of having measures sensitive to variation in the PK *of a given sample*. They also suggested that APK scientific reasoning and/or SAT scores likely would have been high for the highly selective university group and reasoned this "hidden variable" would have correlated with learning. However, given that uniformly high scores on the DPK measure did not correlate with gains, we question whether a different measure on which students scored uniformly high—APK scientific reasoning or PA—would have shown an association with learning. For the current study, we selected measures based on the assumption that a preparation gap will be more detectable by PK measures with the most, rather than least, within-sample variability.

Nakakoji and Wilson [3] evaluated direct and indirect effects of previous math grades and a PA measure on final grades in a lower-level and upper-level physics course at the same institution. With less advanced students, PA both directly affected physics grades and indirectly affected math grades, which in turn affected the physics grade. In more advanced students, PA had no direct effect on physics grade and only had an indirect effect on physics grade, mediated through the math grade. Therefore, PA's direct influence was found with lower-level, but not higher-level, students, whereas the influence of APK math on physics learning was found in both groups.

Salehi *et al.* [5] assessed demographic factors, PA (SAT/ACT Math), and DPK at three universities with different selectivity levels. PA and DPK significantly predicted exam scores in each sample. Effects of DPK were highly consistent across institutions, whereas effects of PA were strong but not as uniform. Smaller or larger correlation PA coefficients were not clearly aligned to the degree of institutional selectivity. As these three studies [3,5,23] show, additional empirical work is needed to develop an understanding of the effects of PK (APK vs DPK) vs PA on learning in differently characterized samples, and indeed how these measures of correlated knowledge types can best distinguish unique aspects of PK or PA that are most related to preparation for physics learning.

In the current study, we included students who were enrolled in introductory physics courses at the postsecondary level. From this population, we selected two samples that differed in important ways. Sample 1 consisted of STEM and non-STEM majors enrolled at a large, selective northeastern state university (NESU) in an algebra-based introductory course, with an instructional focus on concepts and a relatively lesser focus on solving problems. Sample 2 was almost exclusively STEM majors enrolled at a highly selective, midwestern private university (MWPU) in a calculus-based introductory course, with an instructional focus on both concepts and problem-solving.

It is likely that based on the type of course and the nature of the institution and student majors, a given sample's PK profile will differ—students may have little DPK but stronger APK, be relatively strong in both, or be relatively weak in both—and therefore the effect of PK on learning will differ. Examining such variability in starting levels of DPK *and* APK and identifying how particular PK profiles affect learning should bring a clearer understanding of the nature of the preparation gap in various learner groups that leads to recommendations for remedying prior knowledge gaps associated with these profiles.

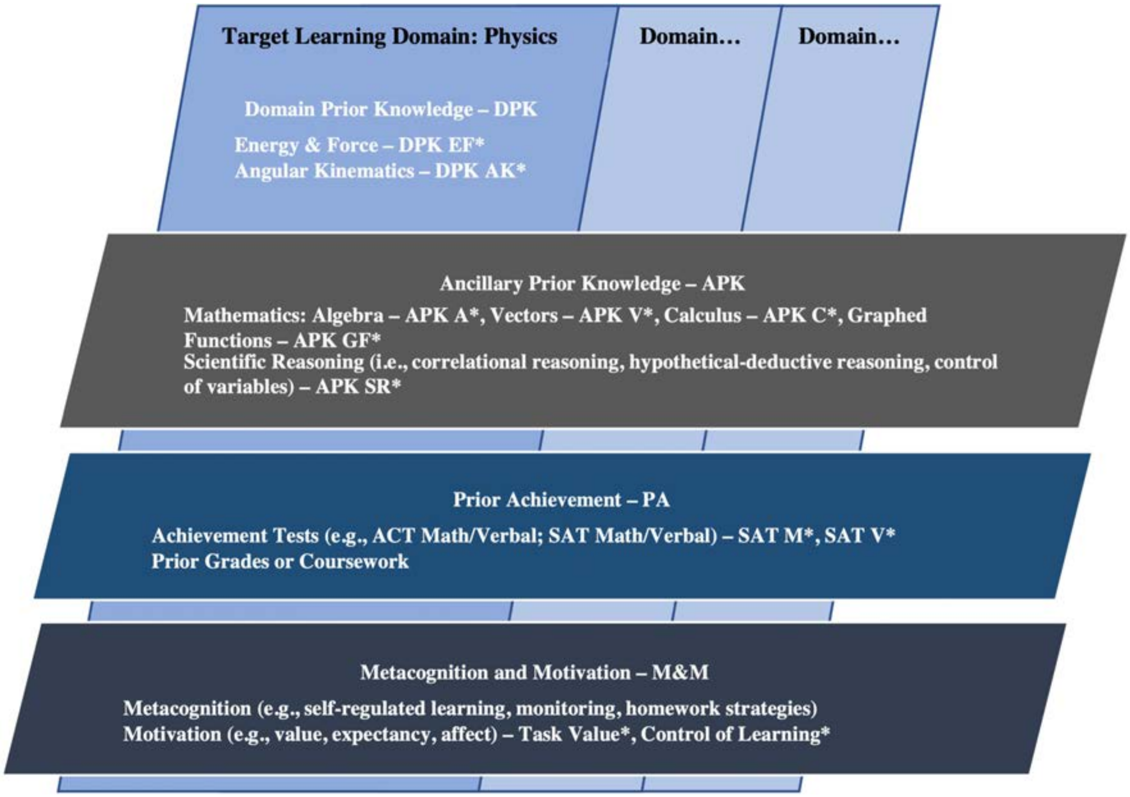
Target Learning Domain: Physics
Domain…
Domain…
Domain Prior Knowledge – DPK
Energy & Force – DPK EF*
Angular Kinematics – DPK AK*
Ancillary Prior Knowledge – APK
Mathematics: Algebra – APK A*, Vectors – APK V*, Calculus – APK C*, Graphed Functions – APK GF*
Scientific Reasoning (i.e., correlational reasoning, hypothetical-deductive reasoning, control of variables) – APK SR*
Prior Achievement – PA
Achievement Tests (e.g., ACT Math/Verbal; SAT Math/Verbal) – SAT M*, SAT V*
Prior Grades or Coursework
Metacognition and Motivation – M&M
Metacognition (e.g., self-regulated learning, monitoring, homework strategies)
Motivation (e.g., value, expectancy, affect) – Task Value*, Control of Learning*

FIG. 1. Four categories of variables that could impact learning in the physics domain: domain prior knowledge (DPK), ancillary prior knowledge (APK), prior achievement or aptitude (PA), and metacognition and motivation (M&M). The specific variables included are noted with an asterisk (*). The horizontal blocks for APK, PA, and M&M signify they have the potential to apply across domains, i.e., in more than one domain, but, importantly, are not general in the sense that each would apply indiscriminately for every domain.

A. The multivariate model

Figure 1 depicts four categories of variables that theory and/or research have suggested affect domain learning: domain prior knowledge (DPK), ancillary prior knowledge (APK), prior achievement or aptitude (PA), and metacognition and motivation (M&M). We focused on these groups of variables because the first three, DPK, APK, and PA, are often placed under the umbrella of PK measures, and the fourth, M&M, was intended to capture any meaningful differences in motivation given key differences in the samples and courses. DPK, PA, and M&M are captured in theories on domain learning, whereas APK has been left out of many theoretical discussions, and yet empirical work has pointed to its role in learning. We included variables from all four categories to determine which would predict learning in the physics domain using multivariate analyses.

While we were comprehensive in our search for potentially relevant PK for the DPK and APK measures, we selected limited measures from PA and M&M. Furthermore, this model does not represent all potentially relevant variables for physics learning in novices, such as background and demographic variables. See Docktor and Mestre [38] for a synthesis of the research approaches and factors to consider in studies of physics learning.

B. Defining and measuring prior knowledge

Dochy and Alexander [39] defined prior knowledge (PK) as all of a person's knowledge available before a given task. This definition is sufficiently broad to cover all of the types of knowledge that may prove useful to a person approaching a new task. However, the breadth of this definition makes it more challenging to operationalize—i.e., establish necessary and sufficient measures of—PK and study its effects on learning.

In this work, we (a) bring attention to a relatively new distinction in PK measures (i.e., APK vs DPK), (b) focus on the content areas of PK (e.g., physics, mathematics) where lack of discrimination between APK and DPK may be particularly problematic, and (c) include as many sources of potential variation in PK as possible, many of which we expect are correlated but nevertheless explain unique variance in learning. At a higher level, these goals aim to address the undesirable metapractices of insufficient measures of PK and ambiguity in the definition of distinct aspects of the PK construct.

1. DPK measures

Researchers who study physics learning in classroom settings often use concept inventories (CIs, *see* http://physport.org [40]) that measure knowledge of the broad domain and/or specific subdomain(s). Frequently used measures include the Force Concept Inventory (FCI) [2,5,22,23] and the Force Motion Conceptual Evaluation (FMCE) [1,2,4,5].

CIs can be used effectively as pretests (i.e., they are often understandable to the novice) and post-tests [22] given at the start and end of an academic term that provides instruction in all or most of the tested topics. CI use facilitates comparisons across institutions, terms, and instruction [41]. Moreover, CIs seem to capture much of the variation in preparation that could be attributed to other background factors, such as the type of high school physics class taken and grades achieved in these classes [1].

