Impact of animation on assessment of conceptual understanding in physics

Melissa H. Dancy

Department of Physics, University of North Carolina at Charlotte, Charlotte, North Carolina 28223, USA

Robert Beichner

Department of Physics, North Carolina State University, Raleigh, North Carolina 27695, USA

(Received 20 June 2005; published 3 March 2006)

This study investigates the effect of computer animation on assessment and the conditions under which animation may improve or hinder assessment of conceptual understanding in physics. An instrument was developed by replacing static pictures and descriptions of motion with computer animations on the Force Concept Inventory, a commonly used pencil and paper test. Both quantitative and qualitative data were collected. The animated and static versions of the test were given to students and the results were statistically analyzed. Think-aloud interviews were also conducted to provide additional insight into the statistical findings. We found that good verbal skills tended to increase performance on the static version but not on the animated version of the test. In general, students had a better understanding of the intent of the question when viewing an animation and gave an answer that was more indicative of their actual understanding, as reflected in separate interviews. In some situations this led students to the correct answer and in others it did not. Overall, we found that animation can improve assessment under some conditions by increasing the validity of the instrument.

INTRODUCTION

Assessment is an integral component of the learning process and has traditionally been accomplished through the use of pencil and paper. Newer technologies offer alternative means of evaluating student understanding. In theory, the computer should be an asset to learning and assessment. After all, it can engage students interactively and provide an approximation of one-on-one interaction by utilizing student input to adjust the presentation of information. It can present movement, graphics, and sound, easily records student input, and facilitates communication that would otherwise be difficult or impossible. But the computer by itself does not educate. As Rieber[1] put it, asking if the computer is a good instructional device is a bit like asking, "Is a hammer a good tool to use? The answer is sure, sometimes, but it all depends. It's great for hammering and pounding nails, but pretty lousy for cutting hair."

Technology can be used to facilitate learning but it can also be an expensive waste of time, or even worse, undermine learning. But how are teachers and instructional designers to know which uses will be valuable? The purpose of this study is to add to the growing body of research in instructional technology, particularly in the use of computer animation to aid in assessing conceptual understanding of physics.

REVIEW OF THE LITERATURE

Although computer animation is theoretically valuable for both learning and assessment, it is only partly understood how it is best used. Most research has focused on differences in learning outcomes between a group that saw animations and a static or text-only group. Since only half of these studies report a significant difference (usually in favor of the animation) between the two groups,[1–3] it is important to consider the specific conditions where animation might be effective or destructive. There are indications that animation will offer the potential for increased learning when there is a need for external visualization and when the content depends on an understanding of motion.[1,4,5] There is also evidence that suggests that students may need help attending to the relevant parts of an animation.[1,6,7] In other words, students are novices at learning and need guidance in learning from animation just as they do with other means.

There has been some research conducted on the use of animation within the context of Newton's laws of motion. Much of this research was carried out by a group of psychologists concerned with perception.[8–10] Their research focused on subjects' ability to correctly identify an object's correct trajectory when viewing either an animation or a static drawing. In some cases the animation improved performance while in others there was no statistically significant difference in performance between the two groups. Although the researchers did offer some suggestions, it is not clear from this body of research why differences were seen in some cases and not others.

Along related lines of perception-oriented inquiry, other researchers have found that a computer simulation can be a credible representation of reality.[11] This was accomplished by comparing student responses to questions about actual balls moving on rails to animations of the same situations. This is an important finding as it indicates that animation (which is usually simpler to present) can be used in place of reality. However, there is evidence[12] that students are not always able to correctly pick an accurate depiction of motion, especially when the correct motion violates their expectations.

Previous research indicates that computer animation can be used to increase student learning in some situations. There are also indications from the research base that students may express their knowledge differently when asked through ani-

mation than when assessed in a more traditional manner. But more research is needed to address the specific conditions when animation can have its greatest effect and more importantly, to address the issue of *why* an animation may or may not be effective. Also, most of the previous research has focused on learning and almost none has been done with assessment, making this area in need of study.

