Testing student interpretation of kinematics graphs

Robert J. Beichner
Physics Department, North Carolina State University, Raleigh, North Carolina 29695

(Received 11 October 1993; accepted 1 March 1994)

Recent work has uncovered a consistent set of student difficulties with graphs of position, velocity, and acceleration versus time. These include misinterpreting graphs as pictures, slope/height confusion, problems finding the slopes of lines not passing through the origin, and the inability to interpret the meaning of the area under various graph curves. For this particular study, data from 895 students at the high school and college level was collected and analyzed. The test used to collect the data is included at the end of the article and should prove useful for other researchers studying kinematics learning as well as instructors teaching the material. The process of developing and analyzing the test is fully documented and is suggested as a model for similar assessment projects.

I. INTRODUCTION

The development and analysis of the Test of Understanding Graphs in Kinematics (TUG-K) will be described in this article. This is intended to serve two purposes. First, the results of a study aimed at uncovering student problems with interpreting kinematics graphs will be reported. This kind of knowledge can be very helpful before, during, and after instruction. Physics teachers tend to use graphs as a sort of second language, assuming their students can extract most of their rich information content. These results indicate that this is often an incorrect assumption. The secondary purpose of this article is to propose a model for creating research oriented multiple choice tests which can be used as diagnostic tools and for formative and summative evaluation of instruction.

There has been a long term interest in creating good physics tests.[1] A series of letters to this journal[2-5] discussed pros and cons of multiple choice testing. Aubrecht and Aubrecht[6] have outlined the steps involved in developing objective tests. Unfortunately, informal discussions with members of the physics community seem to indicate that many physics teachers are still not aware of rigorous test construction and analysis methodology. This article is an attempt to address the problem by describing the steps involved in the development of a specific test. A few of the statistics that can be used while creating a test and analyzing the results will be explained. Of course, multiple choice testing is not the only way to investigate student understanding of physics. For example, the interviewing technique in use at the University of Washington by McDermott and others has proven to be extremely fruitful. The depth of probing and flexibility of questioning provided by interviewing can be a very powerful tool. On the other hand, the ability to statistically analyze data from large numbers of objectively graded multiple choice exams may allow greater generalizability of the findings, albeit with lower resolution than interview-based results. The ideal course of action is probably found in the combination of the strengths of both these research methodologies. Places where this blending would be an appropriate addition to this study will be noted during the discussion.

A considerable effort has been made to examine what physics students learn from their introductory classes dealing with kinematics-the motion of objects. Although it is not clear why this one area of physics instruction has received more attention than others, one might speculate that researchers have recognized the importance of this topic as a "building block" upon which other concepts are based. A more pragmatic consideration is that the early availability of microcomputer-based labs which allowed real-time measurement of position, velocity, and acceleration held the possibility of drastically changing the way these concepts are taught. Researchers were interested in knowing if the new microcomputer-based laboratory approaches to teaching were viable.[7] Regardless of the reason, it is now quite easy to find many studies of students' alternative conceptions in kinematics. The well-known Force Concept Inventory[8] and Mechanics Baseline Test[9] are excellent assessment tools based on this earlier work. Unfortunately, there is less research on students' problems with the interpretation of kinematics graphs. This project was an attempt to replicate those few existing studies, find additional difficulties if they exist, and develop a useful research tool for others interested in working in this area.

II. WHY GRAPHS?

The ability to comfortably work with graphs is a basic skill of the scientist. "Line graph construction and interpretation are very important because they are an integral part of experimentation, the heart of science." (p. 572)[10] A graph depicting a physical event allows a glimpse of trends which cannot easily be recognized in a table of the same data. Mokros and Tinker[11] note that graphs allow scientists to use their powerful visual pattern recognition facilities to see trends and spot subtle differences in shape. In fact, it has been argued[12] that there is no other statistical tool as powerful for facilitating pattern recognition in complex data. Graphs summarize large amounts of information while still allowing details to be resolved. The ability to use graphs may be an important step toward expertise in problem solving since "the central difference between expert and novice solvers in

a scientific domain is that novice solvers have much less ability to construct or use scientific representations" (p. 121)[13]. Perhaps the most compelling reason for studying students' ability to interpret kinematics graphs is their widespread use as a teaching tool. Since graphs are such efficient packages of data, they are used almost as a language by physics teachers. Unfortunately, this study indicates that students do not share the vocabulary.

