Validation and structural analysis of the kinematics concept test

A. Lichtenberger,[1,*] C. Wagner,[1,†] S. I. Hofer,[2] E. Stern,[2] and A. Vaterlaus[1]
[1]Laboratory for Solid State Physics, ETH Zurich, 8093 Zurich, Switzerland
[2]Institute of Behavioral Sciences, ETH Zurich, 8092 Zurich, Switzerland
(Received 20 September 2016; published 10 April 2017)

The kinematics concept test (KCT) is a multiple-choice test designed to evaluate students' conceptual understanding of kinematics at the high school level. The test comprises 49 multiple-choice items about velocity and acceleration, which are based on seven kinematic concepts and which make use of three different representations. In the first part of this article we describe the development and the validation process of the KCT. We applied the KCT to 338 Swiss high school students who attended traditional teaching in kinematics. We analyzed the response data to provide the psychometric properties of the test. In the second part we present the results of a structural analysis of the test. An exploratory factor analysis of 664 student answers finally uncovered the seven kinematics concepts as factors. However, the analysis revealed a hierarchical structure of concepts. At the higher level, mathematical concepts group together, and then split up into physics concepts at the lower level. Furthermore, students who seem to understand a concept in one representation have difficulties transferring the concept to similar problems in another representation. Both results have implications for teaching kinematics. First, teaching mathematical concepts beforehand might be beneficial for learning kinematics. Second, instructions have to be designed to teach students the change between different representations.

DOI: 10.1103/PhysRevPhysEducRes.13.010115

I. INTRODUCTION

To design effective instructional interventions, it is necessary to adequately assess students' conceptual knowledge. The major way to evaluate conceptual knowledge of students in physics education research is by means of a multiple-choice test. To cover kinematics and mechanics, different tests have already been developed, e.g., the Force Concept Inventory [1], the Mechanics Baseline Test [2], the Force and Motion Concept Evaluation [3], and the Test of Understanding Graphs in Kinematics [4]. Although most of the concept tests in mechanics contain items about kinematics, a concept test which systematically explores the basic kinematics concepts in different representations is still missing.

To close this gap we developed the kinematics concept test (KCT), which we present in this article. The KCT is designed for use in research to evaluate students' understanding of seven basic kinematics concepts in three representations. The purpose of the test raises two research questions. (i) Is the KCT indeed a reliable and valid instrument to evaluate kinematics concept knowledge? (ii) Is our model of seven kinematics concepts reflected in the data of student responses?

*lichtenberger@phys.ethz.ch
†wagnercl@phys.ethz.ch

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

This article is mainly divided into two parts according to the research questions. The first part documents the detailed development and validation process of the KCT. It includes the analysis of test data from 338 Swiss Gymnasium students using classical test theory. The Gymnasium is the secondary school in Switzerland that is attended by 20% of high performing students. A final diploma of the Gymnasium allows access to the university. The second part of the paper deals with the structural analysis of students' responses. We applied an exploratory factor analysis to a data set of 664 student answers. The results from both evaluations are discussed thereafter and recommendations to improve teaching in kinematics are presented.

II. BACKGROUND AND MODEL OF THE KCT

Our model to develop the KCT is based on seven concepts in kinematics, which are already described in the literature [4–6]. Thompson [7] specified the concept of velocity as rate. Trowbridge and McDermott [8] investigated velocity as a one-dimensional construct, and Aguirre and Erickson [9] referred to the vector character of velocity in two dimensions. A further concept is the displacement vector as area under the curve in the vt diagram. McDermott, Rosenquist, and van Zee [10] and Nguyen and Rebello [11] used the concept of area under the curve to explore students' understanding and application of the concept in different contexts including kinematics. Acceleration is derived from velocity in a similar fashion as velocity from position. Consequently, four additional

TABLE I. The relation of the KCT between concepts and representations (denoted by "X") and the numbers of the items we used and adapted from the FCI and the TUG-K. Three items of the TUG-K (11, 14, 15) can be assigned to two concepts each.

	Representations		
Concepts	Pictures	Tables	Graphs
C1: Velocity as rate	X FCI 19	X	X TUG-K 5, (11)
C2: Velocity as one-dimensional vector	X		X (TUG-K 11)
C3: Addition of velocities in two dimensions	X FCI 9		
C4: Displacement as area under the vt curve			X TUG-K 4
C5: Acceleration as rate	X	X	X TUG-K 2, 6, 7, (14, 15)
C6: Acceleration as one-dimensional vector	X		X (TUG-K 14, 15)
C7: Velocity change as area under the at curve			X TUG-K 1, 10, 16

concepts for acceleration can be defined: acceleration as rate, acceleration as a one- or two-dimensional vector, and area under the curve of the at graph as change of velocity [12]. Because the acceleration in two dimensions is not discussed in basic physics courses at Swiss Gymnasiums, we are left with three additional concepts for acceleration. The numbering of concepts is presented in Table I. While the concepts are used to construct multiple-choice items, the preconceptions and misconceptions described in the literature [4,10,13,14] are used to form distractors for the multiple-choice items.

An important contribution to analyze the graph representation of the discussed concepts has been provided by Beichner [4], who developed the Test of Understanding Graphs in Kinematics (TUG-K). His results revealed that students at the high school and the college level have great difficulties interpreting graphs in kinematics correctly. Still, there is more potential for researching the role of representations as a reasoning tool. For instance, Ainsworth [15] suggests that using multiple representations might be beneficial for the learning of complex ideas. Therefore, in addition to graphs, we used two other representations in the concept questions of the KCT: tables and pictures. These three representations seem to be the most common ones for physics concept questions out of the eleven types defined by Lohse, Biolsi, Walker, and Rueter [16].