METHOD

For this study, we wanted to learn more about how computer animation could affect the assessment of conceptual understanding of physics topics. We were interested in whether computer animation could provide more insight into students' understanding than can be ascertained by traditional pencil and paper testing. Alternately, would we find situations where animations confound the assessment effort? If the answer to either of these questions was "yes," we wanted to discern the mitigating circumstances and so provide guidance to educators and researchers who want to use technology. In order to answer these questions an animated version of a widely used instrument was developed. We tried to ensure that the animated version was as similar as possible to the static version. Tversky[13] and colleagues suggest that many studies have not been careful about this equivalence, potentially confounding the findings.

Research instrument

The Force Concept Inventory[14] (FCI) is a pencil and paper conceptual test comprised of multiple-choice items. It is based on common misconceptions about forces and motion and has been used extensively for educational research and evaluation purposes.[15,16] The FCI is typically given before and after instruction. Because the test is based on common misconceptions and grounded in everyday situations, most students feel they can answer the questions reasonably well even when taking the test before instruction on the topics it covers.

All 30 questions from version two[17] of the test were rewritten by replacing static pictures and descriptions of motion with student-controlled computer animations. The animations were developed using the PHYSLET ANIMATOR,[18,19] a scriptable Java applet designed for the creation of such animations. An example question is shown in Fig. 1, more sample questions are found in the EPAPS document.[20] Other questions and discussion of the instrument can be found elsewhere.[21]

Data collected

Student responses to animated and static FCI questions were collected in the fall of 1999. The sample consisted of 53 students taking conceptual physics at a private high school and 325 students taking calculus-based, introductory mechanics at a large state university. The students were randomly assigned to a group that answered all 30 questions in animated form, or a group that answered the questions in a static form. Because the number of computers available for testing was limited, there were more students assigned to the

static condition. All students answered the questions before formal instruction began so instructor effect was not a concern. For most students, gender was recorded. ACT scores were available for 241 of the university students.

In order to provide deeper insight into the results of the statistical analysis and to highlight interesting areas for further research, interviews were conducted. During the spring of 2000, 14 university students (who were not involved in the previous data collection) participated in intensive, one-on-one interviews in which they were asked to answer either static or animated questions while verbalizing their thoughts. The interview participants were volunteers who were taking calculus-based mechanics. Interviews were conducted as early in the semester as possible since the statistical data were based on a pretest.

For the majority of the interviews the student was asked to answer all 30 questions in order while verbalizing his or her thoughts. During this period, the interviewer remained quiet except to maintain the flow of the interview. Each student was randomly given one of two mixed versions of the FCI in which about half of the questions were animated and half were in their original form. After the student had given an answer to all questions the interviewer sometimes asked him or her to relook at some questions in the opposite format. This method proved to be enlightening. The particular questions a student was asked to look at again depended on the particular student. For example, if a student answered a particular question in animated form and gave a response that indicated to the interviewer that the animation may have had an effect, then the researcher would later ask that student to look at the same question in its static form. The student was asked if they wanted to change their answer. Although it would have been desirable to have every student relook at every question, students were too tired after answering the original 30 questions and a few additional ones to seriously consider reviewing all questions again.

Three of the 14 students interviewed were asked to first write their answers down instead of verbalizing their thoughts as they went. They then discussed their answers with the interviewer. The information from this format was not very rich and the approach was quickly abandoned.

The interviews were recorded, transcribed, and analyzed predominantly using a grounded theory methodology[22] of qualitative analysis. In this approach analysis is inductive. Theories are allowed to emerge from the data rather than beginning with a particular theory and then looking to the data to prove or disprove it. As described below, we went through several cycles of analysis, deepening our understanding and checking the validity of theories as they emerged.