III. RESULTS FROM EARLIER STUDIES

Physics teachers often report that their students cannot use graphs to represent physical reality. The types of problems physics students have in this area have been carefully examined and categorized.[11-16] Several of these studies have demonstrated that students entering introductory physics classes understand the basic construction of graphs, but have difficulty applying those skills to the tasks they encounter in the physics laboratory.

Kinematics graphs have position, velocity, or acceleration as the ordinate and time as the abscissa. The most common errors students make when working with these kinds of graphs are (1) thinking that the graph is a literal picture of the situation and (2) confusing the meaning of the slope of a line and the height of a point on the line.[11,14] The first of these might occur when a student is asked to draw a velocity versus time graph of a bicycle going downhill, uphill, and then on level road. Many students produce incorrect velocity graphs which look like the hills and valleys traversed by the bicycle. It is easy to see how the path of the bike is mistakenly taken as a cue in drawing the graph. In another situation, students asked to find the point of maximum change in a graph sometimes indicate the point of largest value. In general, students tend to find slopes more difficult than individual data points.[17] They also have a hard time separating the meanings of position, velocity, and acceleration versus time graphs.[18] Regardless of the type of errors students make, it is generally agreed that an important component of understanding the connection between reality and the relevant graphs is the ability to translate back and forth in both directions.[15]

Recognizing the importance of graphing skills and the recent interest in students' interpretation of kinematics graphs leads to the need for assessment of those skills. "The construction of a valid and reliable instrument for assessing specific graphing abilities would be a step toward establishing a base line of information on this skill" (p. 572).[10] The purpose of this study was to produce such an instrument. Most of the tests used in studies of microcomputer-based lab instruction, although usually appearing to be reasonable, would probably benefit from a rigorous development and analysis. Possibly more important is the difficulty raised by differences in each study's tests. A single, consistent assessment instrument would be helpful to researchers trying to compare results from several studies. I examined a wide variety of tests[10,19-22] to see what kinds of questions were being asked by others. Several of the TUG-K items were adapted from these resources.

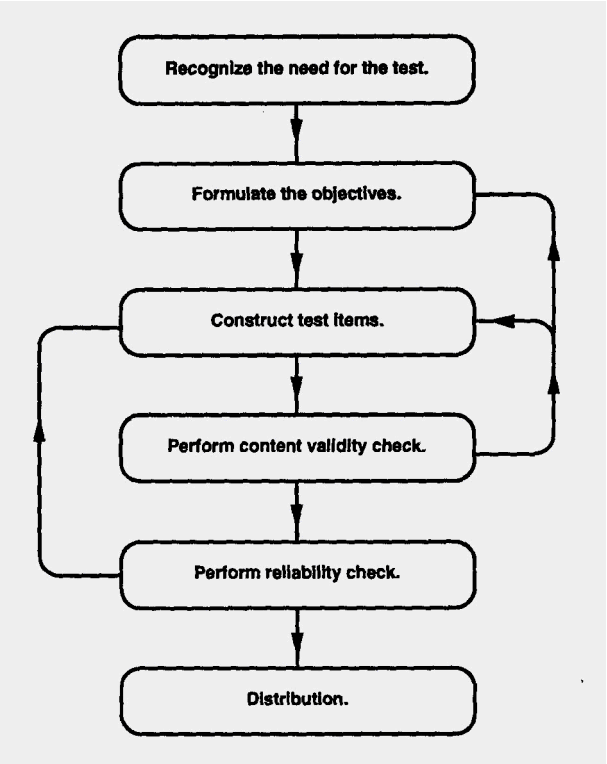

Fig. 1. A flowchart for test development showing feedback loops between steps.

