Assessment of representational competence in kinematics

P. Klein,[1,*] A. Müller,[2] and J. Kuhn[1]

[1]Department of Physics, Physics Education Research Group, University of Kaiserslautern, Erwin-Schrödinger-Straße 46, 67663 Kaiserslautern, Germany

[2]Faculty of Science/Physics Section and Institute of Teacher Education, University of Geneva Pavillon d'Uni Mail, Boulevard du Pont d'Arve 40, CH-1211 Geneve, Switzerland

(Received 22 September 2016; revised manuscript received 26 April 2017; published 26 June 2017)

A two-tier instrument for representational competence in the field of kinematics (KiRC) is presented, designed for a standard (1st year) calculus-based introductory mechanics course. It comprises 11 multiple choice (MC) and 7 multiple true-false (MTF) questions involving multiple representational formats, such as graphs, pictures, and formal (mathematical) expressions (1st tier). Furthermore, students express their answer confidence for selected items, providing additional information (2nd tier). Measurement characteristics of KiRC were assessed in a validation sample (pre- and post-test, $N = 83$ and $N = 46$, respectively), including usefulness for measuring learning gain. Validity is checked by interviews and by benchmarking KiRC against related measures. Values for item difficulty, discrimination, and consistency are in the desired ranges; in particular, a good reliability was obtained (KR20 $= 0.86$). Confidence intervals were computed and a replication study yielded values within the latter. For practical and research purposes, KiRC as a diagnostic tool goes beyond related extant instruments both for the representational formats (e.g., mathematical expressions) and for the scope of content covered (e.g., choice of coordinate systems). Together with the satisfactory psychometric properties it appears a versatile and reliable tool for assessing students' representational competency in kinematics (and of its potential change). Confidence judgments add further information to the diagnostic potential of the test, in particular for representational misconceptions. Moreover, we present an analytic result for the question—arising from guessing correction or educational considerations—of how the total effect size (Cohen's d) varies upon combination of two test components with known individual effect sizes, and then discuss the results in the case of KiRC (MC and MTF combination). The introduced method of test combination analysis can be applied to any test comprising two components for the purpose of finding effect size ranges.

DOI: 10.1103/PhysRevPhysEducRes.13.010132

I. INTRODUCTION

The purpose of developing the KiRC inventory was to provide an instrument for measuring changes in representational competence in the conceptual framework of kinematics. Representational competence (RC) can be defined as the ability to interpret and to construct multiple representations as well as to translate and switch from one representation to another [1]. The term multiple representations (MRs) refers to the many different forms in which a certain physics concept is expressed, demonstrated, depicted, and communicated, such as words, graphs, algebraic expressions, pictures, free-body diagrams, data tables, etc. (cf. Fig. 1 and Ref. [2]). The target population for which the KiRC inventory is constructed is undergraduate students at the beginning of introductory mechanics courses.

A. MRs in general and in kinematics

Physics education research points out that competent handling of representations is a key to successful physics learning [3]. In particular, RC and problem-solving skills are closely connected, i.e., students who are consistently competent in using different representations perform better at problem-solving tasks [4,5].

A similar relation holds for MRs and domain expertise in general, i.e., fluency and consistency of their use are crucial parts of expert knowledge and competence in a certain domain [6].

Even though lecturers and teachers may assume that students' RC (e.g., graph and formula interpretation skills) is adequately developed when enrolling at university, there is a lot of evidence that many students have difficulties with MRs [7]. Because of these considerable difficulties on the one hand, and the essential role of MRs on the other hand, physics educators and researchers advocate explicitly

*pklein@physik.uni-kl.de

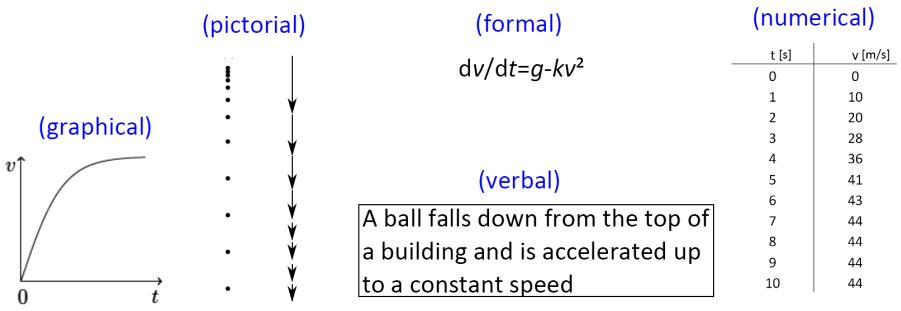


FIG. 1. Coherent representations of an isomorphic kinematics problem: velocity-time graph; strobe picture (the dots record the position at equal time intervals), and vector picture (acceleration or net force vectors); differential equation (g = Earth's acceleration, k = constant); verbal problem statement and data table.

addressing them in physics instruction [4]. For kinematics, in particular, three decades of intense research have shown that the subject poses well-known obstacles for proper understanding [8,9]. Yet, kinematics is an essential basis for understanding the entire domain of mechanics and of the later physics curriculum as it provides conceptual and mathematical foundations for all areas of physics (e.g., concept of force, momentum and energy, symmetry and conservation principles, etc.). For example, misconceptions regarding rotational motion are directly related to improper understanding of linear kinematics [10].

A look into common physics textbooks is sufficient to see that kinematics is a domain that is particularly rich in MR, and in learning difficulties connected to them. For this and the above reasons, information about students' RC in this domain is highly valuable from both a diagnostic and educational perspective. In order to quantify whether instruction is successful in fostering the use of MRs, an adequate tool for assessing students' RC is needed.

B. Concept inventories in kinematics: Gap in existing instruments

Several research-based concept inventories have been developed to test students' understanding in different domains of introductory physics. Regarding kinematics, some of the most widely used tests are the Test of Understanding Graphs in Kinematics (TUG-K) [11], the Force Concept Inventory (FCI) (in particular, the kinematics items of the FCI [12]), and the Force and Motion Concept Evaluation (FMCE) [13]. As Nieminen *et al.* point out, these tests "do not permit comprehensive evaluation of students' skills in using multiple representations" [14]. In fact, RC can be seen as a kind of *hidden variable* in common PER-based instruments such as the FCI [4].

As a consequence, Nieminen and his colleagues developed the representational variant of the FCI (R-FCI) to evaluate students' ability to use different representations consistently across isomorphic, i.e., context- and content-identical, items. In particular, they derived vectorial and graphical variants from nine original FCI items primarily focusing on the concept of force. This development was a very important step towards understanding the role of MRs for learning physics by analyzing representational consistency in a fixed domain. However, kinematics is thoroughly underrepresented within the R-FCI as it is reduced to questions about constant speed and uniform acceleration. As its name states, the focus of the FCI is on the basic understanding of the relationship between force and acceleration and other aspects of the concept of force and, therefore, it does not probe kinematic concepts as such. Furthermore, the R-FCI uses graphical representations of vectors only to distinguish between different magnitudes of forces while orientation is not an issue. Last, five items use bar graphs which are not a current and adequate representation for describing kinematic quantities in university mechanics courses.

The TUG-K was developed to uncover difficulties with interpreting kinematics graphs. While studies using the TUG-K have revealed important misconceptions (such as slope or height confusion, graph as picture, area or slope errors), the TUG-K is limited in its representational formats (primarily graphs, verbal descriptions, and numbers derived from simple calculations). For example, connecting graphs to real-world events described by pictures with embedded data (such as strobe pictures or trajectories) is another important concept that should be tested within the scope of RC [7].

Physics education research (PER) has created several diagnostic tools or items which cover some parts of RC but there is no inventory that addresses RC in kinematics systematically [14] and, in particular, none at the introductory university physics level. Moreover, some important concepts, such as the choice of coordinate systems and how this affects motion graphs, are not explicitly addressed in the inventories mentioned above. Therefore, the present paper proposes an instrument focusing on RC in kinematics, synthesizing and extending previous research, and including new methodological aspects.

The paper is structured as follows: In Sec. II, we report on the development process which was led by the methodology standards of inventory development by Lindell *et al.* [15] and inspired by other work related to

concept inventories—especially by FCI [12], TUG-K [11], and RCI [16]. We discuss the different item formats and the subjects covered by the KiRC inventory based on insights of physics education research concerning learning with MRs. In Sec. III, we give general information about the data collected for test validation. In Secs. IV–VI, we present the methods and results of test validation, including student interviews, item and test statistics, confidence measures, and pre-post comparisons. We conclude with a brief discussion.

II. INSTRUMENT CONTENT AND DEVELOPMENT

A. Development

The development started with a literature review on existing test instruments related to representations and kinematics [4,7,10–12]. Findings of physics education research concerning MRs had a strong influence on the test structure. It has been shown that student performance differs significantly between different representations of nearly isomorphic statements, e.g., graphical and pictorial representations [17]. This implies that basic physical concepts should be assessed including multiple representational formats and relationships between them in order to test explicitly for RC understood in this sense.

Furthermore, even small contextual changes may influence students' problem-solving performance given the same representation format [18]. Therefore, the context of each question was a consideration, as every question has to refer to some object (the referent). We decided to avoid framing a given item with a specific real-world setting, because this might influence student problem-solving performance. Instead, we are testing RC in an abstract context (which can be interpreted as a particular kind of context).

On this basis, an initial pool of over 50 items was developed and then checked by discipline experts (faculty members, lecturers, and postgraduates) for content validity. The test was administered to 80 introductory physics students in Fall 2012 and an item analysis was carried out. Using its result, several items were deleted or modified and the procedure was then repeated with three more cohorts ($\approx$300 students) in the following academic years.

The final version of the KiRC inventory contains 18 items comprising the following representation formats: graphical (g), formal (algebraic expressions and equations) (f), and pictorial (p) as these proved to be the most common and most important representations in kinematics. The final test items are provided as Supplemental Material [19]. Table I presents short descriptions of each item and item characteristics. Each item stem is given in some representation format A and the alternatives are given in representation format B. If $A \neq B$, translation between representations is necessary in order to answer the question. Table I shows which items require either interpretation of a single representation (g,f, or p) or translation between two of them. Textual representations (written language) are included in the problem statement of every single item and there are no text-only items. The agreement between this classification was assessed by four independent raters. After each rater assigned one of the six categories (p, g, f, $p \leftrightarrow f$, $p \leftrightarrow g$, and $g \leftrightarrow f$) to each item, Fleiss'κ_F was calculated as a measure of interrater agreement [20]. We found $\kappa_F = 0.86$ which can be interpreted as an almost perfect agreement [21].