When administered at the beginning and end of the term, CIs provide a good approximation of learning [41]. For example, Coletta and Phillips [23] correlated the FCI as the baseline measure in an introductory mechanics course with the normalized gain score on the FCI as the outcome measure. Meltzer [2] used the Conceptual Survey in Electricity (CSE) in a general physics course in the same manner.

These studies computed a normalized gain score and included the baseline measure as one of the predictor variables—along with the relevant APK measures—a method we adopted in the current study. However, we used normalized change rather than normalized gain as our outcome measure. Normalized change [42] is similar to normalized gain but is used to compare gains for individual students, whereas normalized gain is generally used to compute the average gain at the classroom level. Normalized change also includes a correction so that when an individual's score is lower at post-test than pretest, it decreases the penalty to be more commensurate with similar gains.

2. APK measures

When including outside domain prior knowledge, researchers have most frequently considered the role of scientific reasoning or mathematics in physics learning. Following our plan to include all related PK constructs in the pretest or post-test, we identified measures for each.

(a) *Scientific reasoning*. The CTSR [37] is a commonly used instrument in science education research [26], particularly in physics education research [23]. We included CTSR items that measured correlational reasoning, hypothetical-deductive reasoning, and identifying or controlling variables, all topics that our task analysis identified as relevant to solving target physics problems.

(b) *Mathematics*. To measure mathematics prior knowledge, researchers typically use prior coursework [43] or item-based instruments [2,3]. The latter, quantitative type of measure includes items that assess specific concepts and skills needed to perform physics problems and therefore may be more predictive of physics learning than participation in or grades in past mathematics coursework [44]. We drew items from a diagnostic instrument for algebra used in past studies of physics learning [2], as well as higher-level mathematics (e.g., vectors, calculus) items available from instruments on PhysPort [40].

(c) *Unconfounded APK measurement*. When measuring APK, items meant to measure APK should be independent of DPK, e.g., not contextualized in the domain of physics, to ensure that APK alone is being assessed. Some instruments contain both unconfounded and confounded APK items. For example, the Vector Evaluation Test (VET) [45] has items that test knowledge of vectors in a physics context. If those items are used to measure vectors as a mathematical concept (i.e., APK), student scores do not clearly differentiate between knowledge of vectors (independent of physics), physics domain knowledge, or both. Similarly, the CTSR [37] includes some items that evaluate scientific reasoning in the context of physics conservation concepts. Students could answer these items using scientific reasoning, but they may also answer them using physics knowledge (i.e., having some physics knowledge could help students score better), making it difficult to distinguish the type of PK that is driving predictions of learning. We excluded these types of items from our pretest and post-tests to "unconfound" the measurement of APK from DPK and enhance construct validity.

3. Selecting and piloting PK measures

While it is common to use CIs when investigating physics classroom learning, the selection criteria for such inventories are rarely discussed. Given that we expected PK to vary between, not just within, our two samples, we felt it important to select APK and DPK measures that would align with the level of instruction typically provided to each sample. In particular, we needed to ensure measures were sensitive for the sample(s) under study. If the PK measures were too easy [23], they would not be sensitive enough to detect a relationship of PK to learning should one exist and would not correspond to the level of the more challenging course instruction that was expected to produce a change in learning from pretest to post-test. The MWPU group was the sample for which this "too easy" scenario was more likely, given their generally strong background in science and math.

Given the issues in measuring PK, we needed to identify items for our various pretest instruments that would (a) cover the APK concepts and skills relevant to introductory physics problems, (b) exclude any items on APK measures that were confounded with DPK (e.g., because of physics content in the item), and (c) be appropriately sensitive to the likely APK and DPK variation in our NESU vs MWPU samples. To accomplish this, we used several methods: task analysis, instructor consultation, and piloting.

We conducted two types of task analysis. First, we sat with two students who were recent graduates of the target course and asked them to talk through (i.e., think aloud) the types of math and science knowledge that they would use on physics problems from the various CIs. This process identified the APK subareas that we needed to cover. We then shared these

TABLE I. APK and DPK pretest items for NESU students. Reference items are given for publicly available measures. Items administered to both samples are noted with an asterisk.

Prior knowledge type	Number of items	Source	Reference items
APK			
Algebra (APK-A)	6	Meltzer (2002) [46]	⋯
Vectors (APK-V)	4	Vector Evaluation Test (VET) [45]	VET_2; VET_3; VET_4; VET_24
Scientific reasoning (APK-SR)	8	Classroom Test of Scientific Reasoning (CTSR) [37]	CTSR_11*; CTSR_12;* CTSR_13*; CTSR_14;* CTSR_19*; CTSR_20;* CTSR_21*; CTSR_22*
DPK			
Angular kinematics (DPK-AK)	6	Rotational Kinematics Inventory (RKI) [51]	RKI_4*; RKI_7*; RKI_8*; RKI_9*; RKI_10*; RKI_11*
Energy and force L1 (DPK-EF L1)	12	Next-Gen Physical Science Diagnostic (NGPSD) [49]	UEM2; UEM5; UPC4; UPC5; UPEF4; UPEF5; MIF2; MIF3; MIF4; MIF7; MIF8; MIF9
Energy and force L2 (DPK-EF L2)	6	Mechanics Baseline Test (MBT) [50]	MBT_7*; MBT_11*; MBT_12*; MBT_18*; MBT_23*; MBT_24*
	3	Additional problems written by course professors	⋯

results with course professors to see if they agreed with the identified knowledge areas or would identify additional types of knowledge and found that instructors' and students' task analyses agreed. Next, we worked with three professors to identify items that addressed those areas of knowledge that would be neither too easy nor too difficult for the students in our study as well as be most relevant to course instruction. Finally, we piloted these measures to assess their appropriateness for the two samples.

Via the pilot with MWPU students, we determined the APK algebra pretest items initially selected [46] were too easy. We selected more difficult items from the same measure and supplemented these with items found online. APK vector pretest items selected from the VET [45] also proved too easy, so we chose more difficult items from this measure and added calculus [47] and graphing functions [48] items. Similarly, scores on the DPK energy and force pretest that were initially selected, a subset of items from the Next Gen Physical Science Diagnostic (NGPSD) [49], were near the ceiling and were replaced with a greater number of items from the Mechanics Baseline Test (MBT) [50] for MWPU students.

These "easier" test items were still administered to the NESU students, both the APK algebra items [46] and DPK items from the Next Gen Physical Science Diagnostic (NGPSD) [49], along with the smaller, initial set from the Mechanics Baseline Test (MBT) [50]. While this approach hindered direct comparison between the two samples for every measure in the study, a subset of MBT items were taken by both samples, as were items from the Rotational Kinematics Inventory (RKI) [51] and the CTSR [37]. Items taken by both samples are marked with asterisks in Tables I and II. All other measures differed. For example, the NESU group received only algebra items, whereas the MWPU group received more advanced algebra items plus calculus and graphing function items.

The NESU and MWPU groups' PK measures each included multiple types of DPK measures—energy and force (DPK-EF), angular kinematics (DPK-AK)—and multiple types of APK math measures, which included algebra (APK-A), vectors (APK-V), calculus (APK-C), and graphing functions (APK-GF). Each group also received an APK scientific reasoning (APK-SR) measure. Therefore, the PK constructs of DPK, APK math, and APK scientific reasoning were consistent between the two samples but the measurement of the constructs varied.

The combined PK measures were given at both pretest and post-test. For DPK-EF, we defined three outcome measures called level 1, level 2, and level 3, for which we calculated each individual's normalized change score. Each level was comprised of progressively harder items requiring more problem-solving and computation: level 1 (L1: NGPSD [49]), level 2 (L2: MBT [50]), and level 3 (L3: matter and interactions textbook [52] and instructor-created problems). Levels 1 and 2 were administered to NESU students and levels 2 and 3 to MWPU students.

C. Additional measures: PA and motivation

1. Prior achievement measures

To aid in the detection of PK effects on learning, we advocate for the use of PA measures (in addition to APK

TABLE II. APK and DPK pretest items for MWPU students. Reference items are given for publicly available measures. Items administered to both samples are noted with an asterisk.

Prior knowledge type	Number of items	Source	Reference items
APK			
Algebra (APK-A)	2	Meltzer (2019) [46]	⋯
	3	Google search for "hard algebra problems"	⋯
Calculus (APK-C)	5	The Calculus Concept Inventory (CCI) [47]	CCI_7; CCI_9; CCI_12; CCI_14; CCI_15
Vectors (APK-V)	5	Vector Evaluation Test (VET) [45]	VET_26; VET_27; VET_29; VET_30; VET_31
Graphed functions (APK-GF)	3	Concise Data Processing Assessment (CDPA) [48]	CDPA_3; CDPA_4; CDPA_7
Scientific reasoning (APK-SR)	8	Classroom Test of Scientific Reasoning (CTSR) [37]	CTSR_11*; CTSR_12;* CTSR_13*; CTSR_14;* CTSR_19*; CTSR_20;* CTSR_21*; CTSR_22*
DPK			
Angular kinematics (DPK-AK)	6	Rotational Kinematics Inventory (RKI) [51]	RKI_4*; RKI_7*; RKI_8*; RKI_9*; RKI_10*; RKI_11*
Energy and force L2 (DPK-EF L2)	14	Mechanics Baseline Test (MBT) [50]	MBT_7*; MBT_9; MBT_11;* MBT_12;* MBT_15; MBT_16; MBT_18;* MBT_20; MBT_21; MBT_22; MBT_23*; MBT_24;* MBT_25; MBT_26
Energy and force L3 (DPK-EF L3)	2	Matter & Interactions, Vol. 1 [52]	Ch. 5, Checkpoint 2 Ch. 6, Checkpoint 3
	5	Additional problems written by course professors	⋯

and DPK) to control for prior achievement as a known source of variance. Researchers have found PA to be a nontrivial factor even when combined with other APK or DPK measures [5]. Moreover, accounting for PA in the model can strengthen the claim that adequate PK, in addition to PA, is important for learning.