After the interviews were transcribed they were read through on a question-by-question basis looking for anything that might be interesting or relevant. When possible trends were seen, the entire interview transcripts were reread for supporting or contradictory statements. Finally, the interview analysis was compared to the statistical data for each question to see if there was evidence of a relationship. For example, if in the interviews it was felt that many correct answers to the static version were a false positive but answers to the animated version were an accurate representation of
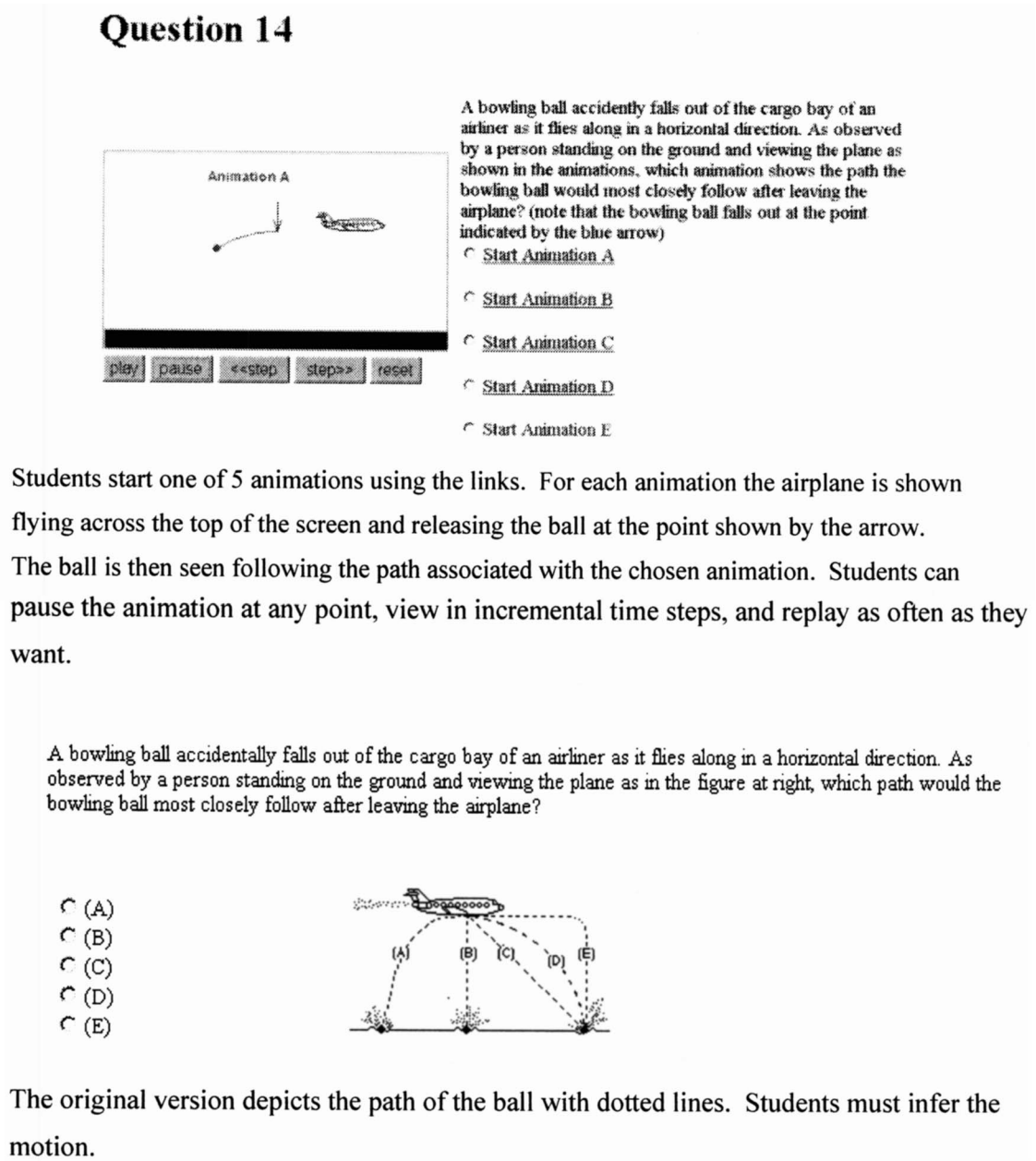
Question 14

FIG. 1. Question 14 shown in animated and original form.

students' understanding, then it would be expected that performance on the static question would be significantly higher than performance on the animated question in the large group analysis.