For example, Fig. 1 depicts the three different representations for velocity as rate. In order to determine the instantaneous velocity in a graph students use the slope of the tangent. In a table or a stroboscopic picture students have to determine the velocity via the distance covered during a time interval divided by this time interval. For velocity as a one-dimensional vector students have to relate the velocity to the introduced coordinate system. Moving forward means the object moves in the positive direction of the coordinate axis and backward in the negative direction. In the case of velocity as a two-dimensional vector, students only have to be familiar with the graphic addition of

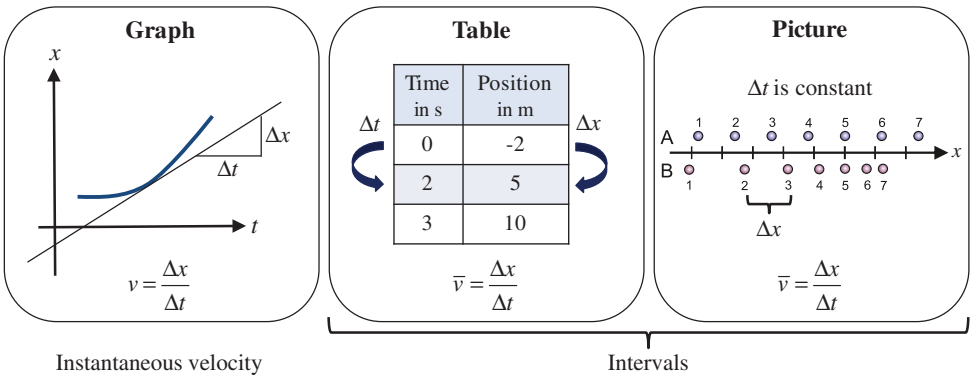

FIG. 1. Velocity as rate. In the case of a graph students must be able to determine the instantaneous velocity using the slope of the tangent. Using a table, students have to read out the difference in time and position in order to determine the velocity during the corresponding time interval. In a stroboscopic picture the time interval is constant and the velocity of the object is proportional to the distance between two subsequent positions.

two-dimensional vectors. Finally, the area under a curve can only be used in combination with a graph.

Table I summarizes the concepts and representations which build the model of the KCT. The concepts and representations already addressed by the FCI and by the TUG-K are also indicated. The major reason for the design of the kinematics concept test was to explore the complete set of concepts and representations shown in Table I.

III. DEVELOPMENT AND VALIDATION OF THE KCT

In this section we address the first research question and show in detail the development of the test and the process of validation. The psychometric data obtained from the administration of the test to 338 high school students can be used as a reference for further applications of the KCT.

A. Methods

The test development is based on the flowchart suggested by Beichner [4]. It includes the formulation of learning objectives, the construction of a test draft, and the performance of reliability and validity checks, which then result in an adaption of the items (feedback loop). The different development and validation cycles are presented in Table II. Overall, we collected and analyzed data from 745 Swiss Gymnasium students. To measure the reliability of our test drafts we calculated the internal consistency of the items using the Kuder-Richardson Formula 20 (KR-20) for dichotomous items [17]. To provide evidence of validity for using the KCT in research we collected and analyzed ratings from experts. As validity is a joint responsibility between the developers and users of a test [18], we also gathered feedback from students in written comments and in short interviews. Furthermore, we examined the psychometric properties of the test. For each item we determined the difficulty, the point-biserial correlation coefficient [4], and the discrimination index. The latter was determined by calculating the difference of the mean scores between two equal-sized subgroups of the sample, one built from the 27% highest scorers, the other from the 27% lowest scorers [19].

To create a first draft of the KCT we constructed a set of items based on the kinematics items 1–2, 4–7, 10–11, and 14–16 of the TUG-K [4] and the items 9 and 19 of the FCI [20]. We then developed new items according to the model given in Table I. We carefully designed each item in a way that it can be associated with exactly one concept and one representation. Eighteen Gymnasium students between 14 and 18 years solved the first set of items (version 1.0) in an open-ended questions format. We analyzed the answers qualitatively in order to work out common misconceptions. Furthermore, we conducted short interviews with the students. The students had to explain their thoughts regarding wrong answers. Based on the interviews and the written answers we formulated a set of distractors for each item. Moreover, we used well-known misconceptions from the literature [4,10,13,14] to build additional distractors. This process resulted in a first multiple-choice version of the KCT with up to seven distractors per item (version 2.0–2.1).

After the analysis of 149 student answers we reduced the set of distractors per item to the three or four most effective ones. Moreover, we added parallel questions (analog items with different context) to most of the test items. We ended up with version 3 of the multiple-choice test which was administered to 110 Gymnasium students. An open text field for comments followed each item. Most students used this option and we received informative feedback. After the application of version 3.0 and 3.1, we slightly changed some of the items. With the data from the last group (test

TABLE II. Development stages of the kinematics concept test. The format MC stands for multiple-choice items. The review methods at each stage are mentioned in the right column. Item analysis includes the calculation of difficulties, item-total correlations, and discrimination indices.

Version	Format	No. Items	No. Students	Review methods
1.0	Open-ended questions	29	18	Qualitative analysis of responses Short interviews with students
2.0–2.1	MC with comments (3–7 distractors)	34	149	Reliability and item analysis Evaluation of student comments Evaluation of distractors
3.0–3.2	MC with comments (3–4 distractors)	56	175	Reliability and item analysis Exploratory factor analysis Evaluation of student comments
4.0–4.2	MC with comments (3–4 distractors, calculations open)	61	65	Reliability and item analysis Evaluation of distractors Evaluation of student comments Evaluation of expert feedback ($N = 3$)
5.0	MC (3–5 distractors)	49	338	Reliability and item analysis Exploratory factor analysis Evaluation of expert feedback ($N = 6$)

TABLE III. Categorization of the 49 items. Each item (except the four in parentheses) can be assigned to a single concept and a single representation.

	Representation			
Concept	Picture	Table	Graph	No. Items
C1	1 25	3 27	10 11 16 20 26 33 42 (7) (37)	11 + (2)
C2	5 41	⋯	12 17 21 31 43 (7) (37)	7 + (2)
C3	4 8 13 18 45	⋯	⋯	5
C4	⋯	⋯	30 35 38 49	4
C5	14 29	15 47	2 6 28 32 39 (19) (23)	9 + (2)
C6	9 36	⋯	24 34 48 (19) (23)	5 + (2)
C7	⋯	⋯	22 40 44 46	4
No. Items	13	4	28 + (4)	45 + (4)

version 3.2, 56 items, 65 students) we determined the psychometric properties of the test. Moreover, we conducted an exploratory factor analysis to check if the intended test structure based in the seven concepts can be found in the students' scoring patterns [21]. On the basis of the results from these analyses we adapted the test again.

Three physics educators reviewed the fourth test version and again 65 students solved it. The final step was to reduce the test to a length that can be processed by at least 95% of the students within a standard lesson of 45 min. We skipped items with too high difficulty indices and deleted those that revealed high correlations with other items in order to prevent losing important information. We also balanced the distribution of the items across concepts and representations. The resulting test version 5.0 consists of 49 items with one correct solution and three to five distractors. The random guessing score of the test is 9.6.

Table III shows the categorization of the items. Forty-five of the items are assigned to exactly one concept and representation. Only four items (7, 19, 23, 37) are assigned to two concepts each. In these items two graphs with different y axes (position, velocity or acceleration) have to be matched. We considered these items too important to be omitted even if they did not meet the criterion of referring to only one concept.