It is becoming more common to include PA measures (e.g., SAT math) with DPK measures (i.e., CIs) in multivariate analyses. Salehi *et al.* [5] and Hewagallage *et al.* [1] found that combining both types of measures explained more variance in final exam scores [5] or CI post-test scores [1] than either alone. However, it is still relatively uncommon to include both APK and PA alongside DPK measures in multivariate analyses. Most commonly, researchers have used either APK or PA measures, but not both, for example, using a math pretest when SAT scores are unavailable [2].

Although we would have liked to include a prior achievement measure for both groups, we were only able to obtain SAT or ACT scores for the MWPU group, and so these scores only appear in the linear regression models for MWPU. (If ACT scores were given, they were converted to SAT scores using a concordance table published by ACT and The College Board [53]).

2. Motivation measures

We included a motivation measure because past research suggests motivation may play a role in physics learning [54,55]. Additionally, we wished to control for potential differences in motivation between samples. Students were enrolled at two institutions with differing levels of selectivity and were in different levels of introductory physics courses. We included two subscales from the Motivated Strategies for Learning Questionnaire (MSLQ) [56]. Task Value (TV) measures how interesting, useful, or important the task is for the student (e.g., "It is important for me to learn the course material in this class."). Control of Learning (CL) measures whether the student believes efforts to study will improve learning and performance (e.g., "If I study in appropriate ways, then I will be able to learn the material in this course.").

D. The current studies

1. Research questions

This study was designed to explore the effects of multiple types and subtypes of PK on introductory physics learning in two learner samples. The research questions were as follows:

(1) Does APK predict learning, over and above the effects of DPK?

(2) Are the significant predictors of learning the same for samples with presumably different PK levels from different institutions?

2. *Plan of analysis*

These studies computed each individual's normalized gain score and included their baseline APK and DPK measures (i.e., pretest scores) as predictor variables in multiple linear regression models [57].

We report results in three sections. Section II details the procedure and results of the NESU sample. The outcome measures were (i) normalized change in L1 DPK and (ii) normalized change in L2 DPK. Section III details the procedure and results of the MWPU sample. The outcome measures were (i) normalized change in L2 DPK and (ii) and normalized change in L3 DPK. Section IV details the results of the direct comparison between NESU and MWPU for those items that were taken by both samples. The outcome measure was normalized change on the shared L2 DPK items.

The pretest and post-test for each sample used identical forms, with the exception of the motivation measures, which were given at the pretest only. We performed a hierarchical linear regression for each outcome measure after testing for normality of the predictor variables, making necessary transformations, computing bivariate correlations of APK, DPK, and the covariates, testing for multicollinearity, and, finally, computing normalized change [42] scores for each repeated measure.

Step 1 of the hierarchical linear regression modeled the effects of the APK variables, DPK variables, and covariates (motivation-control of learning, motivation-task value, Pretest time, and post-test time). For MWPU only, we included the additional covariates SAT math and verbal scores. Step 2 added normalized change (N-change) scores for each APK and DPK score to evaluate if learning on these component measures predicted N-change in the L1, L2, or L3 outcomes, over and above effects of all types of starting PK. Naturally, when a N-change variable was the outcome measure, we removed its component submeasure(s) as a predictor in step 2 (e.g., for the N-change DPK L1 outcome, N-change DPK L1 was excluded as a predictor).

II. STUDY 1 METHOD

A. NESU participants

Participants were undergraduate students in an introductory physics course who completed the study for course extra credit. The course, Physics 101, is algebra-based and covers mechanics and thermodynamics. Participation was open to all enrolled students in Spring 2020. Ninety-one students completed the pretest (time: $M = 29.20$ min, $SD = 14.04$ min) and 71 completed the post-test (time: $M = 36.38$ min, $SD = 20.55$ min). There were 43 participants who completed both pretest and post-test and also spent greater than ten minutes on each, so these matched pairs were included in the analyses. After the course ended, we checked for attrition bias by comparing course grades (prior to the addition of study-related extra credit) of participants who completed the study and those who did not. There was no significant difference between the groups, $F(1, 91) = 1.053$, $p = 0.308$, in course grade achieved.

B. NESU procedure

The tests were administered online, outside of classroom time, using Open Learning Initiative (OLI) software [58]. Students had two weeks at the beginning and end of the semester to complete each test. The course instructor was not blind to the items used but did not alter teaching materials, such as lecture slides, quizzes, or exams, from previous semesters in light of the chosen items.

III. STUDY 1 RESULTS

A. NESU data normality and correlations of predictor variables

Predictor variables were normally distributed. Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern, with all VIF < 2. Correlations between each pair of predictor variables are reported in Table III. The highest pairwise correlation was

TABLE III. NESU bivariate correlations of predictor variables and covariates. Note: $^{*}p < .05$, $^{**}p < .01$.

	APK-A	APK-V	APK-SR	DPK-EF L1	DPK-AK	DPK-EF L2	CL	TV
APK-A	1							
APK-V	0.307^{**}	1						
APK-SR	0.154	0.091	1					
DPK-EF L1	0.168	0.020	0.351^{**}	1				
DPK-AK	0.010	−0.002	0.162	0.214^{*}	1			
DPK-EF L2	0.296^{**}	0.168	0.236^{*}	0.241^{*}	−0.113	1		
CL	0.165	0.105	0.038	0.048	−0.005	0.059	1	
TV	0.119	0.065	0.068	0.048	0.210^{*}	−0.016	0.471^{**}	1

between the two motivation measures: CL and TV, $r = 0.47$. Eight pairs of predictor variables were significantly, positively correlated.

TABLE IV. NESU mean scores and standard deviations (in parentheses) for APK and DPK components. The average of individual pretest scores, post-test scores, normalized change scores are given for each measure. Pretest to post-test changes were compared to zero using *paired t-tests* (two-sided); $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Range	Pretest score	Post-test score	N-change
APK math total	0–10	3.82 (1.72)	5.05 (1.84)***	0.19
APK-A	0–6	2.56 (1.24)	3.07 (1.37)*	0.10
APK-V	0–4	1.26 (0.82)	1.98 (0.91)***	0.22
APK-SR	0–8	2.77 (1.36)	3.07 (1.76)	−0.08
DPK-EF L1	0–12	7.26 (2.31)	8.65 (2.23)**	0.32
DPK-AK	0–6	2.21 (1.17)	2.58 (1.24)	0.07
DPK-EF L2	0–9	4.02 (2.06)	4.60 (2.07)	−0.04

B. NESU learning gains in APK and DPK

We report pretest and post-test scores and N-change scores for each measure in Table IV. NESU students learned the most, relative to what they could have gained, in the areas of DPK-EF L1 (i.e., NGPSD items) and APK-V, and the least in the areas of APK-SR and DPK-EF L2 (i.e., MBT items).

C. NESU hierarchical linear regression analysis

1. Energy and force L1 learning

For N-change EF L1, both models were significant. Model 1, $F(10,30) = 2.794$, $p = 0.014$, $R^2 = 0.482$, $R^2{}_{\mathrm{adj}} = 0.310$, was significantly different from zero. Model 2, $F(4,26) = 1.830$, $p = 0.089$, $R^2 = 0.496$, $R^2{}_{\mathrm{adj}} = 0.225$, was not significantly different from model 1 ($\Delta R^2 = 0.014$, $p = 0.946$), so we report model 1 results (see Table V).

APK-SR and DPK-EF L2 were significant, positive predictors of learning for NESU students. The significant, negative predictor DPK-EF L1 means that students with less starting knowledge generally learned more during the term. The significant, negative predictor APK-A is more challenging to interpret and is discussed in the *post hoc* analysis.

2. Energy and force L2 learning

For N-change EF L2, both models were significant. Model 1, $F(10,30) = 4.318$, $p < 0.001$, $R^2 = 0.590$, $R^2{}_{\mathrm{adj}} = 0.453$, was significantly different from zero. Model 2, $F(5,25) = 3.289$, $p = 0.004$, $R^2 = 0.664$, $R^2{}_{\mathrm{adj}} = 0.462$, was not significantly different from model 1 ($\Delta R^2 = .074$, $p = 0.388$), and so we report model 1 results (see Table VI).

APK-SR, APK-V, and time spent on the post-test were significant, positive predictors of learning for NESU students. The significant, negative predictor DPK-EF L2 means that students with less starting knowledge generally learned more during the term.

