Memory & Cognition
1992, 20 (4), 381-391

The methodology of testing naive beliefs in the physics classroom

RICK D. DONLEY and MARK H. ASHCRAFT
Cleveland State University, Cleveland, Ohio

Many undergraduates harbor a variety of misbeliefs about physical objects in motion—for instance, that a bomb will fall straight down when dropped from a moving airplane. The evidence that these misbeliefs are resistant to correction by college-level physics courses, however, has often been based on methodologies that lack adequate internal validity. We used a quasi-experimental "before and after" design to assess the impact of two college physics courses, and we examined selection-bias, test-retest, and task-format factors directly. Initial accuracy and significant improvements due to instruction varied considerably by problem category and subject group; thus, in several ways, the results refute the general conclusion that conventional physics instruction does little to correct students' misbeliefs. We conclude by advocating the quasi-experimental approach for studies of naive beliefs in physics as well as for other situations in which the impact of classroom instruction is of interest.

People often hold a surprising set of misconceptions, or *naive beliefs*, about physical objects in motion, which they demonstrate under a variety of testing situations. For example, a common error has a marble following a curved trajectory after it leaves a spiral-shaped tube, reflecting the mistaken *curved impetus* belief (McCloskey, 1983). Many descriptions of the trajectories of falling objects—for example, that of a bomb dropped from a moving airplane—demonstrate the erroneous *straight down* belief (McCloskey, Washburn, & Felch, 1983).

Research on such mistaken beliefs has led to two general conclusions. The first, that the observed misbeliefs are quite widespread, is neither in serious doubt nor at all controversial (see, e.g., Caramazza, McCloskey, & Green, 1981; Clement, 1982; Halloun & Hestenes, 1985a; McCloskey, 1983; McCloskey, Caramazza, & Green, 1980; McCloskey et al., 1983; Proffitt, Kaiser, & Whelan, 1990). Indeed, from one perspective, such naive beliefs are to be expected: "Every one of the misconceptions about motion common among students today was seriously advocated by leading intellectuals in pre-Newtonian times.... If the evaluation of common sense was so difficult for the intellectual giants from Aristotle to Galileo, we should not be surprised to find that it is a problem for ordinary students today" (Halloun & Hestenes, 1985a, p. 1056).

The second general conclusion, that these misbeliefs persist despite relevant coursework in physics, is drawn nearly as often as the first: for example, "conventional instruction is ineffective in correcting defects in [students'] knowledge" (Halloun & Hestenes, 1987, p. 455); "it has been shown by a number of studies that students often complete a physics course with some of the same misconceptions with which they began" (McDermott, 1990, p. 7; see also Reif, 1990, p. 92). In short, conventional college-level physics is held to be ineffective to some important degree, for reasons such as students' misinterpretations of classroom material (McCloskey, 1983), the "chaotic variety of contexts" in which terms like *force* are used in everyday language (Halloun & Hestenes, 1985a), and misfocused emphasis on facts as opposed to students' own knowledge structures (Halloun & Hestenes, 1987). Recently, Proffitt et al. (1990) have attributed persistence of misbeliefs to the cognitive complexity of multidimensional phenomena such as wheel rotation.

Although it may indeed be true that classroom physics instruction has sometimes failed to correct peoples' naive beliefs, we suggest that much of the evidence said to support this conclusion is suspect on a variety of methodological grounds. For the question of instructional effectiveness, most of the reported experimental designs lack sufficient control procedures to maintain acceptable internal validity. Thus, in this project we did not intend to challenge or elaborate on the theoretical explanations of naive beliefs offered elsewhere, or to discuss how science education could be reformed (but see papers in Gardner et al., 1990). Instead, our purpose was to highlight the methodological issues of applying cognitive psychology to the classroom, and to provide an acceptable test of the hypothesis that naive beliefs persist despite relevant physics instruction. Our methodological analysis is not based on original insights; indeed, it is based largely