TABLE I. Item format, item content, and representation formats (RF) covered by the KiRC inventory: primarily formal (f), pictorial (p), graphical (g), and transitions between them ($g \leftrightarrow f$, $p \leftrightarrow g$, and $f \leftrightarrow p$). Test items are provided as Supplemental Material [19].

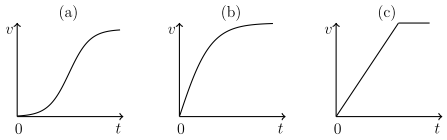
Q.No.	Format	Concept or topic	RF
1	MC3	nonconstant acceleration or relationship between quantities	f
2	MC3	constant acceleration or interpretation of $v(t)$-diagram	g
3	MC3	terminal velocity or air resistance	p
4	MC3	uniform motion (piecewise) or connect $x(t)$ and $v(t)$ graphs	g
5	MC3	uniform motion in a reference or graph interpretation	$p \leftrightarrow g$
6	MC3	reversal of direction or ball on a track	$p \leftrightarrow g$
7	MC3	terminal velocity or air resistance	g
8	MC3	nonconstant acceleration or ball on a track	$p \leftrightarrow g$
9	MC3	angular velocity and speed or rotational kinematics	$p \leftrightarrow f$
10	MC3	vector subtraction or rotational kinematics	p
11	MC3	reversal of direction or ball on a track	$g \leftrightarrow p$
12	MTF	uniform motion (piecewise) or graph interpretation	$g \leftrightarrow f$
13	MTF	vector components or comparing two trajectories	f
14	MTF	vector decomposition or inclined throw	f
15	MTF	terminal velocity or air resistance	f
16	MTF	nonconstant acceleration or relationship between quantities	$g \leftrightarrow f$
17	MTF	vector decomposition or projectile motion in a reference	$p \leftrightarrow f$
18	MTF	angular velocity and acceleration or rotational kinematics	$p \leftrightarrow f$

In the following, we describe item content and item formats covered in the KiRC inventory. Two example items are presented in Fig. 2.

B. Content

We performed a cognitive analysis of the concepts of linear and circular kinematics and identified the most

Item 7. At the instant $t = 0$, a ball is released from rest from the top of a building. The ball accelerates down because of gravity, but eventually attains a constant speed due to air friction. Which of the following $v(t)$-graphs correctly describes the situation?



How confident do you feel about the correctness of your answer?

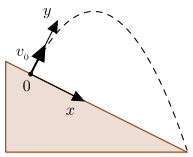
(1) I am very confident in my answer.

(2) My answer is probably right, but I'm not certain.

(3) I feel unsure about my answer.

(4) I guessed.

Item 17. At the instant $t = 0$, a sphere is shot vertically with respect to an incline:



Let v_0, x_d and t_f denote launch speed, throw distance and time of flight, respectively. Consider the Cartesian coordinate system (x, y) as indicated above (with x-axis parallel to the incline) and decide whether the following statements are true or false:

		true	false
(a)	The initial velocity v_0 of the sphere has no component in the x-direction.	◯	◯
(b)	Maximum y-coordinate y_{max} is reached at the instant $t_f/2$, i.e. after half of the time of flight.	◯	◯
(c)	$y(x)$ is a parabola.	◯	◯
(d)	The sphere is accelerated in x-direction.	◯	◯
(e)	The y-coordinates at the start and end are equal, i.e. $y(t = 0) = y(t_f)$.	◯	◯

FIG. 2. Two example KiRC items: Multiple choice item in graphical representation format about free fall with air resistance presenting three alternatives and a confidence rating scale (left), and multiple true-false items dealing with the concept of frame of reference (right).

important concepts about motion according to physics textbooks, including, e.g., the relation between kinematic quantities [position $x(t)$, velocity $v(t)$, acceleration $a(t)$, and time t], several special cases (e.g., uniform motion, inclined throw, fall with air resistance), the superposition principle, explicitness of coordinate system choice, transition and parallels to rotational kinematics, angular velocity and acceleration, etc. The content of all items was chosen in order to match the central concepts treated in an introductory physics course, as based on the literature review and on our findings in the pilot studies. The items necessitate inferences based on understanding of the physics content; memorized procedures are not sufficient to solve them correctly. The items use technical language and their solution requires understanding of the concepts in question and of the representation formats specific to the physics content (in particular, the f and g formats). This makes the KiRC inventory different from concept inventories such as the FCI. In the following, we describe the specific concepts which are addressed by the items.

Students have difficulties with the mathematics of the formal relationship between $a(t)$, $v(t)$, and $x(t)$, with tasks such as substituting and solving equations, differentiation, and integration [22]. These aspects are addressed in questions 1, 13, and 14. In question 1, students have to identify the resulting motion $x(t)$ given a nonconstant acceleration $\dot{v}(t) = b \cdot t$. Apart from the correct answer $x(t) \propto t^3$, two alternatives are presented that look familiar to students, $x(t) = ½bx^2$ and $x(t) = bt$. Questions 13 and 14 show two arbitrary position vectors $\boldsymbol{r}_1$, and $\boldsymbol{r}_2$ and a position vector of an inclined throw $\boldsymbol{r}_3$, respectively. Students have to decide whether a set of given formal properties is fulfilled or not [such as $\boldsymbol{r}_1(t) = \boldsymbol{r}_2(-t)$] and they have to evaluate expressions component-wise [e.g., evaluating $\ddot{\boldsymbol{r}}_3(0)$].

Questions 2, 4, 12, and 16 have been designed to probe students' understanding of graphs in a kinematics context, an area in which previous research has revealed considerable difficulties for students [11]. Given a position-time graph, two velocity-time graphs, and an acceleration-time graph, students are asked to determine (changes in) velocity and acceleration qualitatively, to select another corresponding graph, and to judge verbal descriptions concerning the motion process. Mathematical background such as plotting points and computing slopes is not tested. Items 2 and 4 are adapted versions from two TUG-K questions [11].

Connecting such kinematic graphs to real-world events also causes difficulties for students [7]. This transfer between realistic and schematic representations is addressed in questions 6, 8, and 11 in terms of an object sliding down differently shaped paths. Concepts addressed by the items are the distinction of the shape of a path graph from that of a velocity graph, determination of the rate of change in velocity, and graphical interpretation of a reversal of direction. Incorrect alternatives address well-known misconceptions identified in the development of the TUG-K, such as confusion of height and slope of a graph, confusion of velocity and acceleration, etc.

Questions 3, 7, and 15 deal with the special situation of free fall with air resistance. While the item stem is identical in each question, students have to choose correct pictorial, graphical, and formal representations of the process.

The pictorial representation is given as a strobe picture, such as it is known from some FCI items.

Questions 5 and 17 deal with the concept of coordinate systems and explicitly discuss the choice of a frame of reference—a concept in which students show serious gaps of knowledge [23,24]. In question 5, a reference frame is given and students have to identify the correct $x(t)$ and $y(t)$ graphs. Question 17 shows a ball which is thrown perpendicularly to an incline. The situation has to be described formally within a reference frame oriented along the inclined plane. This task is known to be very difficult for undergraduate students [25].

Mashood and Singh identified difficulties with rotational kinematics and concluded that some misconceptions parallel those reported in linear kinematics, e.g., velocity-acceleration confusion [10]. They developed a concept inventory covering various conceptual aspects of angular velocity ($\boldsymbol{\omega}$) and acceleration ($\boldsymbol{\alpha}$). In questions 9 and 18, we adapted some of these items and put emphasis on the transition between pictorial and formal representations. In question 9, students have to compare the angular velocities of specific points on the minute and second hand of a clock. In question 18, students are shown a particle in circular motion with increasing angular velocity. The item asks for the magnitude and the direction of $\boldsymbol{\omega}$ and $\boldsymbol{\alpha}$, respectively. Question 10 complements this topic by asking about the magnitude and orientation of the acceleration vector while a car accelerates in a bend of the road.

C. Format

We decided to include two different item formats in the inventory: multiple choice items consisting of three alternatives (denoted as MC3) and multiple true-false items (MTF); cf. Fig. 2. MTF items have a stem (lead-in statement) and five independent options that can be answered as either true or false [26]. The MTF format is very efficient in terms of item development and examinee reading time. It has been shown that MTF items produce higher reliability estimates compared with complex multiple choice (CMC) items—such as those used in the original TUG-K—and that students prefer this format compared with CMC [26,27]. However, a better understanding of students' misconceptions might be obtained using MC items in which the false alternatives serve as plausible distractors. Though most concept inventories use a four- or five-option MC format, it has been shown that administering three-option MC items has no detrimental effects on the psychometric quality of test scores [28]. In conclusion, it appears useful to combine both types of formats MC3 and MTF with their complementary advantages.

However, when comparing scores of MC3 and MTF items, it is necessary to take into account the different chance of guessing correctly (33% and 50%, respectively). In this paper, the two item groups are analyzed separately but it seems convenient to calculate a combined score; cf. Sec. V B 1.

The KiRC inventory also assesses how well students are able to estimate the correctness of their answers [29]. For each response i to a MC3 questions, students were asked to rate their confidence in their choice C_i on a four-point Likert-type scale (students were told to choose $C = 0$ if they were guessing completely by chance, $C = 1$ if they were certain, and $C = 0.33$ or $C = 0.66$ as unsure and doubtful values). This choice reflects the examinee's belief in the correctness of the alternative marked and can be interpreted as one aspect of metacognition, viz. students' ability to evaluate their own understanding [16]. Being able to distinguish between right and wrong answers may be a condition to reflect and regulate learning.

III. METHODS: STUDY FRAMEWORK, SAMPLE, AND TEST ADMINISTRATION

In this section, we give general information about the methodology of the study (framework, sample, and test administration). For validity, reliability, and other instrument characteristics, details of the methodology are presented together with the results based on them (Secs. IV–VI).