D. NESU *post hoc* analysis

As a *post hoc* analysis, we wished to determine why greater APK-A presented as a negative predictor of EF L1 learning, as seen in Table V. We thought that perhaps this effect was moderated by learner factors. We did not have SAT scores for NESU students but did have college major data. Students could be grouped by major type: (i) Humanities or social sciences (HSS, $n = 24$; 51%

TABLE V. NESU's N-change EF L1 outcome with APK, DPK, time, and motivation scores as predictor variables. The overall model explains $R^2 = 0.48$; $R^2{}_{\mathrm{adj}} = 0.31$ [$F(10,30) = 2.794$, $p = 0.014$] of the variance in the N-change score. B is the regression coefficient for each variable, CI is the confidence interval for B, β the standardized regression coefficient, t the t statistic, and p the probability of a value as large or larger than t occurred by chance. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	B	95% CI	β	t	p
APK					
APK-SR	0.075	[0.009, 0.141]	0.348	2.333	0.027*
APK-A	−0.088	[−0.172, −0.004]	−0.344	−2.141	0.040*
APK-V	0.043	[−0.068, 0.153]	0.123	0.782	0.440
DPK					
DPK-EF L1	−0.051	[−0.092, −0.011]	−0.393	−2.576	0.015*
DPK-AK	−0.059	[−0.131, 0.014]	−0.242	−1.660	0.107
DPK-EF L2	0.047	[0.004, 0.089]	0.336	2.244	0.032*
Covariates					
TV	−0.052	[−0.130, 0.027]	−0.240	−1.351	0.187
CL	0.058	[−0.031, 0.147]	0.243	1.339	0.191
Pretest time	0.007	[−0.001, 0.016]	0.314	1.722	0.095
Post-test time	0.003	[−0.002, 0.008]	0.204	1.115	0.274

TABLE VI. NESU's N-change EF L2 outcome with APK, DPK, time, and motivation scores as predictor variables. The overall model explains $R^2 = 0.59$; $R^2_{\text{adj}} = 0.45$ [$F(10, 30) = 4.318$, $p < 0.001$] of the variance in the N-change score. B is the regression coefficient for each variable, CI is the confidence interval for B, β the standardized regression coefficient, t the t statistic, and p the probability of a value as large or larger than t occurred by chance. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	B	95% CI	β	t	p
APK					
APK-SR	0.085	[0.005, 0.164]	0.288	2.172	0.038^*
APK-A	−0.097	[−0.198, 0.004]	−0.280	−1.958	0.060
APK-V	0.141	[0.007, 0.275]	0.302	2.154	0.039^*
DPK					
DPK-EF L1	−0.032	[−0.081, 0.016]	−0.184	−1.355	0.186
DPK-AK	0.082	[−0.006, 0.169]	0.248	1.910	0.066
DPK-EF L2	−0.092	[−0.143, −0.041]	−0.487	−3.659	$<0.001^{***}$
Covariates					
TV	−0.044	[−0.139, 0.050]	−0.151	−0.957	0.346
CL	0.016	[−0.091, 0.123]	0.050	0.312	0.757
Pretest time	0.003	−0.008, 0.013	0.080	0.493	0.626
Post-test time	0.007	[0.000, 0.008]	0.336	2.069	0.047^*

architecture, 17% public health, 32% other), and (ii) Science or engineering (STEM, $n = 19$; 63% biology or chemistry, 21% computer science or engineering, 16% other; none were Physics majors).

We found that for STEM majors, greater APK-A and N-change A (i.e., more algebra learned during the term) correlated positively with N-change EF L1. However, the opposite was true for HSS majors, for whom higher scores on both of these measures correlated negatively with N-change EF L1, see Fig. 2.

This interaction of APK with college major suggests that algebra knowledge played out differently for the two groups of majors. In particular, HSS students seemed less able to *apply* greater math APK, and even the algebra they learned over the term, to learning physics than STEM students. This could have been due to better use of math support during the course or applied math instruction in other coursework during the same term. This general finding is worth further exploration to seek replication and further understanding.

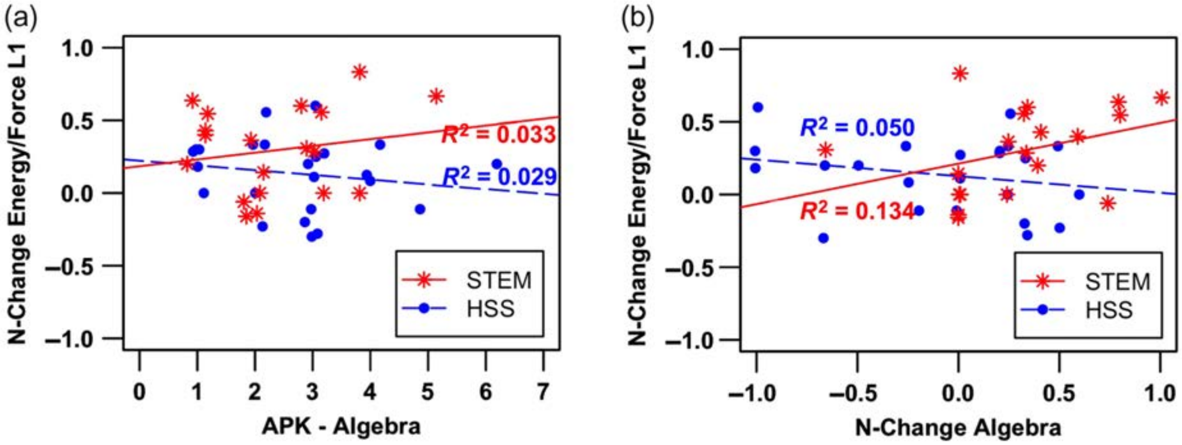

FIG. 2. (a) Interaction (not significant) of NESU students' college major and their APK-A on the N-change EF L1 outcome (i.e., the L1 physics knowledge gained during the term). STEM majors' N-change EF L1 scores positively correlated with better APK-A, whereas HSS (humanities and social sciences) majors' scores negatively correlated with it. (b) This same interaction (not significant) is present for STEM vs HSS majors when considering the relationship between greater N-change A (i.e., the amount of algebra learned during the term) and N-change EF L1. R^2 is the percentage of the variance in the outcome measure that is explained by the predictor variable.

IV. STUDY 2 METHOD

A. MWPU participants

Participants were undergraduate students in introductory physics courses at a large, private university in the Midwestern United States who received a gift card for completing the study. Students in two introductory courses, Physics I for engineering students and Physics I for science students, were included. These courses are calculus-based and cover mechanics and thermodynamics. Participation was open to all students enrolled in Spring 2020 and Fall 2020. Sixty-eight students completed the pretest (time: $M = 68.91$ min, $SD = 28.27$ min) and 60 completed the post-test (time: $M = 66.39$ min, $SD = 20.36$ min). Fifty-seven participants completed both pretest and post-test. These matched pairs were included in the analyses. Minimum times were not a concern. The lowest times were about half an hour. There were some very long post-test times, however, as the browser window continued to accrue time if left open. In cases where the recorded time was longer than 2 h, we set the upper limit to be 2 h. These cases were removed before the M and SD times were computed, but, even after their removal, the actual times are likely lower than the statistics reported.

After the course ended, we checked for attrition bias by comparing the SAT scores of the matched pairs with participants who only did the pretest. We did not find significant differences between the groups on SAT math, $F(1,62) = 3.158$, $p = 0.080$, or SAT verbal, $F(1,62) = 2.713$, $p = 0.105$.

B. MWPU procedure

The pretest and post-test were administered online, outside of classroom time, using Qualtrics [59] software. Students had two weeks at the beginning and end of the semester to complete each test. The course instructors were not blind to the items used but did not alter teaching materials, such as lecture slides, quizzes, or exams, from previous semesters in light of the chosen items.

V. STUDY 2 RESULTS

A. MWPU data normality and correlations of predictor variables

The predictor variables SAT math, SAT verbal, and APK-SR were negatively skewed and kurtotic. These variables were transformed using logarithmic transformations to meet the assumptions of normality. Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern, with all VIF < 2. Correlations between each pair of predictor variables are reported in Table VII. The highest pairwise correlation was between APK-C and APK-SR, $r = 0.501$. Thirty-three pairs of predictor variables were significantly, positively correlated.

B. MWPU learning gains in APK and DPK

We report pretest and post-test scores and N-change scores for each measure in Table VIII. Students learned the most, relative to what they could have gained, in the areas of DPK-EF L3 (i.e., problems), APK-C, DPK-EF L2 (MBT), and APK-A, and the least in the areas of APK-GF and APK-V.

C. MWPU hierarchical linear regression analysis

1. Energy and force L2 learning

For N-change EF L2, neither model was significantly different from zero. Model 1, $F(14,43) = 0.752$, $p = 0.712$, $R^2 = 0.197$, $R^2{}_{\mathrm{adj}} = 0.0$, and model 2, $F(20,37) = 1.154$, $p = 0.343$, $R^2 = 0.384$, $R^2{}_{\mathrm{adj}} = 0.051$. These results suggest that neither variability in APK or DPK nor learning, as measured by N-change scores, significantly predicted gains in EF L2 knowledge. These students were generally able to improve their domain knowledge, as measured by EF L2, by other factors such as class participation or study efforts.

TABLE VII. MWPU bivariate correlations of predictor variables and covariates. Note: $^{*}p < 0.05$, $^{**}p < 0.01$.

