Context sensitivity in the force concept inventory

John Stewart,* Heather Griffin, and Gay Stewart†
University of Arkansas, Physics Department, Fayetteville, Arkansas 72701, USA
(Received 12 May 2005; revised manuscript received 15 December 2005; published 2 February 2007)

The force concept inventory and a 10-question context-modified test were given to 647 students enrolled in introductory physics classes at the University of Arkansas. Context changes had an effect ranging from −3% to 10% on the individual questions. The average student score on the ten transformed questions was 3% higher than the average student score on the corresponding 10 force concept inventory questions. Therefore, the effect of contextual changes on the total of the 10 questions is not sufficient to affect normal use of the force concept inventory as a diagnostic instrument.

DOI: 10.1103/PhysRevSTPER.3.010102 PACS number(s): 01.40.Di, 01.40.Fk, 01.50.Kw

I. INTRODUCTION

Students often leave an introductory physics class with little more conceptual mastery than they had when they entered.[1] Halloun and Hestenes suggest that the difficulty students have in learning Newtonian concepts arises from the fact that before they set foot in any physics class, they already have fixed common sense beliefs learned from everyday experiences. However, these "common sense beliefs about motion are generally incompatible with Newtonian theory. Consequently, there is a tendency for students to systematically misinterpret material in introductory physics courses."[2]

Hestenes, Wells, and Swackhamer introduced the force concept inventory (FCI) in 1992,[3] and it has become a widely used tool for evaluating student comprehension of basic Newtonian concepts. This work is based on the revised version of the FCI included with Mazur.[4] The 30-question multiple-choice test challenges students to answer correctly with the one Newtonian choice over four common misconceptions. The FCI allows instructors to determine the extent to which their instruction addresses the misconceptions held by their students. The FCI has been used to show that conventional physics instruction does very little to alter student misconceptions.[1] This conclusion is consistent with studies using other conceptual instruments.[5]

The low FCI score found at many institutions using traditional modes of instruction[1] suggests that student knowledge after the completion of an introductory physics class is often incomplete, fragmentary, and still contains significant errors and misconceptions. One effect of the incomplete state of student knowledge is that a student will sometimes answer correctly on a question, but incorrectly on a closely related question; the student's application of knowledge is *uncertain*. Substantial research effort has been expended to further understand this uncertainty. Students may apply different reasoning methods based on their beliefs about what type of reasoning is appropriate for the situation[6,7] or their general beliefs about how a physics problem should be addressed.[8] A student's general attitude toward the material may also affect the effort or care used in solving problems.[9] Novice problem solvers group problems differently than expert problem solvers, and sometimes this grouping is based on problem context instead of actual problem structure.[10–12] Studies of the effects of problem context have evolved to investigate the consistency of student misconceptions[13] and the consistency of the reasoning behind those misconceptions.[14]

The unsure state of student knowledge reveals itself in performance differences that depend on the context of a question or evaluation. The misconceptions remaining in a student's knowledge may cause a sensitivity to the physical system used in the question: the physical context. Students may be sensitive to the distractors used, whether the test is multiple choice or free response, the presence of a figure, the order of distractors, or the previous questions in the evaluation: the testing context. A student's performance may be sensitive to the amount the examination is worth toward the class, the placement of the exam with respect to the covered material, or whether the exam was announced in advance: the situational context. All these effects will be considered forms of context sensitivity where a student answers differently to two closely related questions.

This work seeks to answer two questions.

(i) Are FCI questions substantially context sensitive to very restrictive context transformations?

(ii) Can the poor performance on the FCI observed at many institutions be explained by context effects; that is, do the context effects of individual problems tend to accumulate or to cancel?

The issue of the possible context sensitivity of the FCI first arose as a result of an analysis of FCI data by Huffman and Heller[15] where they found that student responses fail to cluster under the conceptual categories proposed by the FCI's creators.[3] As an explanation, Huffman and Heller suggested that students are using "bits and pieces of knowledge" to understand forces. In this case, the pieces of knowledge used may depend on how familiar the student is with the context of the question. "Students may be more familiar with hockey pucks than with rockets, and this experience with the context can affect their understanding of the concept."[15] The authors of the FCI challenge this conclusion, and a discussion was carried out in the literature.[16,17] The issue of context sensitivity is independent of Huffman and Heller's conclusion and is interesting and important in its own right.