RESULTS

Analysis of student answers

When comparing the distribution of item responses between those who saw the animation and those who saw the static version, a full one-third of the questions showed a significant difference at the $p \leqslant 0.01$ level (using a z test for the equality between proportions).

Of these, there were six items for which the significant difference in the distribution of responses was found in the

TABLE I. Results for questions where significant differences were found in performance. The actual number of students answering each question varied because not every student answered every question.

	% Correct (N)			p value
Question	Animation	Static	z score	(two-tailed test)
1	57%(135)	71%(265)	−2.70	0.007
7	80%(134)	64%(262)	3.22	0.0012
14	76%(134)	51%(262)	4.86	<0.0001
19	34%(131)	51%(256)	−3.14	0.0016
20	24%(131)	47%(257)	−4.40	<0.0001
26	21%(126)	5%(251)	5.01	<0.0001

TABLE II. Correlation coefficients between English ACT score and FCI scores.

Format	Correlation	N
Static	0.215[a]	158
Animated	0.066	83

[a]$p < 0.01$ that $r = 0$ on a two-tailed test.

correct answer choice. There were three questions (1, 19, and 20, all found in the EPAPS document[20]) for which the static group performed significantly better and three questions (7, 14, and 26; see Fig. 1 and the EPAPS document[20]) for which the animated group performed significantly better. The results for these questions are shown in Table I.

Of the 30 questions, only 14 actually had critical information displayed during the animation. In other words, for those questions students were required to view the animation in order to correctly answer the question. For the remaining items, all relevant information was given in the problem statement making the animation a superfluous addition. It is important to note that all six questions mentioned above as significant contained animations that displayed crucial information. Thus we see that animation in and of itself will not benefit assessment. The animation must be an integral part of the question.

It is also important to note that the animations sometimes led students to the correct answer and other times did not. In terms of assessment, it is desirable that the answer a student gives is reflective of his or her understanding. An increase in performance is only desired if it is due to a more accurate understanding. A difference between the two groups does not, by itself, indicate which is superior for assessment. A deeper investigation is needed to make that determination.

It is apparent that animation can alter the outcome of assessment under some conditions. It is important to understand how and why this effect occurs. The findings described below allow us to suggest some possible explanations.

Analysis of ACT scores

The sample included 241 students who answered at least 29 of the 30 questions and for whom ACT scores were known. Correlation coefficients were calculated between English/math ACT scores for these students and their FCI scores. The results are shown in Tables II and III.

These data suggest that both verbal and mathematical skills play a significant role in the performance of students on conceptual physics questions when assessed traditionally. Interestingly, the correlation between verbal skill and perfor-

TABLE III. Correlation coefficients between math ACT score and FCI scores.

Format	Correlation	N
Static	0.300[a]	158
Animated	0.357[a]	83

[a]$p < 0.01$ that $r = 0$ on a two-tailed test.

mance disappears when the test is given in animated form. It appears that animation might be used to improve assessment of student understanding by reducing the confounding variable of verbal ability. Although verbal ability is an important skill that should be developed in a physics course, if the goal of the assessment instrument is to measure conceptual understanding then it appears that animation may provide a more accurate measure of conceptual understanding without also measuring verbal ability.

Interview results: Link between verbal skills and test performance

The interviews confirm the correlations noted above.