						DPK-EF		DPK-EF				
	APK-A	APK-C	APK-V	APK-GF	APK-SR	L2	DPK-AK	L3	CL	TV	SAT-M	SAT-V
APK-A	1											
APK-C	0.274*	1										
APK-V	0.291*	0.356**	1									
APK-GF	0.096	−0.090	−0.084	1								
APK-SR	0.319**	0.501**	0.265*	0.142	1							
DPK-EF L2	0.389**	0.348**	0.416**	0.137	0.210	1						
DPK-AK	0.425**	0.324**	0.437**	0.086	0.272*	0.440**	1					
DPK-EF L3	0.436**	0.420**	0.429**	0.016	0.373**	0.477**	0.496**	1				
CL	−0.075	0.148	0.100	0.035	0.122	0.047	0.057	0.185	1			
TV	0.107	0.034	0.215	0.158	0.196	0.240*	0.167	0.254*	0.477**	1		
SAT M	0.346**	0.409**	0.096	−0.225	0.356**	0.166	0.143	0.425**	0.275*	0.163	1	
SAT V	0.156	0.229	0.406**	−0.092	0.151	0.290*	0.268*	0.469**	0.211	0.203	0.359**	1

TABLE VIII. MWPU mean scores and standard deviations (in parentheses) for APK and DPK components. The average of individual pretest scores, post-test scores, normalized change scores are given for each measure. Pretest to post-test changes were compared to zero using *paired t-tests* (two-sided); $^{*}p < 0.005$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Range	Pretest score	Post-test score	N-change
APK math total	0–18	9.84 (3.16)	11.40 (2.41)***	0.16
APK-A	0–5	3.03 (1.31)	3.58 (1.00)**	0.21
APK-V	0–5	2.94 (1.65)	3.22 (1.20)	0.11
APK-C	0–5	3.18 (1.29)	3.79 (1.03)***	0.31
APK-GF	0–3	0.69 (0.74)	0.81 (0.79)	−0.07
APK-SR	0–8	6.03 (2.03)	6.15 (1.85)	0.18
DPK-EF L2	0–14	7.30 (2.78)	9.18 (2.73)***	0.26
DPK-AK	0–6	3.79 (1.53)	4.19 (1.61)	0.18
DPK-EF L3	0–7	4.03 (1.47)	4.97 (1.14)***	0.31

2. Energy and force L3 learning

For N-change EF L3, both models were significant. Model 1, $F(14, 43) = 2.374$, $p = 0.015$, $R^2 = 0.436$, $R^2{}_{\mathrm{adj}} = 0.252$, was significantly different from zero. Model 2, $F(21, 36) = 3.295$, $p < 0.001$, $R^2 = 0.658$, $R^2{}_{\mathrm{adj}} = 0.458$, was significantly different from model 1 ($\Delta R^2 = 0.222$, $p = 0.008$), and so we report model 2 results (see Table IX).

APK-A, N-change A, APK-GF, DPK-EF L2, and N-change EF L2 were significant, positive predictors of EF L3 learning for MWPU students. The significant, negative predictor DPK-EF L3 means that students with

TABLE IX. MWPU's N-change EF L3 outcome with APK, DPK, PA, time, and motivation scores as predictor variables. The overall model explains $R^2 = 0.66$; $R^2{}_{\mathrm{adj}} = 0.46$ [$F(21, 36) = 3.295$, $p < 0.001$] of the variance in the N-change score. B is the regression coefficient for each variable, CI is the confidence interval for B, β the standardized regression coefficient, t the t statistic, and p the probability of a value as large or larger than t occurred by chance. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	B	95% CI	β	t	p
APK					
APK-SR	−0.151	[−0.517, 0.215]	−0.120	−0.836	0.409
N-Ch SR	0.104	[−0.106, 0.314]	0.142	1.006	0.321
APK-A	0.118	[0.025, 0.211]	0.442	2.577	0.014*
N-Ch A	0.248	[0.035, 0.461]	0.294	2.366	0.024*
APK-V	0.015	[−0.052, 0.082]	0.069	0.455	0.651
N-Ch V	0.197	[−0.002, 0.395]	0.252	2.006	0.052
APK-C	0.000	[−0.092, 0.093]	0.002	0.011	0.991
N-Ch C	0.080	[−0.152, 0.312]	0.109	0.700	0.488
APK-GF	0.139	[0.009, 0.268]	0.290	2.174	0.036*
N-Ch GF	0.050	[−0.120, 0.221]	0.080	0.596	0.555
DPK					
DPK-EF L2	0.052	[0.008, 0.096]	0.412	2.387	0.022*
N-Ch EF L2	0.310	[0.043, 0.578]	0.281	2.353	0.024*
DPK-AK	−0.032	[−0.118, 0.054]	−0.134	−0.747	0.460
N-Ch AK	−0.048	[−0.291, 0.122]	−0.127	−0.828	0.413
DPK-EF L3	−0.012	[−0.020, −0.005]	−0.507	−3.223	0.003**
PA					
SAT M	−0.016	[−0.363, 0.332]	−0.013	−0.091	0.928
SAT V	0.054	[−0.256, 0.363]	0.044	0.351	0.728
Covariates					
TV	−0.011	[−0.106, 0.084]	−0.032	−0.240	0.812
CL	−0.002	[−0.113, 0.108]	−0.005	−0.039	0.969
Pretest time	0.001	[−0.002, 0.004]	0.083	0.625	0.536
Post-test time	−0.003	[−0.007, 0.000]	−0.293	−2.139	0.039*

TABLE X. MWPU and NESU mean scores and standard deviations (in parentheses) for APK, DPK, and motivation components. The average of individual pretest scores, posttest scores, normalized change scores are given for each measure. Pretest to post-test changes were compared to zero using paired t tests (two-sided); $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

		Range	Pretest score	Post-test score	N-change
APK-SR	NESU	0–8	2.77 (1.36)	3.07 (1.76)	−0.08
	MWPU	0–8	6.03 (2.03)	6.15 (1.85)	0.18
DPK-EF L2	NESU	0–6	1.50 (0.97)	1.91 (1.35)	−0.03
	MWPU	0–6	2.79 (1.35)	3.31 (1.43)*	0.11
DPK-AK	NESU	0–6	2.21 (1.17)	2.58 (1.24)	0.07
	MWPU	0–6	3.79 (1.53)	4.19 (1.61)	0.18
Motivation TV	NESU	0–7	4.95 (1.31)	⋯	⋯
	MWPU	0–7	5.62 (1.08)	⋯	⋯
Motivation CL	NESU	0–7	5.36 (1.19)	⋯	⋯
	MWPU	0–7	5.54 (0.84)	⋯	⋯

less starting knowledge generally learned more during the term. The significant, negative predictor of post-test time is a bit more challenging to interpret. Because the participants could keep the browser window open indefinitely, which resulted in some very long post-test times, this finding may suggest that participants who completed the post-test more expeditiously were more successful, perhaps due to being more attentive to the task.

VI. DIRECT COMPARISON OF NESU AND MWPU SAMPLES

In addition to analyzing the two groups' datasets separately, it is possible to compare learning as a function of PK in the two samples directly using a subset of the full measures that were taken by both, as shown by the asterisked items in Tables I and II.

A. Comparing amount and predictors of learning

In Table X, we report pretest, post-test, and N-change scores for each institution on the shared items. MWPU students learned a greater percentage than NESU students of what they could have learned, based on N-change scores. *Paired t tests* (two-sided) were run to detect any significant changes from pretest to post-test scores. Only MWPU students' MBT scores $[t(57) = 2.175,\ p = .034]$ were significantly better at the post-test than at pretest.

B. Multiple linear regression and interaction analyses

Next, we performed multiple linear regression on the outcome measure N-change EF L2, i.e., MBT items taken by both groups, with interaction terms as the predictor variables. The interaction terms were created by multiplying each PK measure with the institution, coded 0 (NESU) or 1 (MWPU), to detect whether PK (i.e., APK-SR, DPK-EF L2, and DPK-AK) or motivation scores had significantly different effects on learning depending on the institution. All PK and motivation measures were mean centered to aid in interpretation.

The interaction model was significant and explained $R^2 = 0.35$, $R^2{}_{\mathrm{adj}} = 0.27$, $F(11, 93) = 4.541$, $p < 0.001$,

TABLE XI. Interaction effects of Institution (NESU = 0; MWPU = 1) with APK, DPK, and motivation scores on N-change EF L2. The overall model explains $R^2 = 0.35$; $R^2{}_{\mathrm{adj}} = 0.27$; $F(11, 93) = 4.541$, $p < 0.001$, of the variance in the N-change score. B is the regression coefficient for each variable, CI is the confidence interval for B, β the standardized regression coefficient, t the t statistic, and p the probability of a value as large or larger than t occurred by chance. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	B	95% CI	β	t	p
APK					
Institution × APK SR	−0.122	[−0.221, −0.022]	−0.383	−2.421	0.017*
DPK					
Institution × DPK-EF L2	0.181	[0.043, 0.319]	0.439	2.606	0.011*
Institution × DPK AK	−0.054	[−0.175, 0.066]	−0.150	−0.898	0.372
Motivation					
Institution × CL	0.020	[−0.147, 0.186]	0.028	0.232	0.817
Institution × TV	0.112	[−0.037, 0.262]	0.197	1.496	0.138