IV. METHODS, DATA SOURCES, AND RESULTS

A rough flowchart of the steps involved in developing a test like this is shown in Fig. 1. Each "bubble" in the chart can actually consist of several steps. Once it is established that such an effort is worthwhile, it is necessary to formulate a list of specific objectives which relate to an understanding of kinematics graphs. In this study, eight objectives emerged from an examination of several commonly used test banks and introductory physics books,[20,23-25] and informal interviews with science teachers. After a pilot study, one objective was eliminated. Nearly all students were able to go from a point on a graph to its coordinate pair, and vice versa. Since this study was attempting to uncover student difficulties, items relating to this objective were removed from later versions of the test. The remaining objectives are listed in Table I. It is important to note that no graph construction objectives are included since this was to be a test focusing on interpretation skills.

Three items were written for each objective, producing a test of 21 multiple-choice questions. As noted earlier, several outside sources were useful in supplying items which were adapted for the TUG-K; however, most test items were written by the investigator. An effort was made to ensure that only kinematics graph interpretation skills were measured. For example, an item asking a student to "Select the graph which correctly describes the vertical component of the velocity of a ball tossed into the air," would be inappropriate since it tests knowledge of projectile motion. Items and distractors were deliberately written so as to attract students holding previously reported graphing difficulties. Another way to ensure that common errors were included as distrac-

Table I. Objectives of the Test of Understanding Graphs-Kinematics. The last column comes from data collected with the latest version of the test.

Given	The student will	Percent correct
1. Position-Time Graph	Determine Velocity	51
2. Velocity-Time Graph	Determine Acceleration	40
3. Velocity-Time Graph	Determine Displacement	49
4. Acceleration-Time Graph	Determine Change in Velocity	23
5. A Kinematics Graph	Select Another Corresponding Graph	38
6. A Kinematics Graph	Select Textual Description	39
7. Textual Motion Description	Select Corresponding Graph	43

tors was to ask open-ended questions of a group of students and then use the most frequently appearing mistakes as distractors for the multiple-choice version of the test. Aubrecht and Aubrecht[6] describe the method of classifying objectives as to level of cognitive processing required and using that as a "blueprint" for test question development.

Draft versions of the test were administered to 134 community college students who had already been taught kinematics. These results were used to modify several of the questions. These revised tests were distributed to 15 science educators including high school, community college, four year college, and university faculty. They were asked to complete the tests, comment on the appropriateness of the objectives, criticize the items, and match items to objectives. This was done in an attempt to establish content validity does the test really measure what it is supposed to?

The tests were also given to 165 juniors and seniors from three high schools and 57 four-year college physics students. As was the case with every student who was tested, all had already been exposed to kinematics through traditional instruction. After each student had taken one version of the exam they were randomly assigned to one of four different laboratory activities. These labs were approximately 2 h in length. Within a week of the lab experience, they took an alternate version of the test. This second test contained items which were created by modifying questions from the first test. For example, graph scales were shifted slightly, graphed lines were made superficially steeper or flatter, etc. The Pearson product-moment correlation between the pre- and posttest scores was 0.79, indicating that the two versions of the test were similar. (Tests which attempt to assess the same concepts in similar ways are called "parallel forms.") A paired samples t-test revealed a significant increase in the mean scores between pre- and post-lab testing ($t = 4.864$, $df = 221$, $p < 0.0l$). The *t*-test is a statistical technique that indicates whether two numbers are "significantly" different. Typically the t value, a sort of signal-to-noise ratio, is reported along with the degrees of freedom, *df,* in the calculation. *p*, the probability that the numbers really are the same, is also given. Since test scores increased and the

Table II. Statistical results from the final version of the test, taken from a national sample of 524 post-instruction high school and college students. The mean was 8.5 out of 21 items (40%) with a standard deviation of 4.6 out of 21. Means for the seven objectives are reported in Table I.