A. Framework

The presented KiRC inventory data were obtained from an introductory mechanics course at a German university (experimental physics I). This course is taken in the first term mainly by physics majors and includes two weekly lectures (90 min each) and one weekly recitation (90 min, starting in the third week). The course structure is similar to most introductory mechanics courses at German universities. During recitations, students discuss physics problems which they have solved in the past week as homework assignments. Typically, these homework problems are similar to end-of-the-chapter problems of traditional textbooks, very similar to the standard calculus-based courses in the U.S. The development of the KiRC inventory was part of a research project that aimed at modifying these traditional textbook problems to video-based problems, being particularly rich in MR [30]. Used as pre- and post-test, the KiRC inventory yield information about the effectiveness of the new material in comparison to a control group. It is not the aim of this paper to describe the specific instruction, but examples of the video-based problems used on the intervention can be found in Ref. [31].

B. Test administration

The KiRC inventory was administered after the fourth lecture (in German language which is the native language for the vast majority of test takers). The pretest took place in the very first recitation before any homework assignments were solved by students. Weekly recitations are mandatory for students; thus, the sample can be considered to reflect

the population of physics freshman. There were no incentives for taking the test. Up to this point, the formal basics of kinematics and rotational kinematics had been taught and lecture demonstrations of different types of motion had been shown. The lecture put emphasis on the formal relationship between kinematic quantities (position $\boldsymbol{r}$, velocity $\boldsymbol{v}$, and acceleration $\boldsymbol{a}$) and their representation in diagrams. Several physics paradigm cases were discussed and the superposition principle was illustrated with a cart-ball experiment [32]. The post-test was administered six weeks later, i.e., after 12 lectures. Between both measurements, students engaged in weekly recitations in which they applied their knowledge to solve physical problems.

C. Sample

Data analysis is based on the main sample from the winter term 2014–2015. A total of 91 students were enrolled in the course, from whom 83 took the pretest, 46 took the post-test, and 44 took both tests. The population (61 male, 22 female) can be characterized as high achievers in school mathematics and physics as indicated by averages of about 80% of the maximal score [11.8 points in mathematics and 12.0 points in physics on the scale from 00 (lowest) to 15 points (highest) used for the final grading of the German academic track school "Gymnasium"]. These results refer to the average grade obtained over a period of usually two years.

The KiRC inventory was also administered to students in the winter term 2016–2017. All test items were administered but participation in testing was voluntary ($N = 38$). Because of this limitation, we use this additional data only to base item and test characteristics on a larger sample (replication of results).

IV. VALIDITY

A. Convergent validity, as indicated by correlation to related measures

1. Methods

In addition to the KiRC inventory, seven kinematics items of the FCI were administered to the students to consider a correlation with the KiRC scores. Although research suggests to use nothing but the total FCI score, we decided to use the kinematics items only to keep the total test time reasonable. In addition, we considered performance on the exam (class test at the end of the semester) in order to obtain evidence about convergent validity [33] as an aspect of construct validity. The total administration time (KiRC test and FCI items) was 30 min. The missing values were less than 5%.

2. Results

We calculated the product-moment correlation between KiRC scores and FCI scores, and between KiRC scores and exam performance to check convergent validity; see Fig. 3. The FCI can be considered as one of the most often used and best validated instruments in PER. While sharing the same subject, its kinematic items are different (see Sec. I B) but presumably related to what the KiRC inventory is intended to measure. Therefore, we expected a significant but not too high correlation between the scores. The values of $r = 0.35$ and $r = 0.43$, referring to both item groups MC3 and MTF, respectively, are significant at the $p = 0.01$ level $(df = 83)$ suggesting that both measures address similar latent variables.

The exam was taken at the end of the semester and contained five end-of-chapter problems of introductory mechanics. One of these problems was all about kinematics—students had to derive the trajectory of a projectile (inclined throw) given some initial conditions and were asked to calculate quantities such as time of flight, flight distance, etc. As Fig. 3 shows, performance on this specific kinematics task correlates with both test components, MC3 ($r = 0.44$, $p < 0.01$) and MTF ($r = 0.39$, $p < 0.01$), respectively. We also found surprisingly high correlations between the KiRC scores and the total exam score, emphasizing the important role of representation competence in problem solving.

Moreover, we determined the dependence between the KiRC scores and prior achievement in school mathematics and physics for similar reasons. Mathematical representations are ubiquitous and fundamental in physics; facility

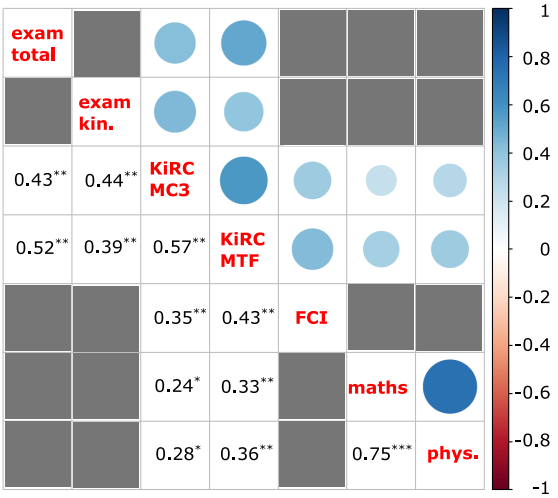


FIG. 3. KiRC test correlations. Note that MC3, MTF, and FCI measures were taken at the same time (pre-instruction). Math and physics scores reflect achievements in school, the exam score refers to performance on the total exam taken at the end of the semester (performance on one exam problem dealing with kinematics exclusively was investigated separately). Pearson's r is reported ($df = 83$, $*p < 0.05$, $**p < 0.01$, $***p < 0.01$)—the p value refers to correlations significantly greater than zero. Irrelevant correlations have not been calculated and are indicated by the gray space.

with mathematical representations (indicated by high preparation scores) is linked to facility with others [34]. In fact, the correlation between math and physics scores is very high ($r = 0.75$, $p < 0.001$). As expected, both scores correlate with KiRC performance ($r = 0.24$, $p < 0.05$ up to $r = 0.36$, $p < 0.01$); see Fig. 3.

Finally, we were interested in the correlation between the two item groups within the KiRC inventory (MTF scores and MC3 scores). The correlation $r = 0.57 (p < 0.01)$ and the fact that both item groups address similar latent variables legitimate combining both scores to generate one total score as mentioned in Sec. V B 1. This makes reporting of KiRC results more practical for instructors.

B. Construct validity, as indicated by student interviews

1. Methods

The construct validity of the KiRC inventory was further evaluated through student interviews. Construct validity refers to the appropriateness of inferences made on the basis of test results. In order to gather evidence for this, we used semistructured interviews and made use of think-aloud procedures to gain insights into student thinking and reasoning when they were answering a subset of test items. If construct validity is given, students should actually be engaging in the thought processes that one would expect for the questions and they should be accessing the appropriate knowledge regarding representations of kinematics. These interviews also provide information if the questions are being understood in the way that was intended.

After taking the pretest, students were divided into high and low achievers according to their test performance. We selected two students of each category as participants for the interviews in order to contrast their problem-solving strategies and representation use (maximum variation sampling [35]).

Each interview lasted approximately 20 to 30 min and was divided into two parts: First, students were presented five KiRC test items (including correct and incorrect alternatives) which they have already answered, and they were asked to elucidate their thought process. Second, students were given item 6 as an open-response item without any alternative. This particular item shows a ball on a racetrack, and students were asked to draw a corresponding $v(t)$ and $a(t)$ diagram. While doing so, students verbalized aloud their thoughts as they emerged. This item is very useful in assessing students' RC as it requires a profound understanding of the relationships between kinematic quantities and how they are visualized in diagrams. Moreover, students have to reason about the coordinate system, and how changes in direction affect motion graphs.

The interviews were audio taped and any notes that students made during the interview were collected.

2. Results

Individual follow-up interviews with four university students were conducted to check construct validity. In the following, we will focus on the reasoning behind three KiRC items (item 10, 15, and 6). We provide parts of the interview transcript and contrast the answers and reasoning of a high achieving student (H1, total KiRC score 0.89) with a low achieving student (L1, total KiRC score 0.45). Then, we sum up the findings during all four interviews in Table II. Please note that the interviews were originally conducted in German language. The protocols were translated afterwards. Language errors have been corrected to improve readability.

Item 10.—Item 10 (see Supplemental Material [19]) was answered incorrectly by L1 in the pretest. He was asked to reread the question carefully and to explain his response.

L1: I think the correct answer is (a).

I: Why do you think so?

L1: The arrows,… they point in the center of the curve. This is very typical for a circular motion.

I: Ok. What about the alternatives (b), and (c)?

L1: I don't know. I've never seen arrows going like this. I'm familiar with (a) because that's how the forces are directed in circular motion and since force is equivalent to acceleration, I think (a) is correct. Otherwise, the car would be thrown off the track.

I: Have a look at the sketch above. Do you know what these arrows represent?

TABLE II. Representational competence indicators related to three KiRC items and if they appeared among the four interviewees. Positive indicators are denoted with a plus (+), whereas indicators opposed to expertlike representation use are denoted with a minus (−).

	L1	L2	H1	H2
KiRC score	0.45	0.50	0.89	0.91
Item 10				
+Correct answer			+	+
+Graphical approach to derive acc.			+	+
+Distinguished vector components		+	+	
−Confusion with uniform circ. motion	−			
−Confusion with linear kinematics	−			
Item 15				
+Correct answers (out of 5)	2	3	5	4
+Referred to a (mental) representation		+	+	+
+Drew an (external) representation			+	
+Wrote down additional equations		+		+
+Verbalized the meaning of equations			+	+
Item 6				
+Correct graphs (out of two)	0	1	2	2
+Defined a coordinate system			+	
+Related graphs and picture	+	+	+	+
+Related graphs among each other			+	+

L1: Yes, they show the speed at certain points, I assume.

I: Can you interpret this picture?

L1: What do you mean? I've read the car goes faster in the U-turn, that's why they become bigger.

I: Correct. How can you relate this sketch with the alternatives below?

L1: Well, if velocity becomes bigger, … maybe the correct answer is (c), because acceleration becomes bigger? Wait a second. If speed increases uniformly, then acceleration must be constant. This means all arrows have the same length. I'm confused.

The response of student L1 is mostly based upon prior experiences with circular motion which has been cued by alternative (a). He fails to connect both vector representations, e.g., he does not try to elaborate changes in the velocity diagram to make conclusions about acceleration. Furthermore, he confuses the relationship between velocity and acceleration, and makes invalid conclusions from linear kinematics.