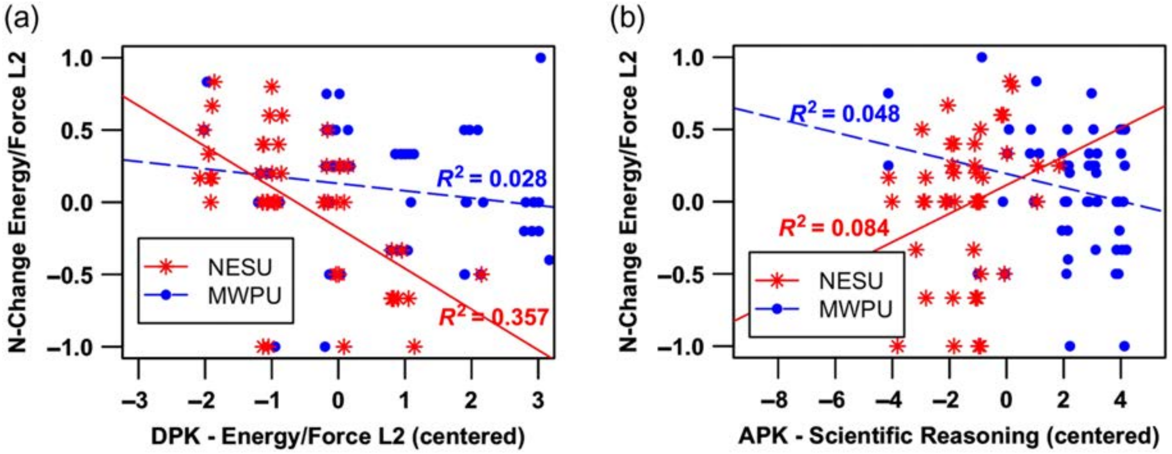

FIG. 3. (a) Significant interaction of institution and DPK-EF L2 on the N-change EF L2 outcome. Both NESU and MWPU students' DPK-EF L2 scores negatively correlated with N-change EF L2, such that students who started with higher scores tended to gain less new knowledge. However, NESU had more students who were below the mean in starting DPK and whose gains were above the mean compared with MWPU students who clustered above the mean in DPK and therefore had less room for growth. (b) A significant interaction of institution and APK-SR on the N-change EF L2 outcome. NESU students' scores were generally better as a function of more APK-SR, whereas MWPU students' scores were not clearly related to APK-SR. A few outlier scores for MWPU students who scored low on APK scientific reasoning but high on N-change energy and force L2 account for the negative relationship in that group and suggest APK-SR was not a factor in MWPU students' domain learning. R^2 is the percentage of the variance in the outcome measure that is explained by the predictor variable. All measures were mean centered.

of the variance in the N-change score, with two significant interaction variables: institution× APK-SR and institution× DPK-EF L2 (see Table XI).

Graphing the two significant interaction terms aids in their interpretation, see Fig. 3. The first significant interaction between DPK-EF L2 and institution on N-change DPK EF L2 scores reflects that, while all learners generally learned less when they started with more domain knowledge, this negative relationship was significantly stronger for NESU students because many started with lower DPK and therefore were able to make larger gains.

The second significant interaction between APK-SR and institution on N-change DPK EF L2 scores reflects that the linear relationship between APK-SR and learning was positive for NESU students and negative for MWPU students. We know from the analyses of the full NESU and MWPU datasets that greater APK-SR was a highly significant predictor of learning for NESU students but not for MWPU students, and this finding is also reflected in the direct comparison.

Mean centering both groups' scores shows the majority of NESU students' APK-SR scores fell below the mean, highlighting a relative deficit in a key skill used to solve physics problems. The few MWPU students whose APK-SR scores fell below the mean still managed to score well on N-change EF L2, a result that highlights the pitfall of using a single variable to predict learning vs comprehensive and discreet measurement of all PK types that may affect learning.

VII. DISCUSSION

A. Research questions

This study investigated two research questions. The first research question asked: Does APK predict learning, over and above the effects of DPK? The evidence suggests it does. In both samples, after variance in the outcome due to DPK was accounted for, APK was strongly predictive of learning. This was the case with NESU students on the L1 outcome (i.e., items from the NGPSD; significant predictors: APK-SR, DPK-EF L2) and L2 outcome (i.e., items from the MBT; significant predictors: APK-SR, APK-V), as well as with MWPU students on the L3 outcome (i.e., problems; significant predictors: APK-A, DPK-EF L2, APK-GF, N-Change A; N-Change EF L2). This finding suggests that gaps in reasoning and math can be a source of learning reduction, just as gaps in physics knowledge can be. In contrast to APK and DPK, neither prior achievement nor motivation explained significant variance in the outcome measures.

The second research question asked: Are the impactful PK gaps the same across the two samples? The evidence suggests they are not. In these two distinct contexts, an algebra-based course taken by a mix of STEM and non-STEM majors vs a calculus-based course taken by almost exclusively STEM majors, a very different mix of APK predictors emerged. Whereas variance in the NESU group's learning was explained by APK scientific reasoning and APK math, variance in the MWPU group's learning was explained by various types of APK math but not APK scientific reasoning. By contrast, DPK, measured by students' scores on the MBT, consistently predicted learning in both samples and across difficulty levels of the outcome measure. Specifically, DPK-EF L2 predicted gains on DPK-EF L1 for the NESU sample and gains on DPK-EF L3 for the MWPU sample.

The differences in the significant predictors of each group's learning can be interpreted through two lenses: Characteristics of each sample's PK and the knowledge demands of the particular outcome measures. For the first lens, characteristics of cohort PK, one of the most striking differences between the samples was the low intercorrelation of the NESU group's PK measures vs the high intercorrelation of the MWPU group's PK measures. For the NESU group, 28.6% of all PK measures were significantly correlated with a median correlation value of 24.1 and a range of 21.0–35.1, see Table III. For the MWPU group, 50.8% of all PK measures were significantly correlated with a median correlation value of 40.3 and a range of 26.5–50.1, see Table VII.

Extensive research into knowledge differences between novices and experts has shown that experts possess not only more knowledge but also knowledge that is more interrelated [19]. For example, a study comparing math teachers with higher and lower levels of mathematical expertise found that the teachers more expert in mathematics had better scores on scales for both mathematics and teaching, despite similar teaching preparation between the two groups. The teachers more expert in math also had higher intercorrelations between their math and teaching knowledge than those less expert in math, reflecting that their math and teaching knowledge was differently structured and showed stronger integration [60]. Similar evidence of a progression toward greater expertise can be detected in the higher number and generally stronger knowledge intercorrelations between APK, DPK, and PA in the MWPU group, indicating their knowledge structures have begun to take on this quality of interconnectedness. For example, MWPU students' DPK-EF L3 scores were significantly correlated with nine other measures (listed highest to lowest): DPK-AK, DPK-EF L2, SAT verbal, APK-A, APK-C, SAT math, APK-V, APK-SR, and TV (see Table VII). Indeed, the highest pairwise correlation for MWPU students was between APK-C and APK-SR, $r = 0.501$, whereas the highest pairwise correlation for NESU students was between the two motivation measures: CL and TV, $r = 0.471$.

In addition to being more intercorrelated, expert knowledge is more locally coherent and differentiated, such that each category of knowledge is rich in features and applications [15,19]. For example, knowing a lot about many specific classes of dinosaurs aids in discriminating between them when given a problem such as classifying a new dinosaur. Similarly, better knowledge of algebra, vectors, and calculus is useful when determining which procedures to apply to solve a physics problem. This may also explain why specific, differentiated math measures (e.g., vectors and graphed functions) explained variance in physics learning in the MWPU group when the more general PA measures (e.g., SAT math and verbal) did not. Moreover, this finding is consistent with Nakakoji and Wilson's [3] result that, for less advanced students, PA had a direct effect on course grade but for more advanced students, PA had only an indirect effect that was mediated by prior math grades.

The NESU group's PK, on the other hand, was characterized by fewer intercorrelations with lower correlation coefficient values. APK-SR was significantly correlated with only DPK-EF L1 and DPK-EF L2 in the NESU group, whereas, in the MWPU group, APK-SR was correlated with almost every other PK measure and *not* predictive of EF learning. By extension, despite generally low scores at pretest and post-test on APK-SR in the NESU group, this measure was the strongest predictor of domain learning, predicting gains on both outcome measures (i.e., EF L1 and EF L2). This finding is consistent with Coletta and Phillip's [23] result that, for less advanced students, but not for more advanced students, APK-SR was a significant predictor of learning.

Considering students' majors in the NESU group further pointed to a unique property of APK—transfer of knowledge—that was not formally included in the scope of this study. Transfer is the ability to apply what you have learned in one context to a similar context. For APK to have its best effect on learning in a new domain, students need to be able to apply it across contexts, e.g., math skills are applied not only in math learning but also in physics, biology, and statistics learning.

We saw that NESU STEM students were better able to apply their APK algebra and their gains in algebra during the term than humanities and social science students, as seen in the positive relationship of both APK-A and N-change A to changes in domain knowledge for STEM students. The first positive relationship—greater APK algebra correlating with greater N-change energy and force—suggests that STEM students were able to apply what they already knew at the start of the term to physics learning. The second positive relationship—greater algebra learning over the term correlating with greater N-change energy and force—suggests that STEM students were able to apply concurrent algebra learning to physics learning during the term. The STEM students, who likely have encountered math skills in multiple prior contexts and may have been taking other STEM courses, were better able to transfer those skills to further their physics learning, whereas the non-STEM students were not. This finding underscores the importance of providing multiple and varied contexts to practice math skills in order to support transfer to physics.