In the last step of our evaluation process six physicists who teach the subject at the Gymnasium or at the university level reviewed the test version 5.0. They solved the test, matched items and concepts, answered several questions to rate the test items and the whole test (e.g. "How appropriate is the item to measure the assigned concept?", "How well does the test cover the kinematics concepts?", "Are the seven concepts represented in a balanced way?") and criticized the distractors. According to the feedback of the physics educators the current version of the KCT is an adequate instrument to test students' overall conceptual knowledge in basic kinematics but also a students' knowledge at the single concept level.

As reevaluating the appropriateness of an instrument's use is an ongoing process [18], we again collected data from 338 Swiss Gymnasium students with the current test version 5.0. We present the psychometric properties of the test based on this last round of data collection in detail in the next sections.

In all applications we used German versions of the kinematics concept test. The non-validated English translation of the current test version 5.0 is provided in the Supplemental Material [22].

B. Data collection

In the whole development and validation process we collected data from 745 Swiss Gymnasiums students in the German speaking part of Switzerland. Remember, the Gymnasium is a public school that provides higher secondary education to above-average achieving students. It constitutes the highest track of the educational system. Gymnasium students are comparable to U.S. high school students attending college preparatory classes.

We distributed the final test version to 343 Gymnasium students (157 boys, 186 girls) from 17 physics courses with ten different teachers at ten Gymnasiums. The mean age of the students was $M = 15.4$ yr (SD = 1.1 yr, range: 13–18 yr). All students took the KCT as a post-test after traditional instruction in kinematics. To control for retesting effects 271 of the 343 students also solved the KCT as a pretest before instruction. The psychometric data presented in the following section only take the post-test data into account.

Prior to the analysis we checked the data set for inconsistencies. As the test was solved on computers we were able to collect the student response times for each item. We evaluated the response times to determine whether or not a student had made a genuine attempt to answer the items. We defined two exclusion criteria based on the response time statistics. First, the time to solve the whole test had to be greater than 830 sec. We derived this criterion from the scatter plot of the KCT scores versus the total response times (see Fig. 2). As the plot shows, only for response times more than 830 sec, test scores higher than the student mean value of 27.8 were reached. Moreover, this time is close to the 2σ deviation from the mean, which

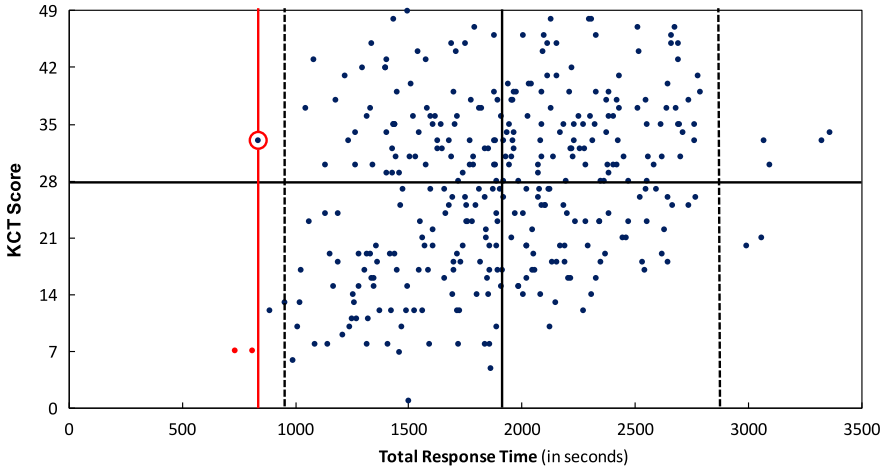

FIG. 2. Scatter plot of the KCT score versus the time needed to complete the test. The solid black lines indicate the mean values, the dashed lines show the 2σ interval (2 standard deviations) for the response time. A time greater than 830 sec was necessary to score higher than the mean (see data point marked with a circle). This time was determined as the minimal time needed to take the test seriously. Data points with a response time shorter than 830 sec (marked red) were removed from the sample.

also indicates that it is a reasonable threshold. The average response time was 1919 seconds (SD = 482 sec). Two students (0.6%) did not meet this criterion.

In addition, we looked at the number of items where the response time was less than 5 or less than 10 seconds. We found that an above-average KCT score was only reached when less than 15 or 27 items were answered in less than 5 or 10 sec, respectively. We used this information to set these numbers as threshold values. In other words, students who spent less than 5 or 10 sec on more than 14 or 26 items were removed from the sample. The mean values of the number of items with response times of less than 5 and 10 sec were 0.5 (SD = 2.4) and 1.6 (SD = 3.7), respectively. Expert ratings actually also suggested that the time to answer an item seriously was at least 10 sec. We set the threshold values rather high not to exclude data unjustifiably. Three more students were removed from the sample due to the second criterion. We ended up with post-test data sets from 338 students and pretest data sets from 266 students, excluding a total of five students (1.5%; three boys, two girls).

C. Test analysis

The psychometric properties of the KCT gained from the answers of 338 Swiss Gymnasium students are summarized in Table IV (mean values) and Table V (individual item values).

The average number of correctly solved items is 27.8. This score is close to the desired value, which is halfway between the maximum and the random guessing score, as suggested by Doran [23]. Almost all items (five exceptions) are in the difficulty range between 0.30 and 0.90, which is reasonable [23,24]. Considering the distribution of item difficulties, we find 14% of the items in the difficult range (0.00–0.35), 37% in the moderately difficult range (0.35–0.60), 47% in the moderately easy range (0.60–0.85), and 2% in the easy range (0.85–1.00). Compared to the suggestions by Doran, there is a small underrepresentation of items in the easy range and a corresponding overrepresentation of items in the moderate range. Thus, the present test especially discriminates the range from very high to moderately low performance.

We found the reliability index for the KCT to be 0.92, which is sufficiently high for both group and individual measurements. We also looked at the contributions of the single items to the scale. An analysis of the "KR-20, if item deleted" values showed that no item lowered the reliability of the test instrument.

The quality of single test items is ascertained by calculating the point-biserial correlation coefficients and discrimination indices [4]. Items are generally considered to be reliable if the point-biserial correlation coefficient is ≥ 0.20. This requirement is met by all items. The discrimination index measures how well an item differentiates between

TABLE IV. Overall test results from the final 49-item-version of the test, taken from a sample of 338 post-instruction Gymnasium students. The score is the number of correctly solved items.

Mean score	Standard error	Mean difficulty	Reliability (KR-20)	Mean point-biserial coefficient	Mean item discrimination index
27.8	0.6	0.57	0.92	0.46	0.53

TABLE V. Statistics for the individual items, taken from a sample of 338 postinstruction Gymnasium students. Values marked with a star are not within the desired range. Values in bold typeface indicate the correct answer.