Name of statistic	Meaning	Possible values	Desired value	TUG-K Value
Standard error of the mean	Uncertainty in the mean	0 to maximum possible test score	As small as possible	0.2 out of 21
KR-20	Reliability of the whole test via calculation of the internal consistency of the items	0 to 1	≥0.70 for measurements of groups, ≥0.80 for individuals	0.83
Post-biserial Coefficient	Reliability of a single test item, defined as the correlation between the item's correctness and the whole test score	-1 to + 1	≥0.20	average 0.74
Ferguson's Delta	Discriminating ability of the whole test via how broadly it spreads the distribution of scores	0 to +1	≥0.90	0.98
Item Discrimination Index	Discriminating ability of a single item, indicating how well it distinguishes top scoring students from poorly performing students	-1 to 1	≥0.30	average 0.36

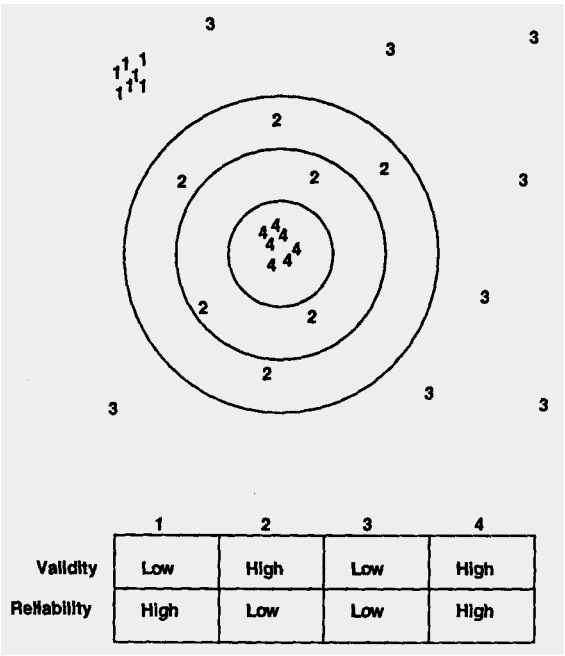

Fig. 2. A comparison of test validity and reliability. Adapted from Doran (1980). The center of the target represents what the test purports to

Table III. Point biserial coefficients and percentages of students selecting a particular choice for each test item. The correct answer is boldface.

		Point biserial	Choice					
Item	Objective	coefficient	A	B	C	D	E	Omit
1	4	0.54	41	**16**	4	22	17	0
2	2	0.80	2	10	24	2	**63**	0
3	6	0.85	8	0	20	**62**	10	0
4	3	0.74	2	14	23	**28**	32	0
5	1	0.78	3	2	**73**	18	4	0
6	2	0.66	45	**25**	6	6	16	1
7	2	0.68	**31**	20	10	28	10	1
8	6	0.75	11	11	37	**37**	5	0
9	7	0.67	7	57	5	7	**24**	0
10	4	0.70	**30**	2	62	3	3	0
11	5	0.82	28	17	11	**36**	8	0
12	7	0.86	14	**67**	8	2	9	0
13	1	0.77	10	15	9	**61**	4	0
14	5	0.90	25	**48**	15	9	3	0
15	5	0.64	**29**	24	13	8	26	0
16	4	0.67	1	39	31	**22**	7	0
17	1	0.59	**21**	46	8	7	19	0
18	3	0.84	7	**46**	32	4	10	1
19	7	0.80	19	9	**37**	12	23	0
20	3	0.82	11	6	10	2	**72**	1
21	6	0.55	**18**	72	2	5	0	2

graphing exercises taken between tests dealt exclusively with kinematics graphs, this was seen as evidence of validity.

A final version of the test was prepared from the items that best discriminated between students. A few of these questions were slightly modified to further examine interesting patterns emerging from the preliminary data analysis. This test was given to 524 college and high school students from across the country. The results are summarized in Table II.

V. DISCUSSION

The test mean score of 40% is quite low considering that the test was taken following instruction in kinematics. The case could even be made that this instruction might be better than the norm since the teachers who administered the test to their students were volunteers (mostly contacted via electronic mail). This might lead to a bias in the student population since it is possible that only good teachers would "risk" giving an outsider the opportunity to closely examine what their students were learning. It can certainly be said that the teachers who volunteered were interested in improving instruction. But the results are clear, whether the instruction was exemplary or ordinary, the students were not able to fluently interpret kinematics graphs.