The second lens for interpreting differences in predictors of learning in the two groups involves considering the demands of each particular outcome measure. Specifically, the alignment of the APK measures to the required tasks on the outcome measures. By design, each "level" of the DPK measures included progressively more applied problems requiring math skill. Therefore, it is not surprising that knowledge gains on the L1 measure, items on the NGPSD taken by the NESU group, were not predicted by APK math but only by APK-SR; knowledge gains on the L2 measure, items on the MBT taken by both groups, were predicted by APK math for the NESU group but not the MWPU group; and knowledge gains on the L3 measure, which included the most difficult problems, were predicted by APK math for the MWPU group. When the outcome measure does not require math, or math that is challenging for a given group, we would not expect math to show up as a significant predictor of learning on that measure, as happened here.

B. Significance and practical implications

We wrote in the introduction that evidence for the impact of PK on learning has been inconsistent. If APK were routinely included *in addition to* DPK in the type of PK measures given, PK would certainly be a more consistent predictor of learning. Without sufficient and differentiated measures of PK, the effect of PK on learning may be missed or, due to intercorrelations in knowledge, one type of measured PK may appear to have a significant impact on learning, but it is simply related to another, unmeasured type of PK that has the true impact on learning [35]. Including measures of APK and DPK together in this study exemplifies an approach to including all relevant types of PK in a multivariate analysis. With distinct subtypes of APK and DPK measured and entered into the analyses, we found APK was strongly predictive of learning in both study groups.

Another goal of including multiple, instruction-relevant APK and DPK measures could be to help identify predictors of learning and other educational outcomes using predictive or learning analytics [61]. These methods combine many variables into a single analysis and apply modern data mining and modeling techniques to identify predictor variables and make quantitative predictions about learners' outcomes (e.g., whether they will graduate on time or pass a given course) [62]. In this approach, datasets tend to be "wide" (i.e., including easy-to-collect data about many students) but not necessarily "deep" (i.e., not including longitudinal and/or detailed data about each student). That said, researchers using this approach have found that adding measures of DPK can improve predictive models compared with the use of demographic data alone [63]. With increasing evidence of the role of APK in predicting learning, we advocate for including both APK and DPK measures in such models.

What are the practical implications of this work? We believe it is important and worthwhile for educators to look for various types of PK gaps and provide support for the particular PK gaps that are likely affecting learning in their courses. The influence of PK (and its variability across students) may not be obvious because, as long as some students are able to keep pace with the instruction, it can be hard to detect that missing PK is diminishing other students' ability to keep up. Moreover, with the bulk of the instructional focus on domain knowledge, gaps in APK easily fall outside the scope of instructors' awareness and, hence, instruction. And yet, APK gaps present a considerable opportunity for remediation of missing knowledge and skills when seeking to improve physics learning outcomes.

Without conducting a full-blown study in their own courses, how can professors hone in on the likely PK (and especially APK) elements to address? The results of this study, and especially the patterns across the two samples, highlight a key message. The most important PK measures are the ones that (a) map to the knowledge characteristics of one's students (i.e., where they have some, but not rock-solid, PK) and (b) align with the knowledge demands of one's course (e.g., the kind or level of math that students must regularly apply to learn the target physics material). These two lenses can guide educators to the most important types of PK to measure *for their learners*. The measurement approach used in this study demonstrates that shorter measures from a wider variety of PK types can be given in a reasonable amount of time (one study period or less). Collecting data specific to one's learners will lead to the most accurate characterization of the gap.

C. Strengths and limitations

This study builds on past research by systematizing and improving the measurement of APK vs DPK in a way that specifies and differentiates multiple forms of PK in order to analyze their correlational structure and unique contributions to domain learning. In future research on the effects of PK on learning, we recommend that researchers continue to use multiple measures of PK that are discrete, specific, and carefully aligned to the outcome measures. Building on this approach, we recommend that future research conduct factor analyses and item response theory analyses for the PK subtype measures to further study their validity and reliability and to inform any refinements. Note that our finding of differing numbers and degrees of intercorrelations between knowledge subtype measures in high- and low-PK learners, supported by research into differences in how novices and experts structure their knowledge [15,19,64], suggests both low- and high-PK learners should continue to be studied when performing dimension reduction. For example, knowledge subtypes will likely load on different factors in low- vs high-PK learners, and items on individual measures will likely have different discrimination and difficulty profiles when tested with these two groups of learners.

Relatedly, it is important that the measures are piloted and tuned appropriately such that they are sensitive enough to detect variation in PK in a given sample of students. The pilot data collected for this study showed that all items on CI measures may not be equally effective for detecting variations in knowledge for all learner groups, particularly those with greater PK. Piloting measures and carefully selecting items with a better ability to discriminate knowledge will help ensure that if variation in PK correlates with learning outcome measures, then these relationships will be detected. In addition, past research suggests that a subset of the full test items on the MBT has sufficient difficulty and discrimination to return a good measure of student skill at a highly selective university and has identified items that do not appropriately discriminate in high-PK learners [64]. A future direction would be to conduct more such studies with often-used CI measures to enhance validity claims.

A significant limitation of these studies was the small sample size of each group. Future studies should include bigger, broader samples and consider methods to achieve high participation and retention rates. In addition, it is important to consider the use of differing measures across our two samples as both a strength and a limitation of this study. Generally, we consider the approach of using targeted measures that can reasonably detect variability in different groups' PK to be a strength. If we had not adjusted the measures across the two groups, we would not be acknowledging differences in their PK and would not have been able to show the generalizability of the role played by APK in learning across groups with diverse PK profiles. The limitation of this approach is that it makes direct comparison between the groups more difficult. A goal of future research on CIs and APK measures should include identifying items that either (a) appropriately discriminate knowledge in both high- and low-PK learners or (b) are relatively matched in their discrimination and difficulty levels across low-PK and high-PK groups to aid in comparison.

VIII. CONCLUSION

In this study, we measured multiple, distinct types of PK, from within and outside the domain of instruction, to determine whether they explained unique variance in novices' learning in introductory physics courses taken by two distinct learner groups. Using correlation and regression analyses, we regressed three, increasingly difficult outcome measures of physics learning on the PK predictor variables. Within-domain PK predicted learning in a similar manner across both types of courses, with items from the MBT predicting learning in both learner groups and on outcome measures of varying difficulty. Outside-domain PK, on the other hand, predicted learning according to somewhat different patterns related to (a) characteristics of the group's prior knowledge and (b) the difficulty of the outcome measures.

When considering characteristics of the group's prior knowledge, MWPU students' PK measures showed a greater number and degree of intercorrelations, suggesting more intercorrelated and differentiated prior knowledge, as well as a greater variety in the types of PK measures that were predictive of their learning, primarily DPK and APK math. The NESU students had a fewer number and lesser degree of intercorrelations in their PK measures, and for this group, APK scientific reasoning was a consistently strong predictor across outcome measures. When considering the difficulty of the outcome measures, APK math showed up as a strong predictor on the more problem-heavy measure (i.e., L3) whereas APK scientific reasoning was a strong predictor on the more conceptual measure (i.e., L1).

These results indicate that all novice learners have meaningful variations in PK that can affect learning, regardless of whether they are more or less advanced domain novices. This finding underscores the importance of measuring more specific, rather than more general, PK constructs, particularly as learners' knowledge becomes more interconnected. The ultimate goal of identifying variations in PK types that impact learning within a peer group is to identify actionable areas where additional instructional support can reduce the preparation gap and improve learner outcomes.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dr. Stephen Garoff for his valuable assistance with item selection and review of both the outcome measures and the manuscript, as well as Dr. Manfred Paulini for his support of this research. The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grant No. R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

A. D. and M. L. conceptualized the work and methods; A. D. conducted the experiments, analyzed the data, wrote the original draft; created visualizations; M. L. supervised the research; A. D., D. A., and S. S. composed the measures; A. D., M. L., D. A., and S. S. reviewed and edited the work.

[1] D. Hewagallage, E. Christman, and J. Stewart, Examining the relation of high school preparation and college achievement to conceptual understanding, Phys. Rev. Phys. Educ. Res. **18**, 010149 (2022).

[2] D. E. Meltzer, The relationship between mathematics preparation and conceptual learning gains in physics: A possible "hidden variable" in diagnostic pretest scores, Am. J. Phys. **70**, 1259 (2002).

[3] Y. Nakakoji and R. Wilson, First-year mathematics and its application to science: Evidence of transfer of learning to physics and engineering, Educ. Sci. **8**, 8 (2018).

[4] J. Stewart, G. L. Cochran, R. Henderson, C. Zabriskie, S. DeVore, P. Miller, G. Stewart, and L. Michaluk, Mediational effect of prior preparation on performance differences of students underrepresented in physics, Phys. Rev. Phys. Educ. Res. **17**, 010107 (2021).

[5] S. Salehi, E. Burkholder, G. P. Lepage, S. Pollock, and C. Wieman, Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics, Phys. Rev. Phys. Educ. Res. **15**, 020114 (2019).

[6] D. P. Ausubel, *Educational Psychology: A Cognitive View* (Holt, Rinehart & Winston, New York, NY, 1986).

[7] F. Dochy, M. Segers, and M. M. Buehl, The relation between assessment practices and outcomes of studies: The case of research on prior knowledge, Rev. Educ. Res. **69**, 145 (1999).