(1) *There were a number of instances where students simply misread a static problem.* With any assessment that requires reading, even statements that are clear and unambiguous to some test takers can be misread or misinterpreted by others. This type of mistake might cause students to either answer incorrectly when they actually had the correct conceptual understanding or to answer correctly for the wrong reasons. The interview data suggest that misreading is much more likely when an animation is *not* present.

Question 28, which asks about the forces between two students pushing off each other (shown in the EPAPS document[20]), provided clear evidence of this. The static version of this item was one of only four questions Janice managed to correctly answer. The reason she gave for her choice was "Both of them are exerting a force but it has equal amounts so neither of them are moving." She clearly did not understand that the students moved after the push even though it is explicitly stated. Later in the interview, she viewed the animated form of this question and was asked if she still felt her original answer was correct. Janice commented that she had originally misinterpreted the question and changed her answer to D, which was more reflective of her actual understanding even though it was an incorrect response.

In contrast, Charlie missed only nine questions, including this particular item. He stated there were no forces between the students because "after the push they're still together, moving…the only way I can see that as feasible is if they didn't produce any force on each other." When this student later saw the animated form of the question he changed his answer to the correct response because, as he stated, "the animation clarified it."

The static version of the questions were more likely to be misread than the animated version because the animated questions require less reading and interpretation of wording. In short, students are more likely to misinterpret text they read than an animation they watch.

It is interesting to note that the assessment of both Janice and Charlie appeared to be improved by the animation. The animation helped Charlie, who generally had a better grasp of the material, to correctly answer the question. It led Janice, who did not have a correct understanding, away from the correct answer. In both cases, the animated question appears to be more valid.

(2) *Even when students correctly read a static question*

they may not be able to understand the situation presented. Animation can be used to show students a situation rather than relying on the students' abilities to understand a written or graphical description. There are some questions (such as questions 19 and 20, shown in the EPAPS document[20]) that may be particularly difficult to describe clearly without giving the answer away. In such cases, an animation can be very helpful as it allows the question to be asked and understood while maintaining its effectiveness as an assessment tool.

Several students, often the weaker ones, had trouble interpreting the strobe photo diagrams used in questions 19 and 20. They were able to interpret the animation, which showed the blocks moving across the screen. When these students answered the static form of the question their answer was either a blind guess or based on a misunderstanding of the question statement. For example, when Beverly read the static form of question 19 she attempted to understand what was being asked of her but could not. Eventually, she concluded, "I honestly have no idea what this question is asking. Or I know what it's asking, but I'm at a loss of how to figure this out so I'm going to try to take an educated guess. And my guess is [long pause] um, at 5, C."

When she later saw the animated version she still gave an incorrect response but she did appear to understand the situation presented: "On that one [the animation] it all looks different. Like the blue one is…[replays animation] I'm going to change that one to D because that time they look like they were going the same speed at one and four."

Although they did not necessarily answer correctly when presented with the animation, the animation was a better reflection of students' actual understanding of the concepts of velocity and acceleration because they were better able to understand the situation presented. As was seen with question 28, discussed above, the animated question appears to have more validity in terms of testing these concepts because it eliminates the extra task of correctly reading and interpreting the strobe representation.

(3) *The animated version is less vague.* An animation gives more information than a description of motion because it shows all aspects of the motion at all times. It is possible for a written description of motion to include all information necessary for an expert to answer the question but neglect to include information a student with misconceptions would consider important. This was apparent in question 4 (see the EPAPS document[20]).

The question statement for number 4 only states "A large truck collides head-on with a small compact car." It does not give any information about the speed of the car and the truck before or after the collision. Out of 14 students interviewed, six specifically mentioned the speed of the vehicles as a factor in their answer choice. The students who viewed the animation used the speed information and the students who were presented with the static question noted that it was important. For example, Maggie saw the static version and stated "I guess it kind of depends on, like how fast they were both going…If you had…a little car going really fast it can be as bad as a big truck going really slow."

The animated and static versions were not identical; those who saw the animation had information about speed and those who saw the static question were left to interpret it for themselves. Again, the animated question appears to be more valid because it is less ambiguous, if the goal is to measure understanding of concepts free of measuring other skills such as ability to make inferences.