The rest of the analysis indicates that TUG-K has content validity and is a reliable test of understanding of kinematics graphs for groups of high school and college level students taking introductory physics. A brief aside into the field of educational assessment is needed to help explain parts of Table II. First of all, testing theory suggests that "ideal" tests have a mean near 50%, maximizing the spread of scores. (Although this might work well for a test used for research or diagnostic purposes, this may not be the best design when the test is to be used in a normal classroom setting.) The small standard error of the mean indicates that there is not much uncertainty in the 40% mean score value. Validity and reliability are jargon from the field (see Fig. 2). Validity is really accuracy-does the test measure what we think it does? Reliability is an indicator of how precisely we made the measurement. The validity of a test is not usually calculated, it is "established." On the other hand, there are several different ways to statistically determine whether a test is reliable or not. The most common method measures the internal consistency of a test and yields a coefficient called the "KR-20." (The name comes from the formula number of the statistic's definition in a paper by Kuder and Richardson,[26] the two statisticians who developed it.) Tests having a KR20 ≥ 0.70 are generally considered to be reliable for group measurements. The quality of individual test questions is ascertained by calculation of their point-biserial coefficients. These values are simply correlations between a single item's correctness and the whole test score. Think of it this way: A good item is one that is answered correctly by those students who do well on the test as a whole and missed by those who perform poorly overall. An unreliable item would do just the opposite, tripping up those who do well on the rest of the test. These correlations usually tend to be small, so any item with a point-biserial coefficient greater than 0.20 is normally considered satisfactory. The TUG-K test items proved to have remarkably high coefficients (Table III), probably because of the care taken to develop them and the narrow, well-defined domain being tested.

Item discrimination indices are another means of measuring how well a particular question differentiates between stu-

dents. Values above 0.30 are normally satisfactory. Ferguson's delta is a whole-test statistic which indicates how broadly student performance is spread out by the test items. 0.70 is an acceptable minimum. Bruning and Kintz[27] and Kline[28] are excellent guides for computing these statistics. Ghiselli, Campbell, and Zedeck[29] provide a more theoretical discussion. A general review of testing can be found in Doran.[30]

These statistics (Tables II and III) indicate that TUG-K is certainly useful for diagnostic purposes and should be a helpful research tool. Additional analyses done within the larger sample yield additional insight. For example, it was found that calculus-based physics students did significantly better on the test (with a mean of 9.8 vs 7.4) than algebra/trigonometry-based physics students ($t = 4.87$, $df = 335$, $p < 0.01$). If the averages were closer, we would not be able to say this since the two values might differ just because of statistical "noise." For example, college students as a whole did no better than their high school counterparts ($t = 1.50$, $df = 522$, $p < 0.13$) since the p value was greater than 0.05, an arbitrarily set standard for deciding whether numbers are significantly different. The average score for a college student was 9.1, while the high school average was 8.3. The spread in scores was large enough that these two numbers could not be distinguished. The logic is similar to the Rayleigh criterion in optics. The degrees of freedom differ from these two examples because math data was not available for all students. Applying this type of analysis once more, it was found that the mean for males of 9.5 was significantly better than the 7.2 value for females ($t = 5.66$, $df = 491$, $p < 0.01$). Disturbingly, other studies have found that, in general, females do not do as well as males in science and math content areas. That appears to be the case here, also. None of the other results are unexpected. The calculus students had taken mathematics classes which dealt specifically with graphs of functions and areas under curves. Their algebra-based counterparts probably did not, and did not perform as well. The interpretation of kinematics graphs does not appear to be beyond the cognitive development of students of this age, so there is no reason to expect high school students to perform at a lower level than college students. However, the fact that many college students had physics in both high school and college-thus spending more time working with kinematics graphs yet doing no better on the test-may indicate that additional exposures to traditional teaching methods do not make much difference in students' understanding of kinematics graphs.

Once a statistical review of the test as a whole has been completed and found to be acceptable, an analysis of individual test questions is indicated. (The items are printed at the end of this article, along with pie charts describing how answer choices were distributed. If you wish to give the test to your students, feel free to photocopy the questions while blocking out the pie charts.) The most interesting items are those where there is at least one white sector that is larger than the shaded correct choice sector. That indicates that not only are many students missing the question, but they are consistently selecting the same wrong answer. A list of percentages for each choice is found in Table III. Notice how much easier it is to interpret the pie charts.