[8] J. Hattie, *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement* (Routledge, New York, NY, 2008).

[9] D. H. Jonassen and B. L. Grabowski, *Handbook of Individual Differences, Learning, and Instruction* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1993).

[10] W. Kintsch, Text comprehension, memory, and learning, Am. Psychol. **49**, 294 (1994).

[11] M. L. Means and J. F. Voss, Star Wars: A developmental study of expert and novice knowledge structures, J. Mem. Lang. **24**, 746 (1985).

[12] Y. Ozuru, K. Dempsey, and D. S. McNamara, Prior knowledge, reading skill, and text cohesion in the comprehension of science texts, Learn. Instr. **19**, 228 (2009).

[13] S. Chandran, D. F. Treagust, and K. Tobin, The role of cognitive factors in chemistry achievement, J. Res. Sci. Teach. **24**, 145 (1987).

[14] H. H. Zeitoun, The relationship between abstract concept achievement and prior knowledge, formal reasoning ability, and gender, Int. J. Sci. Educ. **11**, 227 (1989).

[15] M. T. H. Chi, J. E. Hutchinson, and A. F. Robin, How inferences about novel domain-related concepts can be constrained by structured knowledge, Merrill-Palmer Quart. **35**, 27 (1989), https://www.jstor.org/stable/23086424.

[16] P. A. Alexander and J. E. Judy, The interaction of domain-specific and strategic knowledge in academic performance, Rev. Educ. Res. **58**, 375 (1988).

[17] A. Minnaert and P. J. Janssen, How general are the effects of domain-specific prior knowledge on study expertise as compared to general thinking skills?, in *Alternatives in assessment of achievements, learning processes and prior knowledge*, edited by F. J. R. C. Dochy and M. Birenbaum (Kluwer Academic/Plenum Publishers, New York, NY, 1996).

[18] A. E. Witherby and S. K. Carpenter, The rich-get-richer effect: Prior knowledge predicts new learning of domain-relevant information, J. Exp. Psychol. Learn. Mem. Cogn. **48**, 483 (2021).

[19] M. T. H. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, Cogn. Sci. **5**, 121 (1981).

[20] M. K. Singley and J. R. Anderson, *The transfer of cognitive skill* (Harvard University Press, Cambridge, MA, 1989).

[21] B. A. Simonsmeier, M. Flaig, A. Deiglmayr, L. Schalk, and M. Schneider, Domain-specific prior knowledge and learning: A meta-analysis, Educ. Psychol. **57**, 31 (2022).

[22] 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).

[23] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[24] A. B. Delahay and M. C. Lovett, Distinguishing two types of prior knowledge that support novice learners, in *Proceedings of the 41st Annual Conference of the Cognitive Science Society, Montreal, Quebec, Canada* (Cognitive Science Society, Massachusetts, 2019), pp. 1620–1626.

[25] S. Chandran, D. F. Treagust, and K. Tobin, The role of cognitive factors in chemistry achievement, J. Res. Sci. Teach. **24**, 145 (1987).

[26] M. S. Cracolice and B. D. Busby, Preparation for college general chemistry: More than just a matter of content knowledge acquisition, J. Chem. Educ. **92**, 1790 (2015).

[27] S. E. Lewis and J. E. Lewis, Predicting at-risk students in general chemistry: Comparing formal thought to a general achievement measure, Chem. Educ. Res. Pract. **8**, 32 (2007).

[28] K. Derr, R. Hübl, and M. Z. Ahmed, Prior knowledge in mathematics and study success in engineering: Informational value of learner data collected from a web-based precourse, Eur. J. Eng. Educ. **43**, 911 (2018).

[29] E. K. Bone and R. J. Reid, Prior learning in biology at high school does not predict performance in the first year at university, High. Educ. Res. Dev. **30**, 709 (2011).

[30] F. J. Dochy, M. M Valcke, and L. J. Wagemans, Learning economics in higher education: An investigation concerning the quality and impact of expertise, High. Educ. Europe **16**, 123 (1991).

[31] D. Lombardi and G. M. Sinatra, College students' perceptions about the plausibility of human-induced climate change, Res. Sci. Educ. **42**, 201 (2012).

[32] R. Bhathal, An appraisal of an online tutorial system for the teaching and learning of engineering physics in conjunction with contextual physics and mathematics, and relevant mathematics, Eur. J. Eng. Educ. **41**, 504 (2016).

[33] S. Dehipawala, V. Shekoyan, and H. Yao, Using mathematics review to enhance problem solving skills in general physics classes, in *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, Bridgeport, CT* (IEEE, New York, 2014), pp. 1–4.

[34] H. T. Hudson and D. Liberman, The combined effect of mathematics skills and formal operational reasoning on student performance in the general physics course, Am. J. Phys. **50**, 1117 (1982).

[35] L. R. James, Testing hypotheses in the context of the unmeasured variables problem, Hum. Resour. Manag. Rev. **1**, 273 (1991).

[36] D. P. Maloney, Comparative reasoning abilities of college students, Am. J. Phys. **49**, 784 (1981).

[37] A. E. Lawson, The development and validation of a classroom test of formal reasoning, J. Res. Sci. Teach. **15**, 11 (1978).

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

[39] F. J. Dochy and P. A. Alexander, Mapping prior knowledge: A framework for discussion among researchers, Eur. J. Psychol. Educ. **10**, 225 (1995).

[40] PhysPort, https://www.physport.org.

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

[42] J. D. Marx and K. Cummings, Normalized change, Am. J. Phys. **75**, 87 (2007).

[43] P. M. Sadler and R. H. Tai, The two high-school pillars supporting college science, Science **317**, 457 (2007).

[44] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985).

[45] R. Thornton, Measuring and improving student mathematical skills for modeling, in *Proceedings of GIREP Conference Modelling in Physics and Physics Education, Amsterdam, Netherlands* (GIREP, Boechout/Vremde, BE, 2006), pp. 78–90.

[46] D. E. Melzer, Diagnostic math exam for physics (private communication).

[47] J. Epstein, The calculus concept inventory, in *Proceedings of the National STEM Assessment Conference, Washington, DC* (Drury University, Springfield, MO, 2007), pp. 60–67.

[48] J. Day and D. Bonn, Development of the concise data processing assessment, Phys. Rev. ST Phys. Educ. Res. **7**, 010114 (2011).

[49] P. V. Engelhardt, S. Robinson, E. P. Price, P. S. Smith, and F. Goldberg. Developing a conceptual assessment for a modular curriculum, in *Proceedings of 2018 Physics Education Research Conference* (American Association of Physics Teachers, College Park, MD, 2018), pp. 103–106.

[50] D. Hestenes and M. Wells, A mechanics baseline test, Phys. Teach. **30**, 159 (1992).

[51] K. K. Mashood and V. A. Singh, Rotational kinematics of a rigid body about a fixed axis: Development and analysis of an inventory, Eur. J. Phys. **36**, 045020 (2015).

[52] R. W. Chabay and B. A. Sherwood, *Matter and Interactions: Volume One Modern Mechanics*, 4th ed. (John Wiley & Sons, Inc., Hoboken, NJ, 2015).

[53] The College Board, & ACT, Inc., Guide to the 2018 ACT/SAT Concordance (2018), https://collegereadiness.collegeboard.org/pdf/guide-2018-act-sat-concordance.pdf.

[54] R. Dou, E. Brewe, J. P Zwolak, G. Potvin, E. A. Williams, and L. H. Kramer, Beyond performance metrics: Examining a decrease in students' physics self-efficacy through a social networks lens, Phys. Rev. Phys. Educ. Res. **12**, 020124 (2016).

[55] E. Marshman, Z. Y. Kalender, C. Schunn, T. Nokes-Malach, and C. Singh, A longitudinal analysis of students' motivational characteristics in introductory physics courses: Gender differences, Can. J. Phys. **96**, 391 (2018).

[56] P. R. Pintrich and E. V. De Groot, Motivational and self-regulated learning components of classroom academic performance, J. Educ. Psychol. **82**, 33 (1990).

[57] R. Theobald and S. Freeman, Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research, CBE Life Sci. Educ. **13**, 1 (2014).

[58] Open Learning Initiative, Carnegie Mellon University, http://oli.cmu.edu.

[59] Qualtrics, https://www.qualtrics.com.

[60] S. Krauss, M. Brunner, M. Kunter, J. Baumert, W. Blum, M. Neubrand, and A. Jordan, Pedagogical content knowledge and content knowledge of secondary mathematics teachers, J. Educ. Psychol. **100**, 716 (2008).

[61] N. Sclater, *Learning Analytics Explained* (Routledge, New York, 2017).

[62] T. McKay, K. Miller, and J. Tritz, What to do with actionable intelligence: E2Coach as an intervention engine, in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, BC, Canada* (ACM, New York, 2012), pp. 88–91.

[63] R. Bertolini, S. J. Finch, and R. H. Nehm, Testing the impact of novel assessment sources and machine learning methods on predictive outcome modeling in undergraduate biology, J. Sci. Educ. Technol. **30**, 193 (2021).

[64] C. N. Cardamone, J. E. Abbott, S. Rayyan, D. T. Seaton, A. Pawl, and D. E. Pritchard, Item response theory analysis of the mechanics baseline test, AIP Conf. Proc. **1413**, 135 (2012).