(4) *Animated questions may be less likely to elicit memorized responses.* There was also some indication from the interviews that students may answer a question correctly because they remember the correct answer and not because they have an understanding of the underlying concept. For example, the first question (See the EPAPS document[20]) is about the relative time of fall for two balls of different masses. Most students who saw the static version answered correctly and specifically stated that they remembered the answer from past courses. One student even articulated that her correct answer was not reflective of her beliefs. She commented, "I've always been told that a feather and a bowling ball will fall at the same, you know I still have this little, you know it never settles very right but, I know that's what I've been told but at the same time, I'm like, I really want the heavier ball to fall faster." Mazur[23] reports a similar experience in which a student taking the paper-based FCI asked him "How should I answer these questions? According to what you taught us, or by the way I *think* about these things?"

Although the animated version of the question is not immune to correct responses based on memorization, it appears to be less susceptible. We speculate that this may be because the animated form looks less like something from a textbook and more like real life so students' responses are based more in their "everyday" (as compared to "classroom") understanding. Again, the animation appears to increase validity.

DISCUSSION

The results of this research indicate that animation can be used to increase the validity of assessment, under some conditions. Quantitative and qualitative data both show that performance on animated questions is not as closely linked to verbal ability as traditional static questions. Students were more likely to misread or misinterpret a static question with words and pictures than a question with information conveyed in an animation.

Although verbal ability is important, and does relate to overall success in coursework, it is usually not desired to measure verbal ability on a test of conceptual understanding. There are, after all, better ways of measuring verbal ability. Also, there are particular groups of students for whom assessment measures could be more highly influenced by their verbal skills. For example, an animated version of a test would probably give a better indication of the actual understanding of a non-native language speaking student than a text-based version of the test that depends on the student's ability to read and comprehend. For the same reason, the animated test would probably be more valid for very young students or students without a strong educational background.

While the findings of this study indicate that computer animation can be used to improve test validity, they also indicate that computer animation is preferable only under

certain conditions. Although we saw no indicators that computer animation can decrease the validity of an assessment instrument, the use of animation does present some difficulty. For example, hardware can fail, there are often not enough computers, software can be incompatible, and applets can fail to load. Paper is inherently simpler. Also, although no formal records were kept, it is the researchers' general feeling that the animated questions took longer for students to answer because of the time required to play the animation.

The use of animation for assessment appears to be of most value under the following conditions.

(1) The animation is an integral part of the question and not just a good-looking addition. Students should have a need to use the animation to answer the question. From this it follows that questions about motion are the best candidates for animation. This result supports the findings of earlier research.

(2) It is likely that the static form of the question could be misread or misinterpreted in a way that could be clarified by an animation. If a question is vague, or unclear to a student, then the response that student gives may not be reflective of his or her understanding. Perhaps the greatest benefit offered by animation is that it can significantly decrease such problems. This is especially true for students with poor verbal skills.

(3) Students are likely to answer a question based on what they remember rather than what they know and understand. In this case the animation is helpful if it is not as recognizable to the student as the static question.

FUTURE RESEARCH

As with most research, more questions have been raised than answered. There are two directions that would be particularly interesting to follow. First, how do such things as gender, ability, and background change the effect of anima-

tion? This study provided inconclusive evidence that different students may respond differently to an animation. The majority of students in this study were white, male, and from a middle or upper socioeconomic class. Based on the results of the small number of exceptions[21] it would be worthwhile to investigate other groups more thoroughly.