An examination is a sequence of questions. Most physics questions can be rewritten in multiple ways, changing the wording, the physical system, or the distractors to yield a related question that tests the same physical concept. Therefore, each question can be viewed as a member of a set of

physically similar questions. In this framework, an evaluation can then be viewed as a sample of the question sets from which the evaluation questions are drawn. The effect of context changes and additional effects arising from the combination of questions will cause the total exam score S to vary with different samplings. Therefore, a distribution of exam scores could be formed and the average of the distribution, $\langle S \rangle$, calculated. The context shift of an evaluation, the change in exam score due to the context choices of the questions, can then be defined as $\delta S = S - \langle S \rangle$. If the context shifts of the questions are randomly distributed, then δS should be small. If the context shifts of the questions are not randomly distributed, a substantial context shift for the exam could result. It seems reasonable to take $\langle S \rangle$ as the real measure of student performance on the material, so a large context shift would mean the evaluation either understates or overstates student knowledge. This paper will investigate the context shift of ten questions from the FCI to determine the magnitude of the shift of the FCI by question and the context shift for the total of the ten questions. Since this work uses only two evaluation instruments, the context shift reported, δS, is the difference in the FCI score and a transformed version of the FCI, rather the preferred, but much more difficult to measure, difference between the FCI and the average score on all possible transformations of the FCI.

Many different forms of context sensitivity and its relation to the FCI and related examinations have been investigated: (i) Sensitivity to the physical system, whether the question involves bowling balls or trucks,[18–20] (ii) sensitivity to the form of the question (free response or multiple choice) or content of the distracters,[20,21] and (iii) sensitivity to testing environment, test placement, or test value.[20,22]

Studies of context sensitivity vary greatly in the degree to which the problems are transformed. This study, which makes only local well-defined transformations, uses transformations that are more restrictive than those used in most other studies and so this study should represent a lower limit on the context effect.

Bao *et al.*[23] investigated context sensitivity in the Newton's third law concept. College students in introductory courses of different technical level were given questions that exposed one type of common misconception about Newton's third law. The questions were written independently and were not related by transformation. An average maximum context shift of $\delta S = 53\%$ was observed between questions related to different classes of misconceptions about Newton's third law.

More restrictive transformations were used by Metzler[24] to study the related effect of representational sensitivity. The transformations used in this study that change the responses to symbolic and that add or remove a figure make minor alterations to the representation of the question. Metzler examined a transformation of a gravitation question involving Newton's third law from completely textual presentation to completely graphical and found a shift in pretest score of $\delta S = 7\%$. Representational transformations were also used by Dancy and Beichner[25] for all 30 FCI questions. The static pencil and paper questions in the normal FCI were replaced by questions that used the same physical context but included a computer animation. Six of the 30 questions showed a statistically significant difference between the percentage of correct responses on the two tests. For the questions showing a statistically significant change, a range of from $\delta S = -23\%$ to $\delta S = +25\%$ is reported, where a positive shift represents a higher score on the static problem. The average shift for the six questions was $\langle \delta S \rangle = 0.5\%$ (simple average of reported shifts). The authors note that the six questions showing significant shifts are questions where critical information is contained in the simulation and that the simulations sometimes include more information than the static question. As such, the transformation to a problem including an animation represents a significant transformation of the problem for the six questions reported.

Much larger contextual shifts were observed by Palmer[26] in 10th-grade students after compulsory physical science instruction. The students were asked static Newton's third law questions that differed only by context. A shift in the selection of the correct response of $\delta S = 69\%$ was observed. The different contexts in this study were qualitatively different (a book supported by a table was transformed into a book supported by a spring), and these qualitative differences caused the students to use different rules for the different situations. Studies of context sensitivity in physics education research (PER) have used qualitatively similar systems, the kinds of systems normally found in physics problems, so this result suggests the problem of context sensitivity may actually be broader than current research suggests.