Item	Difficulty	Point-biserial coefficient	Item discrimination index	Choice ($N = 338$)					
				A	B	C	D	E	F
Desired value	0.3–0.9	≥0.20	≥0.30	...	...	...	...	...	...
1	0.70	0.59	0.70	20	12	7	63	**236**	...
2	0.86	0.26	0.20*	27	0	**291**	7	13	...
3	0.70	0.55	0.69	28	65	3	6	**236**	...
4	0.69	0.31	0.33	49	**234**	13	29	13	...
5	0.80	0.34	0.35	2	3	49	**269**	15	...
6	0.75	0.34	0.37	19	1	**255**	50	13	...
7	0.58	0.45	0.56	54	40	**196**	22	26	...
8	0.64	0.39	0.42	33	28	**218**	52	7	...
9	0.42	0.40	0.49	71	**141**	22	99	5	...
10	0.38	0.51	0.56	49	**127**	26	3	133	...
11	0.60	0.59	0.67	11	31	90	2	**204**	...
12	0.78	0.44	0.44	8	10	58	**262**	0	...
13	0.57	0.38	0.45	5	41	**192**	27	73	...
14	0.72	0.62	0.71	20	24	47	5	**242**	...
15	0.79	0.55	0.57	24	1	4	43	**266**	...
16	0.38	0.52	0.63	4	**127**	130	8	65	4
17	0.76	0.41	0.43	11	**256**	70	1	...	...
18	0.63	0.43	0.47	26	37	**212**	47	16	...
19	0.44	0.42	0.52	87	21	**149**	10	71	...
20	0.50	0.56	0.70	5	24	80	**170**	57	2
21	0.77	0.43	0.42	17	19	7	**260**	35	...
22	0.29*	0.49	0.55	10	4	**97**	201	26	...
23	0.57	0.46	0.53	**194**	72	20	25	27	...
24	0.75	0.33	0.33	19	25	5	**254**	35	...
25	0.68	0.60	0.74	26	69	**230**	10	3	...
26	0.39	0.56	0.68	**132**	4	74	35	93	...
27	0.70	0.59	0.69	17	9	70	**237**	5	...
28	0.45	0.49	0.63	46	69	**153**	48	15	7
29	0.71	0.49	0.65	40	**239**	16	29	14	...
30	0.48	0.43	0.52	**162**	14	68	66	28	...
31	0.78	0.46	0.47	3	7	63	**265**	...	...
32	0.57	0.51	0.64	29	**191**	71	10	30	7
33	0.32	0.55	0.63	106	**108**	1	22	101	...
34	0.71	0.24	0.26*	75	5	17	**239**	2	...
35	0.34	0.37	0.45	72	7	66	60	**115**	18
36	0.41	0.48	0.62	12	102	**138**	17	69	...
37	0.54	0.45	0.53	40	33	20	**184**	61	...
38	0.28*	0.41	0.47	14	6	66	90	**93**	69
39	0.59	0.49	0.62	36	**201**	11	74	16	...
40	0.13*	0.43	0.35	24	130	24	97	**43**	20
41	0.63	0.43	0.53	3	61	44	**212**	18	...
42	0.44	0.52	0.67	7	**150**	12	163	6	...
43	0.74	0.45	0.46	26	10	21	32	**249**	...
44	0.20*	0.50	0.49	8	118	67	76	**69**	...
45	0.41	0.46	0.54	51	22	**138**	69	58	...
46	0.23*	0.42	0.43	62	52	32	103	**77**	12
47	0.76	0.47	0.51	20	7	11	**258**	42	...
48	0.69	0.42	0.47	33	16	28	**234**	27	...
49	0.56	0.47	0.59	52	**189**	20	48	29	...

high and low scoring students. With exception of items 2 and 34, the figures of all items exceed the required threshold of 0.30.

Finally, Table V presents the distribution of student answers for all items. Most of the distractors were selected frequently. However, there are good reasons to keep even those distractors, which are hardly selected. The main reason is to minimize the random guessing score by providing generally four distractors. Another reason is to provide a symmetric choice of answers (e.g., choice B at item 2 or choice E at item 12), which might appear like a natural choice without hints to the solution. Our pilot studies have shown that skipping an unused distractor can change the distribution of the answers.

As we show in the following section about the KCT structure, there is evidence that the KCT not only allows to measure the overall kinematics knowledge but also to evaluate the single concept knowledge. For completeness, Table VI displays the reliability and difficulty coefficients for the associations of items according to the single concepts. The last column shows that all concept difficulties are highly correlated to the difficulty of the whole test, $r > 0.6$, $p < 0.001$ (for all concepts).

Table VI also reveals that students have great difficulties to solve items related to concepts C4 and C7 (determining area under the curve). The implications for instruction will be discussed in the last section.

Finally, we also investigated the pretest results of 266 students. If the KCT is solved as a pretest the mean score is 16.0, which is considerably lower than in the post-test. Nevertheless, we found a significant correlation between pre- and post-test results, $r = 0.67$ ($p < 0.001$). The pretest reliability index is 0.86. We found no difference in the post-test results between the 266 students who solved the test as a pretest and the 72 students who solved the KCT for the first time. Thus, there seems to be no test-retest effect. This was expected due to the high number of items, the time between pre- and post-test (in average four weeks) and most of all because students were not given feedback about their pretest results.

D. Description of the exceptions

Items 2 and 34 do not meet the general requirements for the discrimination index. The low item discrimination of item 2 can be explained by its high solution rate. As most of the students solve the item correctly it does not discriminate much between the upper and lower group. Nevertheless, item 2 is considered to be important for the test as it is the only item, which falls into the "easy" category according to the definition by Doran [23]. It is therefore the only item which allows differentiating among very low scoring students. There is no straightforward interpretation of the rather low reliability and discrimination index of item 34. The frequent selection of distractor A shows that there is the widespread misconception of a directionless acceleration in graphs that can also be found in the picture representation of acceleration (see item 9, choice D and item 36, choice E).

E. Conclusions

The analysis of 338 student answers shows that the KCT has the desired psychometric properties. The balanced distribution of items over concepts, the high reliability index and the expert feedback provide evidence of validity for the use of the KCT as an instrument in physics education research to evaluate students' concept knowledge in basic kinematics. Moreover, all concept scores are highly correlated with the sum score of the KCT, indicating that they all significantly contribute to the sum. Therefore, determining the KCT score from the sum of averaged concept scores would not substantially change the students' results.

Finally, we also recommend using the KCT to evaluate the knowledge of the seven concepts separately. Evidence of validity for this use is given by expert feedback, reasonable reliability values and, furthermore, by the structural analysis of the KCT presented in the following section.