VI. ANALYSIS OF INDIVIDUAL ITEMS

Approximately 25% of the students believed that switching between kinematics variables would not change the appearance of the graph. This was detected in items 11, 14, and 15. These items also had the highest discrimination indices. Apparently students who could correctly translate from one kinematics graph to another also had the best overall understanding of kinematics graphs. This might mean that "graph-as-picture" errors are the most critical to address. If students viewed graphs as photographs of the situation, they would see no reason for the appearance of a graph to change, even though the ordinate variable changed. Although this seems reasonable, it cannot be verified from this type of assessment. Interviews or transcripts of students "thinking aloud" might shed light on why students answered as they did. This is currently being pursued and will be discussed in a future article.

As predicted by studies noted earlier, it was found that students have considerable difficulty determining slopes. However, this research indicates that this is only true for "unusual" lines. If the line went straight through the origin, 73% were able to correctly determine the slope. Question 5 required this calculation and was the easiest item on the test. However, if the tangent line did not pass through the origin as in items 6 and 17, correct answers dropped to 21% and 25%. Students very often compute the slope at a point by simply dividing a single ordinate value by a single abscissa value, essentially forcing the line through the origin. Lea[31] found that students often make assumptions about initial kinematics conditions that are incorrect. That may be what is happening here. Items 2, 7, and 17 indicate the previously reported slope/height mix-up for approximately 1/4 of the students taking the test. Students selecting answer B for item 13 might also be displaying this type of problem.

Another possible explanation for item 13 results could be kinematics variable confusion. This is more directly seen in items 9 and 21. These are both situations where a simple change of the vertical axis label from one kinematics variable to another would make the greatly favored student choices correct.

Apparently students also confuse slopes and areas. Question 1 was the hardest item on the test. (It is not recommended that a test begin with the most difficult question. This item was not expected to be so challenging.) Comparing its results with item 10 shows that students consistently select answers referring to slopes rather than area-related choices. This might be exacerbated by use of the word "change" in the questions. Results from question 18 indicate that students can often pick the correct solution of finding an area when words describing that action are presented as one of the choices. But they do much worse when they have to actually perform the calculation. Their tendency is to calculate the slope rather than the area or to read a value from the vertical axis. Both these errors were seen in number 16. There are also situations where the particular problem lends itself to a solution preferred by students. Consider for example, item 20. This was the second easiest question on the test with 72% answering correctly. One might assume from this single item that students can determine areas under curves. However, it appears that students actually noticed that the velocity was constant at 3 m/s and they simply multiplied that value by the length of the time interval-not even realizing they were finding an area! In other words, students were able to recall and use a formula ($d = vt$) to find distance covered, but could not determine the same information by looking at a graph and calculating an area. This becomes obvious by looking at the performance on item 4. If students

Table IV. Student difficulties with kinematics graphs.

Graph as Picture Errors	The graph is considered to be like a photograph of the situation. It is not seen to be an abstract mathematical representation, but rather a concrete duplication of the motion event.
Slope/Height Confusion	Students often read values off the axes and directly assign them to the slope.
Variable Confusion	Students do not distinguish between distance, velocity, and acceleration. They often believe that graphs of these variables should be identical and appear to readily switch axis labels from one variable to another without recognizing that the graphed line should also change.
Nonorigin Slope Errors	Students successfully find the slope of lines which pass through the origin. However, they have difficulty determining the slope of a line (or the appropriate tangent line) if it does not go through zero.
Area Ignorance	Students do not recognize the meaning of areas under kinematics graph curves.
Area/Slope/Height Confusion	Students often perform slope calculations or inappropriately use axis values when area calculations are required.

actually understood that they were finding an area, more than 28% would have answered this item correctly. The large fraction of wrong answers for questions 1 and 10 strengthen this conclusion.

A review of Table I shows that calculating areas to determine change in velocity from an acceleration graph was by far the most difficult objective. The rest of the objectives are in the 40% to 50% range. This is discouraging since these are skills instructors expect their students to have after instruction. The types of problems students have has been categorized in Table IV.