Steinberg and Sabella[20] applied far-reaching context transformations to the FCI, transforming the physical system, the question format, and the value of the examination toward the grade in the class. They observed contextual shifts ranging from $\delta S = -9\%$ to $\delta S = 36\%$. These shifts are measured between the average of one to four FCI questions that probe the same concept such as Newton's third law and an exam question that involves the same concept.

Rebello and Zollman[21] investigated the transformation of FCI questions to free response and transformation to alternate sets of distractors identified from the free-response questions. They found a contextual shift ranging from 3% to 20% for the shift to free response. The questions with transformed distractors yielded both significant increases (7%) and slight decreases (−2%) in scores. The relationship between questions in the study is often complicated, and a single conceptual shift statistic does not fully capture the difference in student responses to the questions.

Finally, Henderson[22] evaluated the sensitivity of the FCI to whether the exam counted toward the class and found a shift of $\delta S = 2\%$ with the graded test having the higher score.

II. MEASUREMENT TECHNIQUE

A test was created which contains ten context-modified FCI questions. This test will be referred to as the FCIT, for FCI transformed, and is included in the Appendix. The full 30-question FCI and the 10-question FCIT were given at the University of Arkansas to students enrolled in the introductory calculus-based mechanics course. The tests were administered at the end of the semester for five semesters ranging

TABLE I. Student performance by question: The table presents the average on ten FCI questions and ten questions formed by applying specific transformations to an FCI question, the FCIT questions. The transformations are: (1) Change a concrete system to an abstract system, (2) change one concrete system to another concrete system, (3) remove redundant distractors, (4) add a figure, (5) remove a figure, (6) reorder responses, (7) remove questions from a group of questions, and (8) make distractors symbolic. The change in the percentage of students selecting the correct response δS and the percentage of students selecting the most commonly selected distractor $\delta D_{\max}$ is also reported. The change in question score, δS, is defined as δS =FCIT score−FCI score where both the FCI score and FCIT score are the percentage of students answering the question correctly. All changes are in reference to the FCI score. A positive change indicates the score on the FCIT was higher than the score on the FCI.

FCI question number	Percent FCIT	Percent FCI	δS (%)	δD_{max}	Transformations
1	92	92	−0.01	+0.62	1,3,6
4	43	42	+1.42	−3.59	2,6,8
6	86	86	−0.00	+0.46	3,7
12	92	87	+5.37	−6.29	2,3
22	74	64	+10.2	−12.6	5,7
24	85	85	+0.15	−0.15	5,7
25	34	37	−3.32	+2.99	1,4
27	78	74	+3.50	−4.13	1,4,7
28	59	55	+4.45	−2.46	2,5,6
30	65	60	+5.42	−0.24	2
Total	71	68	+2.72	−2.53	

from the spring of 2000 to the fall of 2003. Over the five semesters, over 650 students were given the FCI and FCIT. Approximately half of the students were given the FCI first followed by the FCIT test. The other half took the FCIT first. The first test had to be turned in before the second was handed out to ensure that students would not look back to the first test and change answers because of the second test. Students were told that they could earn up to five bonus points for correct responses, but they would not be penalized for incorrect answers. Those students who left many questions unanswered were eliminated from the study. There were very few students who did this. No evidence was found that students were answering using patterns (like answering all C's or ABCABC).

The tests were administered during lab sections of the introductory calculus-based mechanics course at the University of Arkansas. This course is taught in a nontraditional format with two hours of lecture and two lab sessions of two-hour duration each week. The labs use microcomputer-based laboratories and interactive-engagement methods. The class produced an average normalized conceptual gain of $\langle g \rangle = 0.5$ over the course of the experiment.