In contrast, student H1 recognizes that a tangential component of acceleration is necessary to increase the velocity of the car in the way shown.

H1: *Choice (b) must be correct, because (a) represents a uniform circular motion and (c) shows no acceleration which could speed up the car tangentially.*

I: Ok. What can you say about the picture above? Can you relate it to your answer?

H1: Yes, we see that the velocity arrows become longer. If we connect the peaks of two consecutive arrows, we can see that this arrow does not point in the center. Hence, (b) must be correct.

Moreover, student H1 interacts with the graphical representation of vectors and provides an answer based upon the interpretation of the problem sketch. He explains his strategy of vector subtraction to find the resultant acceleration arrow.

Item 15.—We now turn to the discussion of item 15 (see Supplemental Material [19]). The interviewees were asked to step through the five true-false questions and explain their answers, before the interviewer makes any inquiries.

L1: *(a) is true. This equation holds for a free fall and I think the ball falls free during the very first seconds. Air resistance occurs later as the text states.*

(b) *is also true because there is a constant force acting on the ball.*

I'm not sure about (c). This equation is also true for a free fall. But after a long time of falling, I don't think we observe a free fall. As I said before, free fall occurs for short times. I conclude that (c) is wrong. I don't know about (d), I would have to calculate this limit. In the test I marked true but I can't remember why.

(e) is false. The ball has to speed up at first. Then, after a certain time, it reaches maximum acceleration.

I: Okay, thanks. You spoke about constant forces. Can you explain which forces are acting on the ball?

L1: Gravity and some resistance force. Well, resistance is not constant, but gravity is.

I: True. What can you say about air resistance?

L1: Well, it is like … it increases with time. It hinders motion.

The reasoning of L1 again involves considerations of familiar cases, such as the free fall with no air resistance. This clearly helped him to answer questions (a) and (c). He certainly showed a good understanding of the forces acting on the ball, e.g., that air resistance is not constant and opposed to the direction of motion. Given his thoughts on (d) and (e), it seems as if he does not have a coherent picture of the motion process in mind. In particular, he did not refer to any other kind of internal or external representation other than those provided by the text.

In contrast, student H1 verbalized the meaning of the equations and connected the statements to a diagram he had in mind. He pictured the motion process mentally and referred to special motion phases when necessary. As the protocol supports, H1 was able to switch between different types of representations, internally and externally.

H1: *(a) is true. The ball starts with zero velocity and accelerates with g. Since air resistance can be neglected for small velocities, the velocity time-graph would show a straight line. In fact, this could be seen in an item before. Let me just draw the graph [he drew a $v(t)$ diagram].*

(b) This cannot be true since the velocity time graph is curved.

(c) cannot be true either. I don't know how exactly $x(t)$ can be expressed but when force equilibrium is given, the ball drops uniformly. This means, there are equal distances between equal time intervals and nothing quadratic.

(d) means the ball is no longer accelerated after a certain time. This is true.

(e) This is true because air resistance increases with velocity, and it is opposed to Earth's acceleration. The maximum acceleration is g.

I: Okay, thank you. You drew a velocity-time graph. Did it help you answering the question?

H1: Indeed, I had this graph in mind from the very beginning of reading the text. Then some questions became easy.

I: Without thinking about the graph, can you find another argument why question (b) is false?

H1: Yes, this equation means that acceleration changes over time as I said before.

I: Excellent. You also spoke about force equilibrium. Please explain that.

H1: While the ball is accelerated by Earth's gravitation, air resistance increases dynamically. The ball therefore exposes a net force, given by the difference of these two effects. After a certain time, a maximum velocity is reached, just as if a ball is dropped in oil. Then there is no acceleration.

Item 6 (open response).—Students were provided item 6 without alternatives and they were asked to draw $v(t)$ and $a(t)$ diagrams corresponding to the motion.

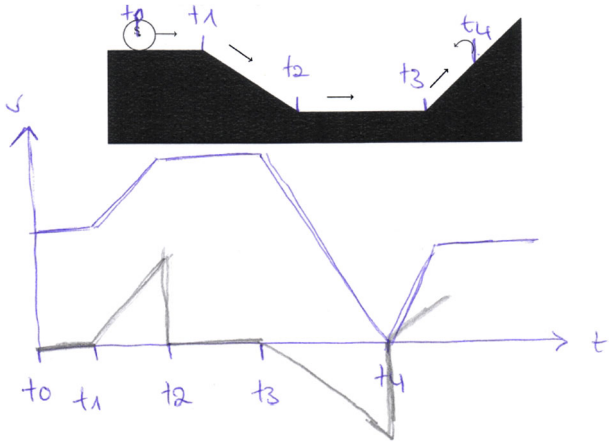


FIG. 4. Item 6 related sketch drawn by student L1 during the interview. Velocity-time graph was drawn first (ballpoint pen, blue), acceleration-time graph subsequently (pencil).

Figure 4 shows the solution from student L1. He started to label crucial points t_0–t_4 in the sketch and drew the coordinate system beneath. First, he drew the velocity-time graph and argued as follows:

L1: We start with some positive initial velocity from zero to one. Then, the ball rolls down the incline until t_2. It gets accelerated, hence velocity increases. Then the path is smooth, hence velocity is constant. Afterwards the ball rolls up the hill until t_4. At this instant velocity is zero.

I: Okay, can you please go on?

L1: Uhhm, then velocity increases again, like this [points to the sketch]. And after some time, it will be constant again.

From the protocol, we can conclude that the student has some basic understanding of the relationship between the shape of the track and the velocity of the ball. However, at the crucial instant t_4, the participant acts hesitantly and draws a wrong solution. When the ball reverses direction, velocity becomes negative. Instead, he considered only the magnitude or speed of the object. The participant did not determine a reference system at all, and started with sketching the graphs intuitively. When L1 drew the acceleration-time graph, major mistakes occurred:

L1: *First, acceleration is zero, because velocity is constant. Then acceleration increases, because it rolls down the hill. Then it's zero again until t_3. When the ball slows down, acceleration becomes negative and then,…*

I: You hesitate. Do you need help?

L1: *I don't know how to proceed here. Because at this instant [he points at t_4] velocity is zero and increases again. I have no idea how to sketch this.*

Again, the interviewee was confused with the relationship between velocity and acceleration. When drawing the graph, he was interacting with the picture and tried to draw a continuous line. He did not attempt to make a connection between the both graphs and did not mention any graphical relationship between them (e.g., acceleration as derivative).

In contrast, participant L1 explicitly mentioned the orientation of a one-dimensional reference system and drew the correct graph. He derived the $a(t)$ diagram directly from his $v(t)$ diagram and related it afterwards to the picture.

Synthesis of interview results including all participants.—During the interviews, we have experienced problem-solving behaviors which clearly indicated differences between representational abilities. Experts (as indicated by high KiRC scores) showed a multiplicity of representation use, very dense in time, and were able to switch between them when necessary. We have listed the most important indicators of this behavior in Table II and checked whether they were present on individual level. We also list common mistakes that are opposed to expert-like thinking.

Upon the interviewees, only the two high performers were able to use representational strategies consistently. They typically begin by describing problem information and using that information to decide on a solution strategy before giving the answer. It is worth noting that none of the interviewees struggled with the symbolic representations dv/dt, $\lim_{t\to\infty}$, etc.

V. RELIABILITY AND INSTRUMENT CHARACTERISTICS

A. Item and test analysis

The purpose of item and test analysis is to examine whether the KiRC inventory is reliable and discriminating, i.e., if the test produces similar results under consistent conditions and whether the results can be used to distinguish high from low achievers.

1. Methodology

Classical test theory provides a number of measures for test evaluation, from which we use three as measures of individual test items j as suggested by Ding and Beichner [36]: item difficulties P_j, discrimination indices D_j, and item-test correlation (point-biserial coefficients) r_{jt}. Furthermore, we report the Kuder-Richardson-Index (KR20) as a measure of internal consistency of all test items. Confidence intervals for each test statistic were computed according to standard procedures. Definitions and a short description of each measure are given in Appendix A, as well as procedures for computation of confidence intervals.

2. Results

Table III shows item difficulty index, discrimination index, and item-test-correlation (point-biserial coefficient) for each item of the KiRC inventory (MTF item satistics consume the averages of the substatements). In addition to these individual measures, the Kuder-Richardson reliability index r_{test} provides insights into internal consistency of the

TABLE III. KiRC inventory pretest item statistics: item difficulty index (P_j), item discriminatory index D_j, and item-test-correlation (r_{jt}) for each item j. For each MTF cluster (items 12–18), statistics were averaged.

MC items				MTF items			
Q.No.	P_j	D_j	r_{jt}	Q.No.	P_j	D_j	r_{jt}
1	0.20	0.48	0.49	12	0.74	0.24	0.25
2	0.70	0.68	0.60	13	0.81	0.30	0.28
3	0.80	0.34	0.40	14	0.74	0.39	0.33
4	0.81	0.43	0.39	15	0.62	0.49	0.38
5	0.65	0.67	0.51	16	0.72	0.41	0.34
6	0.43	0.77	0.63	17	0.54	0.33	0.26
7	0.52	0.59	0.35	18	0.53	0.52	0.43
8	0.71	0.53	0.43				
9	0.62	0.48	0.48				
10	0.41	0.63	0.49				
11	0.60	0.73	0.59				

entire test, respectively [36]. These data as well as average values and confidence intervals are provided in Table IV along with the desired ranges. It is worth noting that the desired item difficulty range suggested in Ref. [36] has been modified because of correction for guessing, see Sec. V B 1. New desired ranges for MC and MTF items were obtained by applying formula scoring to the upper and lower bounds of the original range. In the following, we discuss the results of these five statistics.

Item difficulty.—As Table III shows, the difficulty index ranges from 0.2 (item 1) to 0.81 (item 4, and item 13 average). The KiRC items therefore cover a reasonable range of difficulty with averaged difficulty indices P of 0.59 and 0.67 for both item groups, respectively (see Table IV). These values can be used as an indication of overall test difficulty and they both fall into the suggested range.