TABLE VI. Reliability (KR-20) values and difficulties for the items grouped according to their associated concepts, calculated from a data set of 338 Gymnasium students. The last column shows the correlation of the concept scores to the total test scores. The four items (7, 19, 23, 37) which can be associated to two concepts are only considered in the total score (as they were also skipped in the factor analysis). Adding these items to the corresponding concepts would slightly increase the KR-20 and, obviously, the correlation values. The difficulty values would remain at the same levels (±0.03).

		KR-20	Difficulty	Correlation to KCT total score
All items	(49 items)	0.92	0.57	1
Concept C1	(11 items)	0.84	0.53	$r = 0.90$ ($p < 0.001$)
Concept C2	(7 items)	0.75	0.75	$r = 0.67$ ($p < 0.001$)
Concept C3	(5 items)	0.63	0.59	$r = 0.62$ ($p < 0.001$)
Concept C4	(4 items)	0.56	0.41	$r = 0.64$ ($p < 0.001$)
Concept C5	(9 items)	0.75	0.69	$r = 0.82$ ($p < 0.001$)
Concept C6	(5 items)	0.55	0.60	$r = 0.63$ ($p < 0.001$)
Concept C7	(4 items)	0.65	0.21	$r = 0.64$ ($p < 0.001$)

IV. STRUCTURAL ANALYSIS OF THE KCT

An important aspect of teaching physics is understanding the difference between the ways novices and experts think about the topic [25]. We established seven concepts that were considered to be essential for kinematics. The goal of the KCT is to assess these concepts. Of course, these concepts are constructs, which cannot be measured directly. Thus, we try to measure different aspects of the concepts by using the KCT items. From an expert point of view our test covers the seven concepts and three representations. Moreover, each item (except items 7, 19, 23, and 37) can be clearly assigned to one concept and one representation. Nevertheless, it is not clear if this relation between items and concepts is also valid from a student perspective. In order to explore this issue we performed an exploratory factor analysis (EFA) on test data of 664 Swiss Gymnasium students. Factor analysis provides evidence for the existence of latent variables and allows one to uncover the structure behind the test. In this context, the two important questions are, how many factors are needed to explain most of the variation in the data set? Can we recover the intended model of concepts and/or representations in the structure of the data? Thus, factor analysis can be seen as another source (novice perspective) of evidence in the validation process of the KCT with regard to its intended use as an evaluation tool.

A. Methods

EFA is a standard technique in the construction of questionnaires [26–28]. First we calculated the tetrachoric correlations with the software R [29] by applying the psych package [30]. To determine the number of factors we used Velicer's [31] minimum average partial (MAP) test, Horn's [32] parallel analysis, Cattell's [33] scree test and the related nongraphical optimal coordinate (OC) method [34]. The calculations were done in R using the package nFactors [35]. These calculations were completed by considerations based on the theoretical models in order to obtain the optimal number of factors. The EFA was conducted with the principal axis factoring (PAF) method in the SPSS software [36] and rechecked with other common methods, i.e., the maximum likelihood (ML) procedure in SPSS and the ML-based EFA in M*plus*[37]. Because of the high internal consistency of the test and the specific topic of kinematics, we allowed factors to be correlated. Thus, we applied oblique rotations with the promax method, as generally recommended by Schmitt [28].

To analyze the data we first separated the graph items from the picture and table items in order to compare the results with findings described in the literature. Bektasli and White [38] investigated the TUG-K using factor analysis. They applied principal component analysis and varimax rotation to the data collected from 72 students (12^{th} grade). Their factor analysis revealed two factors, one for determining the slope of a curve and the other one for finding or interpreting the area under a curve. Accordingly, they found two mathematical concepts rather than two physics concepts like velocity and acceleration. In what follows, we show that our results are in line with these findings.

B. Data collection

To analyze the test structure we applied the KCT to 688 Gymnasium students (337 boys, 351 girls) from 37 physics courses with 29 different teachers at 24 Gymnasiums in the German speaking part of Switzerland. The mean age of the students was $M = 15.5$ yr (SD $= 1.0$ yr, range: 13–18 yr). By applying the previously introduced exclusion criteria based on response times, we ended up with a data set of 664 students (320 boys, 344 girls), excluding a total of 24 students (3.5%; 17 boys, 7 girls).

The students took part in a comparison study between three different instructional methods: formative assessment and traditional teaching with and without additional concept tests. As the KCT structure should not depend on instruction, we carried out the EFA with the postinstructional KCT data of all three groups. Taking only data from the traditional teaching into account, the data set would have been too small for an adequate EFA including all 45 items. As a rule of thumb, the inclusion of 45 items requires data from 450 students. Nevertheless, we also carried out the EFA for the three teaching groups separately, by splitting the items according to representations. Indeed, we found no difference in the answer structure between the three groups. In what follows, we report the results of the EFA with a sample including all three groups.

C. Results

For the graph items, the scree test with the OC method suggested a three-factor solution whereas a five-factor solution was appropriate according to the MAP test and the parallel analysis. Figure 3 displays the factor loadings for the three-factor model with PAF analysis and promax rotation. The factor loadings were taken from the pattern matrix representing the regression coefficients between the items and factors. The correlation coefficients between the factors are 0.59 (between factors 1 and 2), 0.63 (between 1 and 3), and 0.42 (between 2 and 3). Four of the 32 graph items were skipped since they cannot be assigned to a single concept (i.e., items 7, 19, 23, and 37).

Figure 3 shows that the items split up into three groups that load on different factors. Regarding the association of items the factors may be interpreted as the mathematical concepts of rate (F1), vector in one dimension (F2), and area under the curve (F3). The correlations between the factors are considerably high, especially between the factors 1 and 3. This is plausible as the rate concept is mathematically related to the area concept. Instead of integrating velocity or acceleration over time, students can alternatively take suggested solutions and differentiate them to get the result.