A preliminary report of these findings was presented at the Northeast Ohio Undergraduate Psychology Research Conference (Donley, 1989). We are indebted to the members and staff of the Physics Department, especially Jearl Walker and Charles Ponyik, for their cooperation in this study, and to Michael McCloskey and three anonymous reviewers for their suggestions on earlier versions of the paper. Address reprint requests to the second author, Department of Psychology, Cleveland State University, E. 24th and Euclid Ave., Cleveland, OH 44115 (email: r0599@csuohio.bitnet).

on Campbell and Stanley's (1963) classic design treatise. It is precisely because such classic issues have often been overlooked in research on naive beliefs about physics that we deal with them here.

A second purpose should be noted as well. Naive beliefs about physics surely make up one of the better known areas of education to which cognitive psychology has been applied. The interest in such beliefs among physics educators is illustrated by the large number of articles and editorials in the *American Journal of Physics* throughout the last decade. As cognitive psychology turns to more frequent applications in the classroom, it is important to focus on the issues that must be considered in that setting.

Two kinds of comparisons, often presented together, characterize the early, exploratory literature on naive beliefs about physics. First, several reports have documented the incidence of errors among students who have completed one or another course in physics or mechanics. For example, 75% of Caramazza et al.'s (1981) college student sample, and "a large number" of Clement's (1982, p. 67), gave incorrect answers to the pendulum problem (described below; see also Gunstone, 1987; Whitaker, 1983). In the second kind of comparison, groups that differed in their physics backgrounds (e.g., college vs. high school vs. no physics coursework) were compared. Although the frequency of errors was somewhat lower for groups with greater exposure to physics coursework (see, e.g., Halloun & Hestenes, 1985b; McCloskey, 1983), subjects still showed the same kinds of errors in all three groups, and they seldom exceeded 80% accuracy.

These reports provide clear evidence of the prevalence of naive beliefs about physics. There are shortcomings, however, in the logic that concludes from this evidence that conventional instruction in physics is ineffective. The first type of design mentioned above has been used to test single groups of subjects previously exposed to the treatment of interest—college physics. Each group's performance is implicitly compared with the accuracy expected from a group free of naive beliefs—that is, universal accuracy. Because such a group is never actually tested, no explicit comparison is provided; this is thus what Campbell and Stanley (1963) called a "one-shot case study" design. Furthermore, such observations are somewhat tangential to the central question; failure to improve after relevant instruction is the issue, not the absolute level of accuracy (but cf. Gunstone, 1987).

In the second type of design mentioned above, groups of subjects with preexisting differences in physics backgrounds have been compared. When group differences have been obtained, the result has been attributed causally to differential physics instruction (notice, though, that less-than-perfect performance is still interpreted with implicit reference to universal accuracy). This design is commonly described as a static-group comparison (see Campbell & Stanley, 1963), in which subjects voluntarily exposed to some experience are compared with nonvolunteers—that is, to students without such a background. Because such a design fails to control for possible selection bias—an effect documented in at least one report (Halloun & Hestenes, 1985b)—the threat to internal validity is unacceptably high, thereby compromising the group comparisons.

In fairness, many of the early studies were conducted to demonstrate the prevalence of naive beliefs, not to assess instructional effects per se. Nonetheless, their results have been cited by others as demonstrating the inadequacy of classroom instruction. More recently, several researchers have directly tested the effectiveness of physics coursework by using "before and after" designs (e.g., Clement, 1982; Hake, 1987; Halloun & Hestenes, 1985a, 1985b; Trowbridge & McDermott, 1981). But aside from Halloun and Hestenes's work, which was pursued further in the present Experiment 2, such procedures are also difficult to evaluate for various methodological reasons. For example, prior physics background has been ignored, often by combining subgroups; prior background has been confounded with pretest and posttest; the same subjects have been pre- and posttested without benefit of an untreated control group. Furthermore, there has been little sensitivity to the possibility of reactivity, demand characteristics, and leading questions. For example, one question was as follows: "Draw one or *more* arrows showing the direction of *each* force acting on the coin when it is at point B. (Draw longer *arrows* for larger *forces*.)" (Clement, 1982, p. 67, emphasis added). The correct answer to this question was "one force."