Only one item (No. 1; nonconstant acceleration in formal representation) has a difficulty index lower than 0.3, which is the chance level for a MC3 item. Such a low P_j value does not necessarily imply that the item is malfunctioning but suggests the presence of at least one strong distractor which is very plausible for the students. This may be the case here since both false alternatives of question 1 represent familiar equations of rectilinear motion (cf. Supplemental Material [19]). We decided to keep the item since an improvement over the course of the semester was to be expected.

Item discrimination index.—The overall discrimination of both item groups is satisfactory ($D = 0.58$ and 0.38, respectively). From Table III, we note that all items except item 12 have a good discrimination value. A detailed analysis of the five statements of this MTF item cluster reveals that the subitem 12d has a very low discrimination value ($D_j = 0.05$) which is probably caused by a very high item difficulty index $P_j = 0.97$ (which means that almost all students answered this item correctly). In case of a strong ceiling effect, the difference in performance between the top and bottom quartile is low and the question should be reconsidered. We would recommend to revise this subitem 12d in a further test revision.

Item-test correlation.—As one can see in Table III, all items fulfill the desirable criterion $r_{jt} \geq 0.2$. The average item-test correlation for the KiRC inventory are 0.49 and 0.32 for both item groups, respectively. We comment on the correlation coefficient and provide a interpretation of such values in the framework of formal assessment tools below.

Kuder-Richardson reliability index.—As we can see in Table IV, both values r_{test} of 0.69 and 0.84 are acceptable. A much higher value for a formative assessment tool such as the KiRC inventory would not guarantee that the test is more reliable; in fact, it would indicate that the test comprises redundant questions. As Day and Bonn point out, mediocre internal consistencies are quite reasonable for a test that does not measure a single construct [37].

TABLE IV. Pretest statistics for each item group and for the total score along with the 95% confidence interval of the combined score as well as possible and desired ranges of item statistics.

Statistic	MC3	MTF	Total	95% C.I.[a]	Possible range	Desired range[b]
Mean item difficulty P	0.59	0.67	0.65	[0.61, 0.69]	[0, 1]	MC: [0.53, 0.93] MTF: [0.65, 0.95]
Corrected P'[c]	0.39	0.34	0.35	[0.32, 0.38]	[−0.5, 1]	⋯
Mean discrimination D	0.58	0.38	0.43	[0.38, 0.48]	[−1, 1]	≥0.3
Mean item-test corr. r_{jt}	0.49	0.32	0.28	[0.07, 0.51]	[−1, 1]	≥0.2
Kuder-Richardson reliability r_{test}	0.69	0.84	0.86	[0.80, 0.90]	[0, 1]	≥0.7

[a]For calculation of confidence intervals, see Appendix A.
[b]These ranges were suggested by Ding & Beichner [36] where the original desired range for P reads [0.3, 0.9]. Because of guessing, we adapted the original desired range for P using formula scoring.
[c]In order to compare the scores with different numbers of alternatives per item, a corrected score can be calculated using formula scoring, see Sec. V B 1, Eq. (3).

Synthesis.—We present an overview of the four measures in Table IV and conclude that the KiRC inventory is a satisfactorily reliable and discriminating test. This holds for each item group and for the total score (i.e., MC3 and MTF items put together). Note that apart from the mean item difficulty, the total score is not a linear combination of both group scores since the distribution of raw scores is different when the whole data matrix is taken into account, hence, changing the scores in the quartiles.

B. Pre-post changes, formula scoring (correction for guessing), and test combination analysis

Next, we explore whether the KiRC instrument is able to measure learning gain (in terms of Hake gain and effect size Cohen's d).

When calculating a total test score from two distinguishable test components (MTF and MC part in the case of KiRC), there is some freedom of how these components should be weighted, and, furthermore, whether statistical results are influences by the relative weights of each test component. We present an analytic result for this question —arising from guessing correction or educational considerations—of how the total effect size (Cohen's d) varies upon combination of two test components with known individual effect sizes, and then discuss the results.

1. Methodology

In order to provide an objective measure of learning in introductory mechanics, Hake introduced the normalized gain factor g_H defined as

$$g_H = g_H(P_{\text{pre}}, P_{\text{post}}) = \frac{P_{\text{post}} - P_{\text{pre}}}{1 - P_{\text{pre}}}, \tag{1}$$

where $P_{\text{pre/post}}$ are the mean item difficulties of the pre- and post-tests, respectively [38]. Please note that while possible gains measure the effectiveness of instruction, it is not our goal here to evaluate the introductory physics course which provides the data. As the course proceeded, the necessity to interpret and to work with MRs was increasing, i.e., we could expect positive gains. From the perspective of instrument development, we identify the questions that show a gain over time.

Another important measure of learning outcome is effect size d, defined as

$$d = \frac{P_{\text{post}} - P_{\text{pre}}}{\text{SD}_{\text{pre}}}, \tag{2}$$

where SD_{pre} denotes the standard deviation of pretest item difficulties [39]. Note that d and g_H can also be calculated from test scores (sum of correct responses) with an obvious modification [40] of Eqs. (1) and (2).

In this section, we present two more analysis procedures used in the present paper: (i) Formula scoring, i.e., calculating scores corrected for guessing is useful to compare absolute test values across concept inventories with different numbers of alternatives. (ii) Test combination analysis often occurs when a combined score from two test components is calculated (in our case MTF and MC3), and one wants to determine the guessing correction and effect sizes for the combined test.

Frary [41] assigned a penalty (formula scoring) to account for the different chance of the correct answer being guessed by calculating a corrected test score,

$$T' = FS(T_{\text{raw}}) = \left(T_{\text{raw}} - \frac{1}{k}n\right)\frac{k}{k-1}, \tag{3}$$

where T_{raw} and T' are the raw and corrected scores, respectively, n the number of test items, and k is the number of alternatives for each item. As a linear transformation of T_{raw}, the corrected test scores T' conserve the statistical properties of T_{raw}, such as normalized gain [42].

Turning to a combination of tests with different answer formats, all quantities in Eq. (3) take on an index i ($i = 1, 2$); in our case $i = 1$ for MC3 and $i = 2$ for MTF ($k_1 = 3, n_1 = 11; k_2 = 3; n_2 = 7 \times 5$ for MTF items). After applying formula scoring, both test scores, T'_1 and T'_2, can be combined to a total test score T by

$$T = T'_1 + T'_2. \tag{4}$$

Equation (4) can be rewritten as

$$T = \alpha T_1 + \beta T_2 + \gamma \tag{5}$$

$$= \alpha\left(T_1 + \frac{\beta}{\alpha}T_2\right) + \gamma \tag{6}$$

with

$$\alpha = \frac{k_1}{k_1 - 1}, \tag{7}$$

$$\beta = \frac{k_2}{k_2 - 1}, \tag{8}$$

$$\gamma = -\frac{n_1}{k_1 - 1} - \frac{n_2}{k_2 - 1}, \tag{9}$$

i.e., the total test score can be written as a linear combination of single (raw) test scores $T_{1/2}$ with weights α and β, plus an offset γ. We now turn to calculating effect sizes of combined test scores from Eq. (2). It can easily be confirmed that the prefactor α and the offset γ cancel out when pre- and post-scores are being compared in terms of calculating Cohen's d [43]. Thus, it is suitable to consider the simplified expression

$$T(\kappa) = T_1 + \kappa T_2 \tag{10}$$

with $\kappa \in [0, \infty]$. By definition, the right-hand side of Eq. (10) allows the total test score to run from "T_1 only" ($\kappa = 0$) through "T_1 and T_2 equally weighted" ($\kappa = 1$) to "T_2 dominates" ($\kappa \to \infty$), and includes the special case of formula scoring ($\kappa = \beta/\alpha$). Inserting Eq. (10) into Eq. (2) the effect size of the combined test score reads

$$d(\kappa) = \frac{T_{\text{post}}(\kappa) - T_{\text{pre}}(\kappa)}{\text{SD}_{\text{pre}}(\kappa)}. \tag{11}$$

For normally distributed T_1 and T_2, it can be shown (see Appendix B) that $d(\kappa)$ can be related to the effect sizes d_1 and d_2 of the components (MC and MTF):

$$d(\kappa) = \frac{1}{\sqrt{1 + \kappa^2\phi^2 + 2\kappa\theta_1}} d_1 + \frac{1}{\sqrt{1 + \kappa^{-2}\phi^{-2} + 2\kappa^{-1}\theta_2}} d_2, \tag{12}$$

with

$$\phi^2 = \frac{\text{SD}_{\text{pre},2}}{\text{SD}_{\text{pre},1}} \tag{13}$$

as the variance ratio of the two test components and

$$\theta_i = \frac{\text{COV}(T_{\text{pre},1}, T_{\text{pre},2})}{\text{SD}^2_{\text{pre,i}}} \tag{14}$$

accounting for the variance both test components share together. Equation (12) allows, in particular, to establish lower and higher bounds of d when different weights κ of the test components are allowed (test combination analysis).

2. Results

The pre-post analysis focuses on the changes in performance whereas other measures (discrimination and reliability coefficients) are not compared. We considered the matched sample of $N = 44$ students who took both the pre- and the post-test. Thus, item difficulty levels may be slightly different from those reported in Sec. V A 2.

Table V reports the mean pre- and post-test difficulty indices, the p values obtained from paired t tests on pre- and post-test scores, the normalized gain indices g_H, and Cohen's d. We calculated these statistics for each test component (MC, MTF) separately, for the equal weighted total test score, and for corrected test scores (after applying formula scoring). As the data show, performance on the KiRC inventory clearly changed over the course of one semester. A t test to establish the statistical significance of differences between average pretest and post-test scores yielded highly significant values corresponding to medium up to large effects.

TABLE V. Matched sample ($N = 44$) pre- and poststatistics for test components and total test scores (FS = Formula scoring).

	MC ($\kappa = 0$)	MTF ($\kappa \to \infty$)	equal weight ($\kappa = 1$)	FS ($\kappa = 4/3$)
P_{pre}	0.60	0.64	0.63	0.32
P_{post}	0.72	0.79	0.78	0.59
p value	$<10^{-3}$	$<10^{-8}$	$<10^{-9}$	$<10^{-9}$
Hake-Gain	0.31	0.42	0.39	0.39
Cohen's d	0.57	0.99	1.04	1.04

As demonstrated in Sec. V B 1, the aforementioned test scores correspond to different weights κ in Eq. (12). Figure 5 present $d(\kappa)$ data for 50 different values of κ ranging from 0 to 100 and thus shows how the effect size is influenced by different weights. The effect size ranges from $d = 0.57$ up to $d = 1.04$ achieving its maximum approximately at equal weight of both test scores ($\kappa \approx 1$). Because MTF and MC test scores correlate with each other (cf. Sec. IV A 2), the combined test score yields larger effect size as compared to one single test component. However, there is not much effect size lost if only the MTF pre- and post-test scores are being compared ($d = 0.99$).