Eight types of transformations were used to create the transformed questions: (1) *Making the system abstract*: This transformation made the system more abstract, such as changing "metal balls" to "spheres" in FCI question 1. (2) *Changing the physical system*: This transformation involves changing a concrete physical system to another concrete system. For example, "truck" was changed to "bowling ball" and "compact car" was changed to "marble" in question 4. (3) *Removing redundant wrong answers*: The third transformation removed redundant wrong answers. For example, in question 6, the correct answer is a straight trajectory, and there are three choices for trajectories curving to the right. The transformation removes two of the right-curving trajectories. (4) *Adding a figure*: The fourth transformation added a figure. (5) *Removing a figure*: The fifth transformation removed a figure. (6) *Reordering multiple-choice answers*: The sixth transformation reordered multiple-choice answers. (7) *Restructure group of questions*: The seventh transformation restructured a group of questions. Questions 25 and 27 were removed from a group of three questions, and questions 22 and 24 were removed from a group of four questions. (8) *Make responses symbolic*: The eighth transformation converted textual responses to symbolic responses.

For most questions, multiple transformations were applied to insure that the question was not recognizable to the students and could not be answered by recalling their previous answer. Table I shows which transformations were applied for each question.

Studies of context sensitivity vary greatly in the degree to which the questions are modified, from making the questions free response[21] to completely rewriting the question.[20] Great care has been taken in this study to maintain the original wording of the FCI question, so that only the specific transformations listed affect the student's response to the question. The questions selected for transformation were questions that allowed transformation without extensive rewriting.

III. RESULTS

The average score out of 100% on the ten questions in the FCIT and the corresponding ten questions on the FCI is pre-

TABLE II. Change in question score by transformation. The results of Table I are summarized by context shift. The average context shift is the average of the change in question score in percent for all the questions transformed. A positive change indicates the score on the FCIT was higher than the score on the FCI.

No.	Transformation applied	Number of questions transformed	Average context shift in percent
1	Change concrete system to abstract	3	+0.06
2	Change concrete system to another concrete system	4	+4.17
3	Remove redundant distractors	3	+1.79
4	Add a figure	2	+0.09
5	Remove a figure	3	+4.93
6	Reorder responses	3	+1.95
7	Remove questions from a group	4	+3.46
8	Make responses symbolic	1	+1.42

sented in Table I. If the scores on the ten questions are representative of the FCI as a whole, then the effect of the applied transformations results in a shift of test score of δS =FCIT score−FCI score of +2.72%, so students would score about 3% higher on the transformed test.

Observation of Table I shows that while one question showed a context shift of 10% there was also one question with a negative shift and three questions with approximately zero shift. When averaged, the four questions with negative or zero shift partially canceled effects of problems with larger shifts.

Some context effect is expected for any evaluation; the $\delta S = +2.72\%$ observed is too small to account for much of the extremely poor performance of classes taught with traditional methods.[1] For the limited class of transformations applied, the FCI score should be a good estimate of the true state of student knowledge $\langle S \rangle$ as defined in the Introduction.

Table I also presents $\delta D_{\max}$, the shift of the percentage of students choosing the most commonly chosen distractor. When $\delta S + \delta D_{\max} \approx 0$, then the effect of the context change is to cause students to shift from the correct answer the most common wrong answer. When $\delta S + \delta D_{\max} \neq 0$, the effect of the context change is to cause students to consider other distractors.

To determine whether the observed differences in total scores represent a statistically significant context effect, a t-test was applied to the difference in individual student total scores with null hypothesis that the mean of the distribution of differences was zero. This showed that the difference in total score was significant at the 1% level, $t(647)=5.08$, $p < 0.01$. The statistical significance of the responses to the individual questions was also investigated using McNemar's statistic for marginal homogeneity. The differences in the student responses to each of the ten questions were all significant at the 1% level. Note that the differences in individual question average scores were only significant when treated as paired data. When treated as unpaired data, the average score by question of the FCI and FCIT was statistically significant at the 5% level only for questions 12, 22, and 30. The difference in significance between the paired and unpaired tests shows the average score on a problem obscures the effect of contextually different choices by individual students on individual problems. Students are answering differently to the paired questions but on average the different choices cancel in the problem average.