Campbell and Stanley's discussion of designs for educational research, and of issues such as selection bias, is ideally suited to the physics classroom, as is their advice that one conduct a quasiexperimental study. By fortunate circumstance, we were able to follow this advice. The first author was one of several regular laboratory instructors for the Basic Physics Lab course at our university, who could thus administer the tests with minimal disruption of normal classroom procedure, minimal reactivity, and minimal difficulties in enlisting volunteers from the course. Further, knowing the schedule of topics within the course, and having access to other relevant groups, enabled us to collect the assessments at advantageous times.

To assess instructional effectiveness while controlling for test-retest effects, we used the "separate sample pretest-posttest" design (Design 12 in Campbell & Stanley, 1963, p. 53, also known as the "simulated before and after design"), in which *different* but comparable groups provide the pre- and posttest scores. This was possible because presumably equivalent sections of the laboratory were staffed by the same lab instructor, the first author, who taught two sections in the fall and two in the winter quarter of the school year. Statistical comparisons of performance on pretests indicated that the groups were indeed comparable, though obviously not by virtue of random assignment. One section per term was randomly selected to receive the pretest, the other to receive the posttest only. Pretested groups were also posttested, so that effects of retesting could be examined directly.

Several aspects of this procedure deserve mention. First, customary professional care was exercised to avoid experimenter/instructor bias. This task was simplified because lab instructors do not lecture or introduce new materials, but instead assist students with the exercises in a standard lab manual. Second, because we were familiar with the course curriculum, we knew that none of the questionnaire items duplicated course material in any exact or literal way; that is, none of our items was "taught" in the course. Further, the design itself also mitigates concern over bias, in that all subjects received the treatment, college physics, and all were assessed with the posttest after relevant instruction. As such, the lab instructor would not have expected any particular Basic Lab group to outperform any other. Finally, recall that our foremost purpose in the study was to conduct a methodologically acceptable test of the effects of classroom instruction. That is, neither of us was partial to one or the other possible outcome concerning the effectiveness of instruction.

The design was replicated in the Advanced Physics Lab course, which was staffed by student instructors other than the first author. Along with results from an introductory psychology group, these observations permitted a second test of instructional effectiveness as well as an explicit test of selection bias.

EXPERIMENT 1

Method

Subjects (n = 159) were recruited from student groups at Cleveland State University during the school year 1988-1989, with no regard to gender or race; three withdrew during testing.[1] All subjects gave informed consent, supplied information regarding prior physics coursework, and then completed the seven-page questionnaire booklet. The booklets were numerically coded, to maintain anonymity. What follows is a brief description of the undergraduate physics curriculum and students, with scheduling details regarding the subgroups and classroom information relevant to the design of the study. Table 1 presents a schematic diagram of the testing schedule for the various groups.

Physics Curriculum and Students

Two separate sequences introduce undergraduates to physics, Basic Lab, a noncalculus course, and Advanced Lab, with a calculus prerequisite. These appear to be fairly conventional courses, comparable to the College Physics and University Physics courses, respectively, described in Halloun and Hestenes (1985b). Each sequence consists of three consecutive courses, taken across three academic quarters, with the 5 academic credits per course divided 4:1 between lecture and laboratory. Students in Basic Lab share a common lecture section, but enroll in one of several laboratory sections; the same is true of Advanced Lab. Laboratory sessions involve no lectures; instead, the students work problems and complete conventional laboratory demonstrations. None of these problems or demonstrations directly involved the items on our questionnaire.

Basic Physics Lab. The first sequence is Basic Physics Lab, or Physics 211, 212, and 213, with approximately equal numbers of students starting the sequence in the fall and winter. Basic Physics Lab satisfies the physics requirement in programs such as biology, occupational or physical therapy, and so forth; it is normally taken in the sophomore or junior year. Lectures relevant to the laws of motion occur in Physics 211 during Week 6 of the 10-week academic quarter.