VI. IMPLICATIONS FOR INSTRUCTION

What can be done to address the difficulties students have with the interpretation of kinematics graphs? The first step is for teachers to become aware of the problem. Knowing that students cannot use graphs as "fluently" as they should means that in-class discussions of kinematics situations and variables cannot start by simply referring to their graphs. Students need to understand graphs before they can be used as a language for instruction. Teachers may want to utilize Arons' idea of operationally defining kinematics concepts.[32] It is possible-and probably even desirable-to use graphs to help students grasp the meaning of kinematics variables. But instruction incorporating these graphs must include thorough explanations of all the information each one relates. This study indicates that teachers must choose their own words carefully-for example, the word "change" does not automatically signify "find a slope"-and be alert for similar mistakes when students are involved in discussions amongst themselves or with the instructor.

Teachers should have students examine motion events where the kinematics graphs do not look like photographic replicas of the motion and the graph lines do not go through the origin. Students should be asked to translate from motion events to kinematics graphs and back again. Instruction should also require students to go back and forth between the different kinematics graphs, inferring the shape of one from another. ("Graphs and Tracks,"[33] David Trowbridge's popular computer simulation, does an excellent job of this). Finally, teachers should have students determine slopes and areas under curves and relate those values to specific times during the motion event. All these suggestions for modifying instruction can be summarized by one phrase-teachers should give students a large variety of "interesting" motion situations for careful, graphical examination and explanation. The students must be given (1) the opportunity to consider their own ideas about kinematics graphs and then (2) encouragement to help them modify those ideas when necessary. Teachers cannot simply tell students what the graphs' appearance should be. It is apparent from the testing results that this traditional style of instruction does not work well for imparting knowledge of kinematics graphs. Instruction that asks students to predict graph shapes, collect the relevant data, and then compare results to predictions appears to he especially suited to promoting conceptual change.[34] This is especially true when microcomputer-based labs allow real-time collection and graphing of data and is probably the main reason for the success of that particular instructional technique.

VIII. SUMMARY

This article attempted to present a model for the development of research-oriented assessment tests. This was done in the context of an actual study of student ability to interpret kinematics graphs. It is hoped that not only will the test development techniques described here be useful to others wanting to carry out similar studies, but the findings of this particular test will help teachers modify their instruction to better address student difficulties with kinematics graphs.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No MDR-9154127. Any opinions expressed in this article are the author's and do not necessarily represent the opinions of the NSF. I would like to express my appreciation to all the teachers and students who volunteered their time to administer or take the exam. Thanks also to the N. C. State Physics Education Group-especially K. Benjamin, P. Englehardt, D. Haase, K. Johnston, and J. Risley-for helpful comments and discussion. R. Doran and J. Mallinckrodt provided exceptionally thorough reviews of the paper and test. Minor revisions of the test are being made from time to time. Copies of the latest version are available from the author on paper or computer disk. If possible, direct requests via the Internet: Beichner@NCSU.edu or telephone 919-515-7226.