Table II summarizes the context effect by the transformation type. The average presented represents the average context shift for all questions to which the same type of transformation was applied. The context shift is positive if the score on the FCIT question was higher than the score on the FCI question. All the transformations generated a positive shift. A strong shift was observed when changing one concrete physical system to another. This is consistent with the theory that student knowledge is formed of a number of incomplete models built out of different experiences and different models are applied based on the physical context of a question. Based on this model, we had expected that the transformation from a concrete to an abstract system would yield the largest positive effect, since the abstract system would give the abstract material of the class the highest chance of being used. This was not the case, and the transformation to an abstract system yielded the weakest context effect. Presenting only part of the questions in a group of questions caused a strong context effect and seems to indicate that the responses to questions in a group cannot be treated as completely independent. Removing a figure also caused a strong contextual effect, perhaps indicating that the chance of choosing an incorrect model was enhanced by the presence of a figure.

All the above conclusions are suggestive but preliminary. It is the unfortunate nature of measurements of context sensitivity that a limited number of questions can be used and that the questions must be sufficiently different to prevent simple recall of a previous answer. Much additional experimentation is needed to understand the effect of even the very limited transformations used in this study, let alone the much broader transformations used in other studies. The primary result—that the average context shift of the full ten questions is 2.72%—is much stronger and indicates that the FCI can be used with confidence as an estimator of the average state of student knowledge.

FCI question 22 showed a 10% context shift, almost twice that of any other question. This question was the second question in a four-question group in the FCI. The first question in the group, FCI question 21, was not included in the FCIT. FCI question 21 asks the student to choose a trajectory, and FCI question 22 asks the student to find the change in speed along that trajectory. The presence of question 21 seems to cause the students to answer incorrectly more often on question 22. Question 21 presents the students with additional information, a set of possible trajectories. It appears this additional information should be considered as part of the context of question 22. Therefore, the contextual transformation of question 22 was more severe than those of other questions leading to the comparatively large δS. This also indicates that the responses to problems within groups of problems on the FCI are not independent. A figure was also removed from question 22 on the FCI, but due to the nature of the question, it seems unlikely that this would cause the students to answer incorrectly more often.

As detailed in the Introduction, measurements of contextual shifts vary greatly in the degree to which problems are transformed. Very large shifts are observed in other works when the qualitative nature of the problem is changed—for example, moving from a static to a dynamic problem in Newton's third law.[23] Large shifts are also observed when the qualitative nature of the physical system is greatly changed, for example by changing from solids to liquids.[26] The addition of animation to FCI questions generated substantial positive and negative shifts.[25] Substantial shifts are also observed if the multiple-choice distractors are changed qualitatively.[21]

In the work presented in this paper, very restrictive transformations that maintain problem wording where possible were applied; therefore, static systems remain static. The qualitative nature of the distractors is maintained, and the physical systems used are like those found in most physics problems. The contextual shifts observed for the individual problems, from −3% to 10%, are consistent with those observed in the cited works where relatively restrictive transformations are used.

No work cited calculated the average shift of a substantial group of randomly chosen FCI problems. As such, before this work it was impossible to tell if the observed contextual shifts tended to add creating a substantial shift in the total FCI score or if they tended to cancel leaving the FCI score as a good estimate of the actual state of student knowledge. Dancy and Beichner[25] transformed all problems by adding animations, and while no overall average shift is reported, their results show substantial cancellation of contextual shifts. The results presented in this paper also suggest that contextual shifts tend to cancel and the total score on the FCI is a good estimate of the state of student knowledge on Newtonian mechanics.

IV. CONCLUSION

This study sought to answer two questions.

(i) *Are FCI questions substantially context sensitive to very restrictive context transformations ?* The context shift of the ten FCI questions ranged from −3% to 10%. All ten FCI questions were statistically sensitive to context shifts when treated as paired data.

(ii)*Is the poor performance on the FCI observed at many institutions substantially explained by context effects; that is, do the context effects tend to accumulate or to cancel ?* The total context shift was only +2.72% out of 100%, so substantial cancellation of context effects takes place and the low total score on the FCI produced by traditional instruction cannot be attributed to contextual effects. The FCI should be a good estimate of the actual state of student knowledge.