In addition to the total test scores, the fraction of correct answers for every individual question increased from the pretest to the post-test, as we can see in Fig. 6. Hence, all normalized gain factors were positive with medium-sized averages (cf. Table V). We conclude that the KiRC inventory indeed works as a tool to capture gain in RC in kinematics.

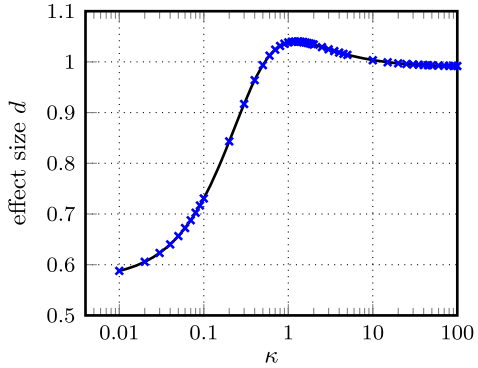


FIG. 5. Test combination analysis: Pre-post effect size (Cohen's d) of composite test score $T = T_1 + \kappa T_2$ as a function of κ. According to Eq. (11), data points were obtained from raw pre- and post-test scores by calculating the composite test score for each student using 50 different values of κ. The fit line was determined by Eq. (12) using θ_1, θ_2, and ϕ calculated from the data. Note that the κ axis is logarithmic in order to illustrate the relevant values of κ according to Table V with sufficient resolution.

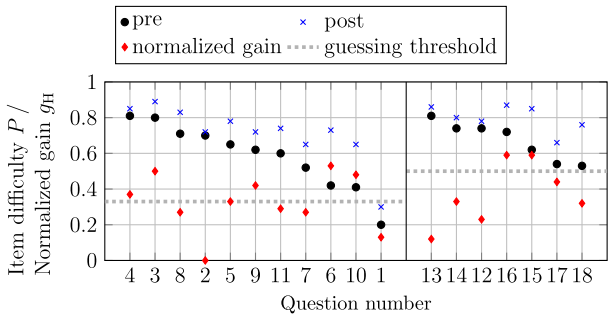


FIG. 6. KiRC results by question for the introductory physics course ordered by pre-instruction item difficulty. The left part of the diagram shows the MC3 items; the right part shows the MTF items.

C. Replication of results

In addition to the winter term 2014–2015 (WT14/15) sample, the KiRC inventory was administered to students in the winter term 2016–2017 (WT16/17) to enlarge the pool of data (cf. Sec. III C). The population of students who enroll in the introductory mechanics course is quite similar from year to year as the university maintains constant admissions criteria. Each year, students have similar preparation for the course, similar learning experience in school and are of similar demographic composition.

TABLE VI. KiRC pretest statistics obtained from the replication sample in winter term 2016/2017 ($N = 38$).

Statistic	Total test value	95% C.I.
Mean item difficulty P	0.68	[0.64, 0.72]
Mean discrimination D	0.39	[0.36, 0.42]
Mean item-test corr. r_{jt}	0.37	[0.05, 0.60]
Kuder-Richardson reliability r_{test}	0.80	[0.69, 0.89]

TABLE VII. KiRC item difficulties obtained in both samples. ANOVA p values indicate chance levels for differences found between the two samples.

	Item	WT14/15 $N = 83$	WT16/17 $N = 38$	ANOVA p-value
MC3	1	0.20	0.40	0.01
	2	0.70	0.63	0.54
	3	0.80	0.71	0.56
	4	0.81	0.82	0.87
	5	0.65	0.82	0.19
	6	0.43	0.55	0.20
	7	0.52	0.66	0.15
	8	0.71	0.76	0.55
	9	0.62	0.68	0.51
	10	0.41	0.42	0.36
	11	0.60	0.50	0.32
MTF	12	0.74	0.78	0.57
	13	0.81	0.88	0.06
	14	0.74	0.81	0.17
	15	0.62	0.67	0.35
	16	0.72	0.68	0.38
	17	0.54	0.54	0.42
	18	0.53	0.44	0.26

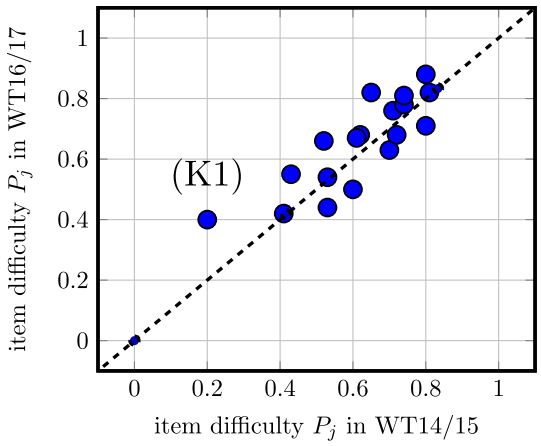


FIG. 7. KiRC item difficulties obtained in winter term WT14/15 and WT16/17. Each point represents one item (item 1 being highlighted). The dashed line represents equal item difficulties. $R^2 = 0.70$.

Table VI shows the mean test statistic obtained in the replication sample. All test values lie within the desired range, and within the 95% confidence intervals calculated from the data of the main sample (cf. Table IV). Because of the smaller sample, confidence intervals reported in Table VI are larger.

We have also compared item difficulties between both cohorts on the level of individual items using analysis of variance (ANOVA). The results in Table VII show that only one item differed significantly in terms of item difficulty between the different cohorts of students (item 1). The similarity of item difficulties between the cohorts is also expressed in Fig. 7, with item 1 being an outlier. Item 1 was solved correctly by 20% of the WT14/15 students and by 40% of the WT16/17 students ($p = 0.01$). There is no evident reason for a difference in this specific item, but it could be argued that the WT16/17 sample performed better in general (average item difficulty of total KiRC test: 0.65 vs 0.68). This may be due to the fact that test participation was voluntary in WT16/17 as stated above. However, item 1 is still the hardest item among all other items in each cohort.

In conclusion, the replication study confirmed a satisfactory discrimination and reliability of the total test, even though the sample size is rather small.

VI. CONFIDENCE ASSESSMENT

The next objective of our analysis was to identify how confident students are with their responses. This information is potentially useful for gauging the quality of students'

understanding, and recently is attracting attention within the PER community [16,44,45].

A. Methodology

From a methodological perspective, an accurate judgment of cognitive skills can be assumed if item difficulty and confidence are sufficiently correlated. In this sense, Shaughnessy provided the CAQ measure (confidence-judgment accuracy quotient), defined by

$$\mathrm{CAQ} = \frac{C(P=1) - C(P=0)}{s_p}, \tag{15}$$

where the numerator is the difference between the mean confidence judgment assigned to items the student answered correctly ($P = 1$) and the mean confidence judgment assigned to items answered incorrectly ($P = 0$), and the denominator is the pooled standard deviation of confidence judgments [46]. If the student could not distinguish between items answered correctly and those answered incorrectly, the CAQ measure would be expected to be close to zero. Several investigations found that the accuracy of confidence-judgments varies among different ability levels [44,47,48]. While well-performing students show CAQ levels significantly greater than zero, low performers do not reliably distinguish between correct and incorrect responses. This phenomenon is well studied in educational psychology and is called the Dunning-Kruger effect [49]. Recently, this effect was investigated and validated in a physics learning context [44]. By analyzing students responses and confidence ratings to the KiRC inventory, we will provide existing research about metacognition with findings related to representations. Furthermore, we will discuss the relationship between confidence and knowledge on the individual item level.

B. Results

The mean overall CAQ score (0.45) is significantly greater than zero, the expected value if students could not discriminate between correct and incorrect answers, $t(69) = 4.00$, $p < 0.001$. To investigate the Dunning-Kruger effect, overall test scores were broken into quartiles and the mean CAQ score for each quartile was determined. The values are 0.87, 0.54, 0.26, and 0.07 from the top through the bottom quartile, respectively. Decreasing mean CAQ scores across performance quartiles indicates that the two measures are positively correlated. In fact, the correlation between overall test scores and overall CAQ scores was significant, $r(69) = 0.29$, $p < 0.05$. This finding is consistent with the Dunning-Kruger effect. Conclusions and implications will be discussed after the presentation of some additional data.

As shown in Table VIII, individual item confidence levels cover a range from about 0.5 to nearly 0.9. The averaged confidence index value C of all items is 0.75 for the KiRC inventory and suggests that students were overall quite confident in answering the questions. This may be due to the fact that kinematics is a quite familiar topic even to freshmen, as argued in Sec. I. This moderately positive level of confidence is also supported by low values for pure guessing ($C = 0$) with questions 9 and 10 identified as outliers (7 and 14 guesses out of 83 responses, respectively).

The high level of guessing in items 9 and 10 may be due to the fact that these questions address rotational kinematics, a subdomain often treated briefly in high school and in corresponding textbooks (which applies to our sample) [10]. These two items also obtain the lowest confidence index of all items, followed by item 1. As one might expect, these questions have low item difficulties P as well; i.e., they are tough (see Table III). We split up the cohort into the students who were certain about their response ($C = 1$) and those who were not ($C \neq 1$) and calculated item difficulties $P_{C=1}$ and $P_{C\neq1}$ for these subgroups, respectively [50]. For items 9 and 10, these specific item difficulties differed significantly, which means that students were aware of their weak conceptual understanding (cf. Table VIII). This is also supported by rather high correlations between item difficulty and confidence. In such cases, a correct instruction would probably be successful.

However, confidence and performance are not related for all items, as Table VIII shows. The average of the Pearson's r correlation between students' confidence of choice and their total score is $r = 0.4$ ($p < 0.01$). We conclude that students are well calibrated in general—on the macrolevel—but some individual items show very low correlations on the microlevel, e.g., items 2, 6, 7, and 11. For these items,

TABLE VIII. KiRC inventory confidence statistics: average confidence index (C), number of guesses ($C = 0$) out of 83 responses, and correlation between confidence and item difficulty r_{CP}. The item difficulties $P_{C=1}$ and $P_{C\neq1}$ refer to answers given with 100% certainty and with less than 100% certainty, respectively. The meaning of the p values is the usual statistical significance level, i.e., the greater p, the higher the probability that the differences between $P_{C\neq1}$ and $P_{C=1}$ were found due to chance.