In models with more than three factors, we find two interesting effects. First, the group of items from the

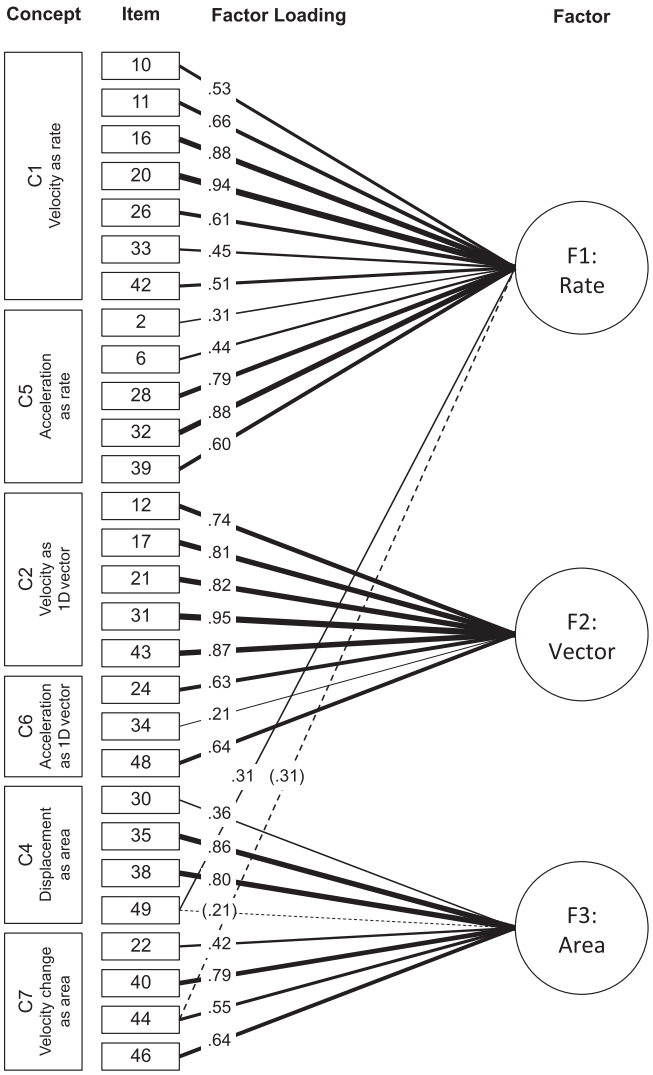

FIG. 3. Factor loadings of the graph items on three factors. The left side shows the assignment of the items to the corresponding concepts. The EFA was conducted with PAF. For clarity, only loadings above 0.20 are shown in the figure. Loadings above 0.30 are commonly considered to be significant. The solid lines mark the highest loadings of the individual items. Lower loadings are indicated with dotted lines.

three-factor model split up into qualitative items (rate, vector, and area) and quantitative items (rate and area). Second, the qualitative items further divides according to the physics concepts. Hence, in an eight-factor model, we find six groups of qualitative items according to the six kinematics concepts shown in Fig. 3 and two groups of quantitative items, one consisting of the rate items (16, 20, 28, 32) and the other one consisting of the area items (35, 38, 40, 46).

The remaining 17 items of the test belong to the picture and table representations. Here the scree test with OC and the MAP test suggested two factors whereas parallel analysis resulted in three factors. We conducted the EFA for both the two- and three-factor models. It turns out that

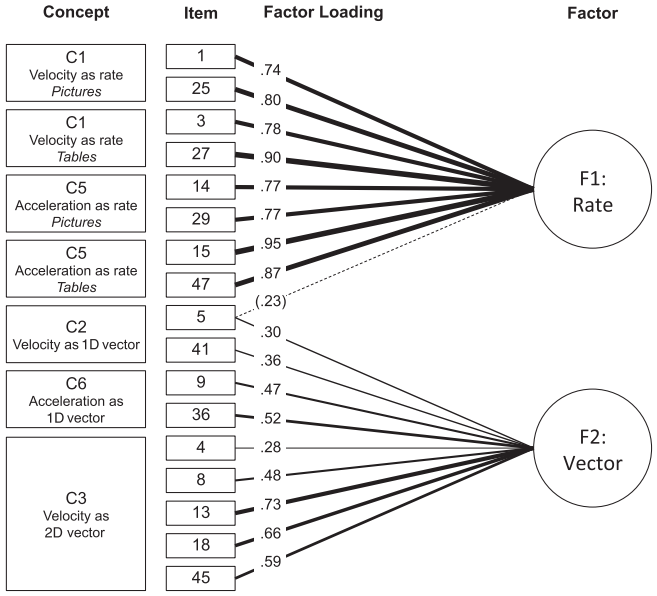

FIG. 4. Factor loadings of the picture and table items on two factors. The left side shows the assignment of the items to the corresponding concepts. The EFA was conducted with PAF. Again only loadings above 0.20 are shown in the figure. Loadings above 0.30 are considered to be significant. The solid lines mark the highest loadings on the individual items. Lower loadings are indicated with dotted lines.

with two factors the items group again according to the mathematical concepts of rate and vectors in one and two dimensions. Figure 4 displays the factor loadings of the EFA. Again, the regression coefficients between the items and factors are presented. The correlation between the factors is 0.45. However, in the case of three factors the vector factor splits up in items related to vectors in one (items 5, 41 9, 36) and two dimensions (items 4, 8, 13, 18, 45).

To summarize the results above, the items group according to the mathematics concepts of rate and vector if we analyze the representations separately (graphs; pictures and tables). The question remains whether in the analysis of all 45 items the rate and vector items from different representations load on the same factor or not.

The examination of the number of factors for the whole data set (45 items) revealed a range from three up to eight factors. We conducted a PAF analysis with promax rotation for every number of factors within this range. The five-factor model allowed the best interpretation considering the previous results. The factor loadings are presented in Table VII. The items group exactly the same way as in the previous analysis. The factors 1 (pictures and tables) and 2 (graphs) may be interpreted as the rate factors, 3 (graphs) and 5 (pictures and tables) as the vector factors and 4 (graphs) as the area factor. This result suggests that learners do not use the common conceptual structure of items with different representations. They rather process items with graphs separately from items with pictures and tables.

TABLE VII. Pattern matrix for the five-factor model gained from PAF analysis with promax rotation. All 45 items are included in the analysis. Values below 0.20 are shown only in the absence of significant values (≥0.30).

			Factors				
Representation	Concepts	Item	1	2	3	4	5
Items with graphs	C1 Velocity as rate	10		0.71			
		11		0.73			
		16		0.73			
		20		0.77			
		26		0.72			
		33		0.56			
		42		0.64			
	C5 Acceleration	2	(0.23)	(0.18)			
	as rate	6	(0.26)	(0.26)			
		28		0.68			
		32		0.70			
		39		0.58			
	C2 Velocity as	12			0.71		
	1D vector	17			0.77		
		21			0.79		
		31			0.91		
		43			0.81		
	C6 Acceleration	24			0.58		
	as 1D vector	34			(0.18)		0.53
		48			0.60		
	C4 Displacement	30				0.30	
	as area	35				0.86	
		38				0.88	
		49				(0.19)	(0.20)
	C7 Velocity change	22				0.36	
	as area	44				0.48	
		40				0.73	
		46				0.57	
Items with pictures	C1 Velocity as rate	1	0.73				
and tables		25	0.81				
		3	0.73				
		27	0.82				
	C5 Acceleration	14	0.73				
	as rate	29	0.79				
		15	0.92				
		47	0.88				
	C2 Velocity as 1D	5	(0.14)	(0.19)			(0.15)
	vector	41			0.34		(0.20)
	C6 Acceleration as 1D	9					0.75
	vector	36					0.74
	C3 Velocity as 2D	4	(0.16)	(0.28)			(0.13)
	vector	8					0.30
		13					0.48
		18					0.40
		45					0.43

If we perform an EFA with higher numbers of factors again the items group according to their assigned physical concepts. In a 13-factor model we were able to resolve all physical concepts, although the factor loadings in many cases were rather small. The two additional factors (with respect to the eleven groups in the left column of Table VII) are formed again from splitting the rate and area items in the graph representation into qualitative and quantitative items.