ACKNOWLEDGMENT

This work is supported in part by the National Science Foundation as part of the evaluation of improved learning for the Physics Teacher Education Coalition, Grant No. PHY-0108787.

APPENDIX: FCI AND FCIT

See separate auxiliary material for the FCI questions and the corresponding FCIT questions.

*Electronic address: johns@uark.edu
†Electronic address: gstewart@uark.edu

[1] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).
[2] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985).
[3] D. Hestenes, G. Swackhamer, and M. Wells, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[4] E. Mazur, *Peer Instruction* (Prentice-Hall, Englewood Cliffs, NJ, 1997).
[5] E. Kim and S.-J. Pak, Students do not overcome conceptual difficulties after solving 1000 traditional problems, Am. J. Phys. **70**, 759 (2002).
[6] L. Lising and A. Elby, The impact of epistemology on learning: A case study from introductory physics, Am. J. Phys. **73**, 372 (2005).
[7] W. A. Sandoval, Understanding students' practical epistemologies and their influence on learning through inquiry, Sci. Educ. **89**, 634 (2005).
[8] D. B. May and E. Etkina, College physics students' epistemological self-reflection and its relationship to conceptual learning, Am. J. Phys. **70**, 1249 (2002).
[9] K. Spall and M. Stanisstreet, Development of school students' constructions of biology and physics, Int. J. Sci. Educ. **26**, 787 (2004).
[10] M. T. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices,

Cogn. Sci. **5**, 121 (1981).
[11] G. S. Gliner, College students' organization of mathematics word problems in relation to success in problem solving, Sch. Sci. Math. **89**, 392 (1989).
[12] G. S. Gliner, College students' organization of mathematics word problems in terms of mathematical structure vs. surface structure, Sch. Sci. Math. **91**, 105 (1991).
[13] E. E. Clough and R. Driver, A study of consistency in the use of students' conceptual frameworks across different task contexts, Sci. Educ. **70**, 473 (1986).
[14] D. H. Palmer, The effect of context on students' reasoning about forces, Int. J. Sci. Educ. **19**, 681 (1997).
[15] D. Huffman and P. Heller, What does the Force Concept Inventory actually measure?, Phys. Teach. **33**, 138 (1995).
[16] P. Heller and D. Huffman, Interpreting the Force Concept Inventory. a reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).
[17] D. Hestenes and I. Halloun, Interpreting the Force Concept Inventory, Phys. Teach. **33**, 502 (1995).
[18] S. Itza-Ortiz, S. Rebello, and D. Zollman, Students' models of Newton's second law in mechanics and electromagnetism, Eur. J. Phys. **25**, 81 (2004).
[19] D. H. Palmer, Measuring contextual error in the diagnosis of alternative conceptions in science, Issues Educ. Res. **8**, 65 (1998).
[20] R. N. Steinberg and M. S. Sabella, Performance on multiple-choice diagnostics and complementary exam problems, Phys. Teach. **35**, 150 (1997).
[21] S. Rebello and D. Zollman, The effect of distractors on student performance on the Force Concept Inventory, Am. J. Phys. **72**, 116 (2004).
[22] C. Henderson, Common concerns about the Force Concept Inventory, Phys. Teach. **40**, 542 (2002).
[23] L. Bao, K. Hogg, and D. Zollman, Model analysis of fine structures of student models: An example with Newton's third law, Am. J. Phys. **70**, 766 (2002).
[24] D. E. Metzler, Relation between students' problem-solving performance and representational format, Am. J. Phys. **73**, 463 (2002).
[25] M. H. Dancy and R. Beichner, Impact of animation on assessment of conceptual understanding in physics, Phys. Rev. ST Phys. Educ. Res. **2**, 010104 (2006).
[26] D. H. Palmer, Investigating the relationship between students' multiple conceptions of action and reaction in cases of static equilibrium, Res. Sci. Technol. Educ. **19**, 193 (2001).