Q.No.	C	$C = 0$	r_{PC}	$P_{C=1}$	$P_{C\neq1}$	p value
1	0.68	3	0.22	0.27	0.15	0.18
2	0.89	0	0.04	0.71	0.68	0.81
3	0.82	0	0.20	0.87	0.69	0.04
4	0.84	1	0.46	0.89	0.67	0.01
5	0.76	2	0.20	0.70	0.61	0.38
6	0.80	1	0.02	0.41	0.45	0.73
7	0.80	1	−0.02	0.52	0.51	0.90
8	0.69	2	0.20	0.80	0.67	0.25
9	0.68	7	0.31	0.81	0.51	0.01
10	0.51	14	0.18	0.64	0.37	0.06
11	0.76	3	0.08	0.62	0.58	0.69

the students who rated their confidence as certain ($C = 1$) show almost equal solution probability as to those who rated their confidence as uncertain or were guessing. Given low solution probabilities, this characteristic may be an indicator for misconceptions. Indeed, items 6,7, and 11 show the lowest values of P apart from items 10 and 1; i.e., they are tough. A closer look at the response rates of these questions reveals that students struggled with the concept *reversal of direction* (response rates are given in the Supplemental Material [19]). In question 6, 56% of students incorrectly chose the graph which has a "V" with a vertex marking the turnaround. They failed to recognize that a reversal in direction is marked only by crossing the time axis. In question 11, the situation is reversed: Students have to relate a given velocity-time graph in which the time axis is crossed to a picture in which an object reverses its direction. Again, 24% of students chose alternative (b) which would be correct if the velocity-time graph were interpreted as not respecting the sign of velocity. This misconception with regard to the representation of a negative velocity on a v vs t graph has also been reported in other studies [7]. If students were better aware of the concept of constant acceleration, they would probably succeed in these items. The false alternatives in question 7 also address the concept of acceleration. If acceleration uniformly decreases as in this case, the corresponding $v(t)$ graph cannot possess an inflection nor can it be discontinuous. These alternatives were chosen by 37% and 11% of students, respectively. However, the latter misconception assumption is further supported by the fact that items 6,7, and 11 show large gaps between confidence and solution probabilities, as discussed above. A possible interpretation is that students tend to overestimate their performance on these items. In contrast, they are aware of their lack of conceptual understanding with regard to questions 9 and 10.

VII. CONCLUSION AND OUTLOOK

The KiRC inventory was developed as an assessment instrument to probe students' understanding and handling of multiple representations in kinematics. Such skills involve interpreting given representations and translating between them as well as creating their own representations. The latter is not included in the present version of the test, due to its format (MC3, MTF). The current version comprises 18 questions in two different formats: multiple choice questions with three alternatives (MC3 items) and multiple true-false items (MTF). Classical test analysis suggests that the KiRC items have satisfactory difficulty levels, discrimination, and item-instrument correlation. Moreover, the instrument as a whole has satisfactory reliability ($r_{\text{test}} = 0.69$ and 0.84 for MC3 and MTF, respectively, and 0.86 for the total test). Detailed results of the five different statistics are provided in Table IV. A few items appear to be at the lower end of the desirable value range for the population of introductory physics students, but we kept them in the instrument as they probe important components of kinematics-related RC. In order to explore convergent validity, we have shown that KiRC scores correlate with other relevant measures such as the kinematics items of the FCI, exam scores, and prior achievement in mathematics and physics. Problem-solving interviews with four students indicated construct validity of the test: Students with high KiRC scores used representations consistently and changed flexibly between different external representations. In contrast, low performing students failed to incorporate representational strategies in their problem-solving approach.

Furthermore, a pre-post comparison of test scores indicates that the KiRC instrument is sufficiently sensitive and reliable for capturing learning progress and, thus, also for comparing the effectiveness of different instructional approaches in its domain. We have presented an analytic expression to relate the effect size d of the total test to those of the test components (MC, MTF), allowing in particular to establish lower and higher bounds of d when different weights of the test components are allowed (test combination analysis). This procedure can be applied whenever two test components are combined to a total test score and might be useful in other measurement contexts.

Moreover, we used confidence scores to identify misconceptions and misunderstanding on the individual item level. With the ability to quantify the degree of certain misconceptions, future research is able to investigate effects of metacognitive-aware treatments on the confidence-judgment accuracy. Given a larger pool of students, it would be interesting to investigate the stability and persistence of misconceptions under different treatment conditions. On the macrolevel, we have evaluated the overall calibration (taking all items into account) of confidence judgments with respect to performance using the CAQ measure. Our findings support the Dunning-Kruger effect, which states that low performers show no adequate estimation of their cognitive skills but tend to overestimate their ability. It can be assumed that students are better able to develop an appropriate understanding if they reflect correctly on their performance [44]. Evaluation of and confrontation with confidence judgments thus could help them to identify knowledge gaps and could prompt them to deal more extensively with the learning content. As argued in Sec. I this is of particular importance concerning representational skills. An interesting aspect of future research is to investigate the relationship between confidence judgments and other measures such as self-efficacy expectation or self-concept.

We have also looked for potential gender differences and found little to none. Since other researchers reported *gender biases* in concept inventories, it might be interesting to study this aspect on a larger sample with more (female) students.

In conclusion, we feel that KiRC is a useful diagnostic instrument for an important component of introductory university physics learning (and, in some countries, also of the last years of preuniversity learning). Further validation for various target groups will be one of the next steps. We are also considering developing an expanded version of the test which includes the creation of own representations.

ACKNOWLEDGMENTS

P. K. thanks the Wilfried-and-Ingrid-Kuhn Physics Education Foundation for generous financial support.

APPENDIX A: DEFINITION OF ITEM AND TEST MEASURES, COMPUTATION OF CONFIDENCE INTERVALS [36]

Item difficulty.—The item difficulty index P, defined as

$$P = \frac{N_1}{N}, \tag{A1}$$

is a measure of the difficulty of a single question, where N denotes the number of students taking the test and N_1 is the number of correct responses. P can obtain values from 0 (no correct response) to 1 (only correct responses) while the suggested range is between 0.3 and 0.9. However, this range should be adapted for MTF items as mere guessing already yields $P = 0.5$.

Confidence intervals for this measure can be obtained from the standard error of a mean.

Item discrimination index.—The discrimination index D is a measure of the discriminatory power of an item and is defined as

$$D = \frac{N_t - N_b}{N/4}, \tag{A2}$$

where $N_{t/b}$ is the number of correct responses in the top or bottom quartile. D can obtain values between -1 and 1; the higher the value, the better the item distinguishes between more and less competent students. Negative values indicate items that can be better solved by low achievers than by high achievers. These items should be eliminated and the majority of test items should have a good discriminatory power $D \geq 0.3$.

As D is in fact a difference of two item difficulty indices, a confidence interval of D can be obtained from the standard errors of the mean.

Item-test correlation.—The item-test correlation (also referred to as the point-biserial coefficient) is a measure of the consistency of a single item with the entire test and can be expressed as

$$r_{jt} = \frac{\langle X_1 \rangle - \langle X_0 \rangle}{\sigma_x} \sqrt{P_j(1 - P_j)}, \tag{A3}$$

where $\langle X_1 \rangle$ ($\langle X_0 \rangle$) is the average test score for those who correctly (incorrectly) answer the item and σ_x is the standard deviation of the test score. Values of this metric range from -1 to 1, where greater values mean better consistency between an item and the test score.

Since the sampling distribution of the point-biserial coefficient is not normally distributed, the computation of a confidence interval is done by (i) converting r_{jt} to an approximately normally distributed z score, (ii) computing a confidence interval in terms of z, and (iii) converting the confidence interval back to r_{jt} [51].

Kuder-Richardson reliability index.—The Kuder-Richardson reliability index estimates the degree of correlation between students' responses. It is a special case of Cronbach's α, computed for dichotomous scores, and defined as

$$r_{\text{test}} = \frac{K}{K-1}\left(1 - \frac{\sum_i P_i(1 - P_i)}{\sigma_x^2}\right), \tag{A4}$$

where K is the number of test items. A Kuder-Richardson reliability index close to 1 indicates that all questions measure the same concept, which is not desirable for the KiRC inventory since we assume that performance differs across different representations. However, since all questions require an understanding of kinematics, a low index is not satisfactory either.

For calculating confidence intervals of r_{test}, we can refer to the calculation of confidence intervals for Cronbach's α; see Ref. [52].

Pearson product-moment correlation coefficient.—As a measure of the degree of linear dependence between two variables X and Y, the Pearson product-moment correlation coefficient r_{XY} is defined as

$$r_{XY} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}, \tag{A5}$$

where $\bar{X}$ is the mean of X.

APPENDIX B: PROOF OF EQ. (12)

We recall some basic properties for the sum of two normally distributed random variables X and Y (a, b are scalars):

$$\begin{aligned}\text{VAR}(aX + bY) = a^2\text{VAR}(X) + b^2\text{VAR}(Y) \\ + 2ab\text{COV}(X, Y),\end{aligned} \tag{B1}$$

$$\text{SD}^2(X) = \text{VAR}(X). \tag{B2}$$

Applied to $T(\kappa) = T_1 + \kappa T_2$, we obtain

$$\mathrm{SD}(\kappa) \equiv \mathrm{SD}(T_1 + \kappa T_2) = \sqrt{\mathrm{SD}_1^2 + \kappa^2 \mathrm{SD}_2^2 + 2\kappa \mathrm{COV}(T_1, T_2)}. \quad \text{(B3)}$$

Note that $\mathrm{SD}(\kappa)$ and SD_i refer to the standard deviation of the total test and of the test component i, respectively.