Second, and most important, how can computer animation be used to assess students in ways that have no paper parallel? We currently use a paradigm of testing that is centered on the possibilities offered by paper. This method of question delivery and response has certainly shaped the type and nature of the questions we ask. In fact, paper assessment is so much a part of our education system that it is difficult to even imagine alternatives. But we are no longer limited to paper based assessment and can begin to think outside the proverbial box. Computer animations open new doors and offer additional possibilities. For example, technology can test students interactively, constantly adapting to the student's input. Computers can also ask questions in unique ways. For example, an animation could be used to ask students "Is this object speeding up?" Also, animations can mimic real situations by giving students more information than they need and by not stating necessary information explicitly. By breaking from our current mode of assessment, we could potentially make drastic improvements in our ability to measure student understanding.

[1] L. Rieber, *Computers, Graphics, & Learning* (Wm. C. Brown Communications, Dubuque, IA, 1994).

[2] L. Rieber, Animation in computer-based instruction, Educ. Technol. Res. Dev. **38**, 77 (1990).

[3] M. Szabo and B. Poohkay, An experimental study of animation, mathematics, achievement, and attitude toward computer-assisted instruction, J. Res. Comput. Educ. **38**, 390 (1996).

[4] R. Beichner, Impact of video motion analysis on kinematics graph interpretation skills, Am. J. Phys. **64**, 1272 (1996).

[5] A. Titus, Ph.D. dissertation, North Carolina State University, 1998.

[6] B. Y. White, Designing computer games to help physics students understand Newton's laws of motion, Cogn. Instruct. **1**, 69 (1984).

[7] R. E. Mayer and R. B. Anderson, The instructive animation: Helping students build connections between words and pictures in multimedia learning, J. Educ. Psychol. **84**, 444 (1992).

[8] M. K. Kaiser, D. R. Proffitt, and K. Anderson, Judgments of natural and anomalous trajectories in the presence and absence of motion, J. Exp. Psychol. Learn. Mem. Cogn. **11**, 795 (1985).

[9] M. K. Kaiser, D. R. Proffitt, S. M. Whelan, and H. Hecht, Influence of animation on dynamical judgments, J. Exp. Psychol. Hum. Percept. Perform. **18**, 669 (1992).

[10] M. McCloskey and D. Kohl, Naïve physics: The curvilinear impetus principle and its role in interactions with moving objects, J. Exp. Psychol. Learn. Mem. Cogn. **9**, 146 (1983).

[11] A. A. Zietsman and P. W. Hewson, Effect of instruction using microcomputer simulations and conceptual change strategies on science learning, J. Res. Sci. Teach. **23**, 27 (1986).

[12] T. C. Thaden-Koch, R. J. Dufresne, W. J. Gerace, J. P. Mestre, and W. J. Leonard, A coordination class analysis of judgments about animated motion, in *Proceedings of the 2003 Physics Education Research Conference*, edited by J. Marx, S. Franklin, and K. Cummings, AIP Conference Proceedings No. 720 (AIP, Melville, New York, 2004), p. 57.

[13] B. Tversky, J. Morrison, and M. Betrancourt, Animation: can it

facilitate?, Int. J. Hum. Comput. Stud., **57**, 247 (2002).

[14] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[15] R. Hake, Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[16] R. Reif, Guest comment: Standards and measurements in physics–Why not in physics education?, Am. J. Phys. **64**, 687 (1996).

[17] The second version of the FCI is unpublished but can be obtained by contacting the authors at http://modeling.la.asu.edu/R&E/Research.html

[18] W. Christian and M. Belloni, *Physlets* (Prentice-Hall, Upper Saddle River, NJ, 2001).

[19] http://webphysics.davidson.edu/Applets/Applets.html

[20] See EPAPS Document No. E-PRSTCR-2-003601 for animated and traditional versions of questions. This document can be reached via a direct link in the online articles HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).

[21] M. Dancy, Ph.D. dissertation, North Carolina State University, 2000.

[22] A. Strauss and J. Corbin, *Basics of Qualitative Research: Grounded Theory Procedures and Techniques* (Sage Publications, Newbury Park, CA 1990).

[23] E. Mazur, *Peer Instruction* (Prentice-Hall, Upper Saddle River, NJ, 1997), p. 4.