A total of 88 Basic Physics Lab students volunteered to participate as part of their weekly laboratory class. They were reassured that their scores would not be used in any way to determine course grades. They reported that they had received no prior exposure to the physics curriculum at the college level. Approximately half of our sample was enrolled during the fall quarter, and half during winter. Drop-out rates in Physics 211 seldom exceed 10%.

Advanced Physics Lab. The Physics 231, 232, and 233 sequence is designed for declared or intended physics majors, students in engineering, computer science, and other such fields, so it provides a more thorough treatment of basic physics concepts than does the Basic Lab sequence. The typical drop-out rate is approximately 5%. Advanced Lab is also normally taken during the sophomore or junior year. A total of 49 of these students agreed to participate, 23 in Physics 233 during the spring term, well after the presentation of the laws of motion during the previous fall, and 26 in Physics 231 the *following* fall, prior to their classroom exposure to the laws of motion. All students were reassured that their performance would not affect course grades. For spring students in Physics 233, prior exposure to college-level physics consisted uniformly of the first two courses in the sequence.

Introductory Psychology. Nineteen students enrolled in Psychology 121 participated in order to satisfy a class requirement; the only selection restriction was no prior exposure to college-level physics.

Materials

Five diagram problems, taken from McCloskey's (1983) report, were used to test subjects' knowledge of objects in motion—two curvilinear impetus problems (Spiral and Softball), and three falling object problems (Bomb, Cliff, and Pendulum). In each problem, the physical situation was described, a diagram of the situation was presented, and the subject was asked to complete the diagram by drawing the trajectory of the moving object. The sixth problem in each packet was an algebra problem, as in Reed's (1984) report, presented both in written and in diagram form. The final page contained a task in which subjects rated simple subtraction problems; this was included merely as a filler. We will refer throughout the paper to the six problems of interest by their names—Spiral, Softball, and so forth. The verbatim statement of each problem, and our scoring criteria, are presented in the Appendix. Illustrations of the problems are presented in Figure 1, with accompanying patterns of responding.

Our selection of this diagram task was not merely intended to facilitate comparisons with the existing literature, but was based on methodological grounds as well. In particular, interviews or combined questionnaire/interviews are often recommended over paper-and-pencil diagrams alone, for at least two related reasons. First, interviews may reveal more subtle misbeliefs; many of McCloskey's (1983) subjects, for example, reverted to naive impetus explana-

Table 1
Schedule of Observations (O_n) and Relevant Classroom Lectures on Motion (X) by Academic Quarter, Experiment 1

Group	Week 2	Week 6	Week 7
Basic Lab			
Physics 211a, fall and winter	O_1	X	O_2
Physics 211b, fall and winter		X	O_3
Advanced Lab			
Physics 231, fall*	O_4	X	
Physics 233, spring		X†	O_5
Introductory Psychology			
Winter		O_6	

*Tested in fall of subsequent year.
†Motion lectures during Week 6, previous fall.

tions when they were asked, in apparently neutral fashion, to explain their answers. Second, a student can presumably "rote memorize" a correct pathway for a diagram, giving the appearance of having overcome the naive belief by parroting a correct answer. These may be valid arguments (though parroting is possible in interviews as well). Nonetheless, interviews introduce measurement difficulties in objective scoring (see, e.g., Trowbridge & McDermott, 1981), and they can be conducted with inadequate attention to reactivity, leading questions, and demand characteristics. For example, compare McCloskey's finding with the following: "After long discussions, most students who showed obstinate beliefs were able to come to adequate justifications ... because they came to realize the inconsistency of their thinking when asked to reflect on their own arguments" (Halloun & Hestenes, 1985a, p. 1059). We would argue that the latter finding may represent teaching or response acquiescence, rather than nonreactive assessment of beliefs. Given all of these issues, we elected to use the familiar pencil-and-paper task. For generality's sake, we used Halloun and Hestenes's (1985b) multiple-choice diagnostic test in Experiment 2.