Expanding Eq. (11) yields

$$d(\kappa) = \frac{T_{\mathrm{post}}(\kappa) - T_{\mathrm{pre}}(\kappa)}{\mathrm{SD}_{\mathrm{pre}}(\kappa)} \quad \text{(B4)}$$

$$= \frac{T_{\mathrm{post},1} + \kappa T_{\mathrm{post},2} - T_{\mathrm{pre},1} - \kappa T_{\mathrm{pre},2}}{\mathrm{SD}_{\mathrm{pre}}(\kappa)} \quad \text{(B5)}$$

$$= \frac{T_{\mathrm{post},1} - T_{\mathrm{pre},1}}{\mathrm{SD}_{\mathrm{pre}}(\kappa)} + \kappa \frac{T_{\mathrm{post},2} - T_{\mathrm{pre},2}}{\mathrm{SD}_{\mathrm{pre}}(\kappa)} \quad \text{(B6)}$$

$$= \frac{T_{\mathrm{post},1} - T_{\mathrm{pre},1}}{\mathrm{SD}_{\mathrm{pre},1}} \frac{\mathrm{SD}_{\mathrm{pre},1}}{\mathrm{SD}_{\mathrm{pre}}(\kappa)} + \kappa \frac{T_{\mathrm{post},2} - T_{\mathrm{pre},2}}{\mathrm{SD}_{\mathrm{pre},2}} \frac{\mathrm{SD}_{\mathrm{pre},2}}{\mathrm{SD}_{\mathrm{pre}}(\kappa)} \quad \text{(B7)}$$

Using the definition of Cohen's d in the ith test component

$$d_i = \frac{T_{\mathrm{post},i} - T_{\mathrm{pre},i}}{\mathrm{SD}_{\mathrm{pre},i}}, \quad \text{(B8)}$$

Eq. (B7) reads

$$d(\kappa) = d_1 \frac{\mathrm{SD}_{\mathrm{pre},1}}{\mathrm{SD}_{\mathrm{pre}}(\kappa)} + \kappa d_2 \frac{\mathrm{SD}_{\mathrm{pre},2}}{\mathrm{SD}_{\mathrm{pre}}(\kappa)}. \quad \text{(B9)}$$

Finally, we obtain Eq. (12) by inserting Eq. (B3) into Eq. (B9).

[1] M. de Cock, Representation use and strategy choice in physics problem solving, Phys. Rev. ST Phys. Educ. Res. **8**, 020117 (2012).

[2] S. Ainsworth, The functions of multiple representations, Comput. Educ. **33**, 131 (1999).

[3] A. Van Heuvelen, Learning to think like a physicist: A review of research-based instructional strategies, Am. J. Phys. **59**, 891 (1991).

[4] P. Nieminen, A. Savinainen, and J. Viiri, Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning, Phys. Rev. ST Phys. Educ. Res. **8**, 010123 (2012).

[5] D. E. Meltzer, Relation between students' problem-solving performance and representational format, Am. J. Phys. **73**, 463 (2005).

[6] P. B. Kohl and N. D. Finkelstein, Patterns of multiple representation use by experts and novices during physics problem solving, Phys. Rev. ST Phys. Educ. Res. **4**, 010111 (2008).

[7] L. C. McDermott, M. Rosenquist, and E. van Zee, Student difficulties in connecting graphs and physics: Examples from kinematics, Am. J. Phys. **55**, 503 (1987).

[8] L. C. McDermott, Research on conceptual understanding in mechanics, Phys. Today **37**, 24 (1984).

[9] R. Duit, H. Niedderer, and H. Schecker, Teaching physics, in *Handbook of Research on Science Education*, edited by N. Lederman and S. Abell (Routledge/Taylor & Francis, London, New York, 2014).

[10] K. K. Mashood and V. A. Singh, An inventory on rotational kinematics of a particle: Unravelling misconceptions and pitfalls in reasoning, Eur. J. Phys. **33**, 1301 (2012).

[11] R. J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

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

[13] R. K. Thornton and D. 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).

[14] P. Nieminen, A. Savinainen, and J. Viiri, Force Concept Inventory-based multiple-choice test for investigating students' representational consistency, Phys. Rev. ST Phys. Educ. Res. **6**, 020109 (2010).

[15] R. S. Lindell, E. Peak, and T. M. Foster, Are they all created equal? A comparison of different concept Inventory development methodologies, AIP Conf. Proc. **883**, 14 (2006).

[16] J. S. Aslanides and C. M. Savage, Relativity concept inventory: Development, analysis, and results, Phys. Rev. ST Phys. Educ. Res. **9**, 010118 (2013).

[17] P. B. Kohl and N. D. Finkelstein, Student representational competence and self-assessment when solving physics problems, Phys. Rev. ST Phys. Educ. Res. **1**, 010104 (2005).

[18] J. Stewart, H. Griffin, and G. Stewart, Context sensitivity in the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **3**, 010102 (2007).

[19] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.13.010132 for the English and the German version of the KiRC inventory.

[20] J. L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. **76**, 378 (1971).

[21] J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, Biometrics **33**, 159 (1977).

[22] J. A. Marshall and D. J. Carrejo, Students' mathematical modeling of motion, J Res. Sci. Teach. **45**, 153 (2008).

[23] J. Bowden, G. Dall'Alba, E. Martin, D. Laurillard, F. Marton, G. Maters, P. Ramsden, A. Stephanou, and E. Walsh, Displacement, velocity, and frames of reference: Phenomenographic studies of students' understanding and some implications for teaching and assessment, Am. J. Phys. **60**, 262 (1992).

[24] E. C. Sayre and M. C. Wittmann, Plasticity of intermediate mechanics students' coordinate system choice, Phys. Rev. ST Phys. Educ. Res. **4**, 020105 (2008).

[25] P. Klein, S. Gröber, J. Kuhn, A. Fleischhauer, and A. Müller, The right frame of reference makes it simple: an example of introductory mechanics supported by video analysis of motion, Eur. J. Phys. **36**, 015004 (2015).

[26] D. A. Frisbie and D. C. Sweeney, The relative merits of multiple true-false tests, J. Educ. Measure. **19**, 29 (1982).

[27] A. Mobalegh and H. Barati, Multiple True-false (MTF) and Multiple-choice (MC) test formats, J. Lang. Teach. Res. **3**, 1027 (2012).

[28] M. C. Rodriguez, Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research, Educ. Meas. Issues Practice **24**, 3 (2005).

[29] D. Treagust, Evaluating students' misconceptions by means of diagnostic multiple choice items, Res. Sci. Educ. **16**, 199 (1986).

[30] P. Klein, J. Kuhn, A. Müller, and S. Gröber, Video analysis exercises in regular introductory mechanics physics courses: Effects of conventional methods and possibilities of mobile devices, in *Multidisciplinary Research on Teaching and Learning*, edited by W. Schnotz, A. Kauertz, H. Ludwig, A. Müller, and J. Pretsch (Palgrave Macmillan, Basingstoke, UK, 2015).

[31] S. Gröber, P. Klein, and J.Kuhn, Video-based problems in introductory mechanics physics courses, Eur. J. Phys. **35**, 055019 (2014).

[32] R. Dilber, I. Karaman, and B. Duzgun, High school students' understanding of projectile motion concepts, Educ. Res. Eval. **15**, 203 (2009).

[33] L. J. Cronbach and P. E. Meehl, Construct validity in psychological tests, Psychol. Bull. **52**, 281 (1955).

[34] 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).

[35] C. Marshall and G. Rossman, *Designing qualitative research* (Sage Publications, Thousand Oaks, CA, 1999).

[36] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Phys. Rev. ST Phys. Educ. Res. **5**, 020103 (2009).

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

[38] 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).

[39] J. Cohen, Statistical power analysis, Curr. Dir. Psychol. Sci. **1**, 98 (1992).

[40] In particular, one has to substitute the item difficulties with the sum of correct responses and the "1" by the total number of questions.

[41] R. B. Frary, Formula Scoring of multiple-choice tests (correction for guessing), Educ. Meas. **7**, 33 (1988).

[42] J. Stewart and G. Stewart, Correcting the normalized gain for guessing, Phys. Teach. **48**, 194 (2010).

[43] In particular, one obtains

$$d = \frac{T_{\text{post}} - T_{\text{pre}}}{\text{SD}_{\text{pre}}} = \frac{\alpha Z_{\text{post}} + \gamma - (\alpha Z_{\text{pre}} + \gamma)}{\text{SD}_{\text{pre}}} = \frac{\alpha(Z_{\text{post}} - Z_{\text{pre}})}{\text{SD}_{\text{pre}}},$$

where $Z = Z(T_1, T_2, \alpha, \beta, \gamma)$ replaces the term in brackets in Eq. (6) for simplicity. Using the rule $\text{VAR}(\alpha Z + \gamma) = \alpha^2 \text{VAR}(Z)$ and $\text{SD}^2(Z) = \text{VAR}(Z)$, it follows that the prefactor α cancels out which proofs our claim. Of course, α is still present in Z.

[44] B. A. Lindsey and M. L. Nagel, Do students know what they know? Exploring the accuracy of students' self-assessments, Phys. Rev. ST Phys. Educ. Res. **11**, 020103 (2015).

[45] J. E. Dowd, I. Araujo, and E. Mazur, Making sense of confusion: Relating performance, confidence, and self-efficacy to expressions of confusion in an introductory physics class, Phys. Rev. ST Phys. Educ. Res. **11**, 010107 (2015).

[46] J. J. Shaughnessy, Confidence-judgment accuracy as a predictor of test performance, Journal of research in personality **13**, 505 (1979).

[47] P. Bell and D. Volckmann, Knowledge surveys in General Chemistry: Confidence, overconfidence, and performance, J. Chem. Educ. **88**, 1469 (2011).

[48] M. D. Sharma and J. Bewes, Self-monitoring: Confidence, academic achievement and gender differences in physics, J. Learn. Des. **4** (2011).

[49] J. Kruger and D. Dunning, Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments, J. Pers. Soc. Psychol. **77**, 1121 (1999).

[50] We preferred a $C = 1$ vs $C \neq 1$ split over a $C > 0.5$ vs $C < 0.5$ split due to the high average confidence score of 0.75.

[51] D. M. Lane, Online Statistics Education: An Interactive Multimedia Course of Study, Chapter 10: Estimation—Correlation, http://onlinestatbook.com/ (accessed January, 17th. 2017).

[52] X. Fan and B. Thompson, Confidence intervals about score reliability coefficients, in *Score Reliability—Contemporary Thinking on Reliability Issues*, edited by B. Thompson (SAGE, Thousand Oaks, CA, London, New Dehli, 2003).