[1]AAPT Committee on Tests, "The 1933–34 College Physics Testing Program," Am. Phys. Teacher **2**, 129–148 (1934). You probably do not want to look this up as there is not a lot of useful information in it. It is just interesting to note how long AAPT has been involved in this issue.
[2]Alfred Bork, "Letters to the Editor," Am. J. Phys. **52**, 873–874 (1984).
[3]Robert N. Varney, "More remarks on multiple choice questions," Am. J. Phys. **52**, 1069 (1984).
[4]T. R. Sandin, "On not choosing multiple choice," Am. J. Phys. **53**, 299–300 (1985).
[5]Bruce L. Scott, "A defense of multiple choice tests," Am. J. Phys. **53**, 1035 (1985).
[6]Gordon Aubrecht II and Judith Aubrecht, "Constructing objective tests," Am. J. Phys. **51**, 613–620 (1983).
[7]Ron Thornton and David Sokoloff, "Learning motion concepts using real-time microcomputer-based laboratory tools," Am. J. Phys. **58**, 858–867 (1990).
[8]David Hestenes, Malcolm Wells, and Gregg Swackhamer, "Force concept inventory," Physics Teacher **30**, 141–158 (1992).
[9]David Hestenes and Malcolm Wells, "A mechanics baseline test," Physics Teacher **30**, 159–166 (1992).
[10]Danny L. McKenzie and Michael J. Padilla, "The construction and validation of the Test of Graphing in Science (TOGS)," J. Res. Sci. Teaching **23**, 571–579 (1986).
[11]Janice R. Mokros and Robert F. Tinker, "The impact of microcomputer-based labs on children's ability to interpret graphs," J. Res. Sci. Teaching **24**, 369–383 (1987).
[12]J. Chambers, W. Cleveland, B. Kleiner, and P. Tukey, *Graphical Methods for Data Analysis* (Wadsworth International Group, Belmont, CA, 1983).
[13]Jill Larkin, "Understanding and problem solving in physics," in *Research in Science Education: New Questions, New Directions*, edited by J. Robinson (Center for Educational Research and Evaluation, Louisville, CO, 1981), pp. 115–130.
[14]William L. Barclay, "Graphing misconceptions and possible remedies using microcomputer-based labs." Paper presented at the 1986 National Educational Computing Conference, University of San Diego, San Diego, CA (1986, June).
[15]Lillian C. McDermott, Mark L. Rosenquist, and Emily H. van Zee, "Student difficulties in connecting graphs and physics: Examples from kinematics," Am. J. Phys. **55**, 503–513 (1987).
[16]Emily H. van Zee and Lillian C. McDermott, "Investigation of student difficulties with graphical representations in physics." Paper presented at the Second International Seminar in Misconceptions and Educational Strategies in Science and Mathematics, Cornell University, Ithaca, NY (1987, July).
[17]Jay R. Price, Victor R. Martuza, and James H. Crouse, "Construct validity of test items measuring acquisition of information from line graphs," J. Ed. Psychol. **66**, 152–156 (1974).
[18]Ibrahim Halloun and David Hestenes, "Common sense concepts about motion," Am. J. Phys. **53**, 1056–1065 (1985).
[19]Heather Brasell, "The effect of real-time laboratory graphing on learning graphic representations of distance and velocity," J. Res. Sci. Teaching **24**, 385–395 (1987).
[20]Ontario Institute for Studies in Education, *The Ontario Assessment Instrument Pool—Physics, senior division.* (The Minister of Education, Toronto, Canada, 1981).
[21]Harvard Project Physics, "Unit 1: Concepts of Motion," in *The Project Physics Course Test* (Holt, Rinehart, and Winston, New York, 1970).
[22]Physical Science Study Committee, *Tests of the Physical Science Study Committee, Test 1: Space, Time and Motion* (Educational Testing Service, Princeton, NJ, 1959).
[23]David Halliday and Robert Resnick, *Physics*, 3rd ed. (Wiley, New York, 1978).
[24]F. Sears, Mark Zemansky, and Hugh Young, *University Physics*, 5th ed. (Addison Wesley, Reading, MA, 1980).
[25]Joseph Kane and Morton Sternheim, *Life Science Physics* (Wiley, New York, 1978).
[26]G. F. Kuder and M. W. Richardson, "The theory of the estimation of test reliability," Psychmetrika **2**, 151–160 (1937).
[27]James L. Bruning and B. L. Kintz, *Computational Handbook of Statistics*, 3rd ed. (Scott, Foresman, Glenview, 1987).
[28]Paul Kline, *A Handbook of Test Construction* (Methuen, New York, 1986).
[29]Edwin Ghiselli, John Campbell, and Sheldon Zedeck, *Measurement Theory for the Behavioral Sciences* (Freeman, San Francisco, 1981).
[30]Rodney L. Doran, *Basic Measurement and Evaluation of Science Instruction* (National Science Teachers Association, Washington, D.C., 1980).
[31]Susanne Lea, "Assessing degrees of student understanding of acceleration." Paper presented at the summer meeting of the American Association of Physics Teachers, Boise, ID (1993, August).
[32]Arnold Arons, *A Guide to Introductory Physics Teaching* (Wiley, New York, 1990), especially the first three chapters.
[33]David Trowbridge, *Graphs and Tracks* (Physics Academic Software, Raleigh, NC, 1994).
[34]Dewey Dykstra, "Studying conceptual change in learning physics," Sci. Ed. **76**, 615–652 (1992).