D. Description of the exceptions

There are seven items that do not perfectly fit our interpretation. Items 2 and 6 both load on factors 1 and

2. This can be explained by the correlation of 0.55 between the two factors, which are both interpreted as rate even though in different representations. We find a similar explanation for items 34 and 41, which load on the vector factors in both representations. The relatively high difficulty coefficients and low discrimination indices of items 4 and 5 seem to be plausible reasons for the low loadings of these items. If an item is solved by 80% of the students then the correlation with other items is typically low. Therefore, item 4 and item 5 load in a random fashion and only little on several factors. Concerning the low loadings of item 49 we conducted interviews with students. The evaluation revealed that some of the students had difficulties to decide whether the area under the curve for runner *A* or the area under the curve for runner *B* was larger. As a consequence some of the 94 students picked the wrong answer (*A*) even though they applied the right concept. Therefore, we assume to have false negatives for item 49. The distribution of answers supports this assumption: distractor *A* was chosen frequently (94 times, 14%). We suggest changing item 49 by slightly decreasing the area for runner *A*. A nonvalidated adapted version of the item is included in the KCT in the Supplemental Material [22].

E. Conclusions from the EFA

The EFA shows that the physical concepts, we have initially defined, can actually be identified in the student data. We therefore suggest that the KCT not only allows measuring the overall kinematics knowledge, but also the knowledge regarding the seven single concepts listed in Table I. Consequences of the EFA results for instruction are discussed in the following section.

V. DISCUSSION

The KCT presented in this paper is a test designed for education research to evaluate the conceptual knowledge of Gymnasium and high school students in kinematics. It consists of 49 multiple-choice items, which can be assigned to seven different kinematics concepts. It was developed and optimized in several cycles with feedback from students and experts. To evaluate the psychometric properties of the current version of the KCT we applied it to 338 Swiss Gymnasium students and computed the reliability (KR-20), item difficulties, point-biserial correlation coefficients and discrimination indices (see Table V). The characteristics were in the standard range, with a few exceptions, discussed in the text.

The structure of the test was analyzed applying exploratory factor analysis to the test results of 664 Swiss Gymnasium students. The major result of the EFA is that we indeed found evidence for the seven kinematics concepts in the response data of the students. In addition, we found two more important results. First, the students' knowledge structure in kinematics seems to be mainly based on the mathematical concepts of rate, vector, and area under the curve. Second, students perceive problems with pictures or tables differently from problems with graphs. These results of the EFA are illustrated in Fig. 5. The KCT items can be divided into two groups of representations: "pictures and tables" and "graphs." Within these representations the items mainly collocate according to the mathematical concepts. The physics concepts of velocity and acceleration can also be seen in a more fine-grained analysis. Bektasli and White [38] found a similar hierarchy for the TUG-K. Rather than two physics concepts they extracted two mathematical concepts. Their concepts of finding or interpreting the slope of a curve and of finding or interpreting the area under a curve can be associated with the concepts of rate and area described in the present work.

The results of the KCT difficulty and structure analysis have several implications for teaching. First, students who are able to correctly answer questions about velocity are also able to correctly answer questions about acceleration. This holds for the rate, the vector, and the area concept in different contexts. Therefore, we infer that the mathematics concepts are crucial for the understanding of kinematics. It means to efficiently teach kinematics the mathematical requirements first have to be settled on solid ground [5]. These results are in line with findings by Christiansen

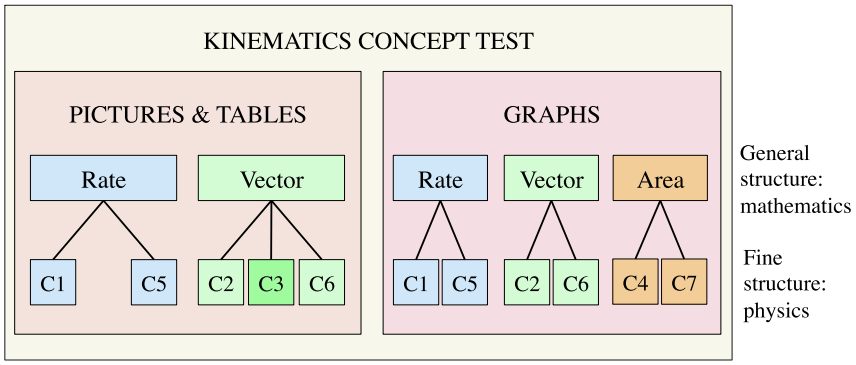

FIG. 5. Structure of the kinematics concept test.

and Thomsen [39] and by Bassok and Holyoak [40]. The former authors investigated the graphical representations of slope and derivative among third-semester students and the latter examined the interdomain transfer between isomorphic topics in physics and algebra. Both found that students who had learned arithmetic progressions were very likely to spontaneously recognize the application of the algebraic methods in kinematics. In contrast, students who had learned the physics topic first almost never showed any detectable transfer to the isomorphic algebra problems. But, even if the understanding of the mathematical concepts seems to be a requirement for learning kinematics, it is just a prerequisite and does not guarantee a successful transfer to physical concepts [5].

An unexpected result was the grouping of representations with graphs being separated from tables and pictures. Thus, as a second implication for teaching, we suggest that switching between representations should be explicitly practiced. Students should realize that the basic physics concepts are the same for the different representations although the mathematical approaches might be different. A possible explanation for this specific grouping is the use of different solution strategies necessary to solve the problems. Let us consider concept C1 as an example (velocity as rate). As shown in the left panel of Fig. 1, to obtain the velocity in a graph, first the tangent and second the slope of the tangent has to be determined by a slope triangle $\Delta x/\Delta t$. This is different from finding the velocity in tables and stroboscopic pictures. In these representations, the differences Δx and Δt can be directly read out from tables or pictures to assess the velocity by the fraction $\Delta x/\Delta t$ (Fig. 1, central and right panels). Thus, solving strategies might be one of the key factors to understand the grouping of items due to representations. However, further work has to be done to investigate the relationship between solving strategies and representations.

As a last point, it turned out that students had great difficulties with the general concept of "area under the graph" to determine either the displacement or the velocity change. Planinic, Ivanjek, and Susac [5], who investigated students' understanding of graphs in different contexts, made the same observation. We agree with their conclusion that during teaching kinematics the interpretation of the area under the graph is often neglected compared to the other concepts like rate and vector. Therefore, as a third implication for teaching, we suggest that instructions should put more emphasis on the area concept.

Further implications for teaching might be gained from the analysis of distractors used in the KCT, which allows identifying common misconceptions of students before and after teaching.

The evidence of validity we have collected so far suggests that the KCT can be used in physics education research to evaluate the overall student conceptual knowledge in kinematics and also the knowledge of students according to the seven concepts separately. By giving access to the KCT in the Supplemental Material [22], we hope that it might prove useful for other researchers in physics education.

ACKNOWLEDGMENTS

This work is supported by the Swiss National Science Foundation under Grant No. 146038. We would like to express our appreciation to all the teachers and students who volunteered their time to participate in our project. For copies of the latest KCT version please direct requests to us via email.

[1] D. Hestenes, M. Wells, and G. Swackhammer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[2] D. Hestenes and M. Wells, A mechanics baseline test, Phys. Teach. **30**, 159 (1992).
[3] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).
[4] R. J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).
[5] M. Planinic, I. Lana, and S. Ana, Comparison of university students' understanding of graphs in different contexts, Phys. Rev. ST Phys. Educ. Res. **9**, 020103 (2013).
[6] M. L. Rosenquist and L. C. McDermott, A conceptual approach to teaching kinematics, Am. J. Phys. **55**, 407 (1987).
[7] P. W. Thompson, in *The Development of Multiplicative Reasoning in the Learning of Mathematics*, edited by G. Harel and J. Confrey (SUNNY Press, Albany, NY, 1994).
[8] D. E. Trowbridge and L. C. McDermott, Investigation of student understanding of the concept of velocity in one dimension, Am. J. Phys. **48**, 1020 (1980).
[9] J. Aguirre and G. Erickson, Students' conceptions about the vector characteristics of three physics concepts, J. Res. Sci. Teach. **21**, 439 (1984).
[10] L. C. McDermott, M. L. Rosenquist, and E. H. van Zee, Student difficulties in connecting graphs and physics: Examples from kinematics, Am. J. Phys. **55**, 503 (1987).
[11] D. H. Nguyen and N. S. Rebello, Students' understanding and application of the area under the curve concept in physics problems, Phys. Rev. ST Phys. Educ. Res. **7** (2011).
[12] D. E. Trowbridge and L. C. McDermott, Investigation of student understanding of the concept of acceleration in one dimension, Am. J. Phys. **49**, 242 (1981).

[13] J. M. Aguirre, Student preconceptions about vector kinematics, Phys. Teach. **26**, 212 (1988).
[14] S. E. Tejeda Torres and H. Alarcon, A tutorial-type activity to overcome learning difficulties in understanding graphics in kinematics, Lat. Am. J. Phys. Educ. **6**, 285 (2012).
[15] S. Ainsworth, DeFT: A conceptual framework for considering learning with multiple representations, Learn. Instr. **16**, 183 (2006).
[16] G. L. Lohse, K. Biolsi, N. Walker, and H. H. Rueter, A classification of visual representations, Commun. ACM **37**, 36 (1994).
[17] G. F. Kuder and M. W. Richardson, The theory of the estimation of test reliability, Psychometrika **2**, 151 (1937).
[18] K. A. Douglas and Ş. Purzer, Validity: Meaning and relevancy in assessment for engineering education research, J. Eng. Educ. **104**, 108 (2015).
[19] T. L. Kelley, The selection of upper and lower groups for the validation of test items, J. Educ. Psychol. **30**, 17 (1939).
[20] Force Concept Inventory (1995 revision), http://modeling.asu.edu (Accessed 20.09.2016).
[21] A. Lichtenberger, C. Wagner, and A. Vaterlaus, in *E-Book Proceedings of the ESERA Conference, Nicosia, 2013*, edited by C. P. Constantinou, N. Papadouris, and A. Hadjigeorgiou (ESERA, Nicosia, 2014).
[22] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.13.010115 for the English version of the kinematics concept test.
[23] R. L. Doran, *Basic Measurement and Evaluation of Science Instruction* (National Science Teachers Association, Washington, DC, 1980).
[24] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).
[25] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).
[26] A. P. Field, J. Miles, and Z. Field, *Discovering Statistics Using R* (Sage, London, 2012).
[27] P. R. Merrifield, Factor analysis in educational research, Rev. Res. Educ. **2**, 393 (1974).
[28] T. A. Schmitt, Current methodological considerations in exploratory and confirmatory factor analysis, J. Psychol. Assess. **29**, 304 (2011).
[29] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, 2013.
[30] W. Revelle, R Package psych: Procedures for Psychological, Psychometric, and Personality Research, 2015.
[31] W. F. Velicer, Determining the number of components from the matrix of partial correlations, Psychometrika **41**, 321 (1976).
[32] J. L. Horn, A rationale and test for the number of factors in factor analysis, Psychometrika **30**, 179 (1965).
[33] R. B. Cattell, The scree test for the number of factors, Multivar. Behav. Res. **1**, 245 (1966).
[34] J. Ruscio and B. Roche, Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure, Psychol. Assess. **24**, 282 (2012).
[35] G. Raiche, R Package nFactors: Parallel Analysis and Non Graphical Solutions to the Cattell Scree Test, 2011.
[36] IBM Corp., IBM SPSS Statistics for Macintosh, Version 23.0, IBM Corp., Amonk, NY, 2014.
[37] L. K. Muthén and B. O. Muthén, *Mplus User's Guide*, 6th ed. (Muthén & Muthén, Los Angeles, CA, 1998–2011).
[38] B. Bektasli and A. L. White, The relationships between logical thinking, gender and kinematics graph interpretation skills, Euras. J. Educ. Res. **48**, 1 (2012).
[39] W. M. Christiansen and J. R. Thomsen, Investigating graphical representations of slope and derivative without a physics context, Phys. Rev. ST Phys. Educ. Res. **8**, 023101 (2012).
[40] M. Bassok and K. J. Holyoak, Interdomain transfer between isomorphic topics in algebra and physics, J. Exp. Psychol. Learn. Mem. Cogn. **15**, 153 (1989).