Procedure

The students were tested in groups, either during normal class sessions for the lab students or during a scheduled group testing session for the Introductory Psychology students. After completing informed consent procedures, the students were given the questionnaires and were asked to complete all seven questions as carefully as possible. No time limits were imposed; the average completion time was approximately 10-15 min. All subjects were told that the study was being conducted by a member of the psychology department and his assistant, and that results of the study would be available at a later date.

Response Scoring

We first examined two groups' questionnaires in order to define our scoring procedures (see the Appendix for the criteria). After this, we both scored all questionnaires, using the categories correct, incorrect, and omitted. We were blind to subject identity and group membership at the time of scoring.[2] Although we were prepared to request a third judge's assistance, all disagreements were resolved by discussion and/or comparison of a disputed diagram against another pathway drawn by the same subject. Our disagreement rate averaged 4.7% across groups, ranging from 0% to 10.4% by group. The majority of disputes involved the Bomb and Cliff problems, whether or not the arc-shaped pathway seemed clearly to have lost all forward motion by the time of impact ("wobbles" due to handwriting inaccuracies were ignored).

Results

Accuracy

Figure 1 illustrates the five motion problems, first with their correct answers, and then with columns showing the two most frequent incorrect responses.

Initial analyses of pretest scores for all six problems revealed no differences between the Basic Lab groups (Physics 211) in fall and winter (*ps* > .25), confirming that the groups were largely comparable. Thus, the groups were combined for all further analyses. In keeping with Campbell and Stanley's (1963) notational scheme, we refer to tests as Observations (*O*), subscripted to indicate the group/testing occasion. A summary of the data, expressed as the percentage of subjects giving correct responses per group or testing occasion, is presented in Table 2 (faculty performance is also presented in Table 2, though not discussed further; see Note 1).

We conducted four families of chi-square tests, with each questionnaire problem tested separately. Correct versus incorrect was always one of the two dimensions in these tests, and group or test occasion was the other. The first two families of analyses addressed the methodological issues of (1) test-retest and (2) selection bias effects. The last two families of tests examined the effects of classroom instruction in (3) the Basic Lab and (4) Advanced Lab groups. To simplify the presentation, all obtained chi-square values are shown in Table 3.

Note that we intend the customary statistical connotation of the term *significant*, rather than any connotation regarding mastery of course material as sufficient or significant in the sense of an educational goal. Stated differently, we do not speculate about which other effects might have been significant if sample size, and hence power, had been greater, nor do we dismiss significant gains that are modest in absolute terms (see the introduction to Experiment 2 and also Note 4 for further discussion of effect size and significance). Accordingly, effects described here as significant achieved at least the .05 level of significance by conventional tests. Because several groups' data were used in several different tests, we were concerned about a possible inflation of the alpha rate. But when more stringent significance levels were used, the clear majority of the seven significant chi-square values remained significant, six of them at the .025 level, and five at .01. Finally, because questions about selection bias and test-retest effects rely on demonstrating the *lack* of significant differences, we also note those chi-square values in the range $.05 < p < .25$—that is, results that are somewhat inconsistent with the conclusion of "no differences."

Test-retest effects. To determine whether exposure to the questionnaire at pretest affected posttest performance, we compared the posttest scores of all Basic Physics Lab subjects—those who had received the pretest versus those who had not (O_2 vs. O_3). These results indicated no significant test-retest effects (all *ps* > .25). In other words, accuracy was approximately the same, regardless of prior familiarity with the questionnaire, suggesting few if any reactivity or "test learning" effects in our testing situation (i.e., a 5-week retest interval; see also Hake, 1987; Halloun & Hestenes, 1985b).

Selection bias. We compared three sets of observations to check the possibility that accuracy would differ by group *prior* to any relevant coursework in physics—that is, to determine whether selection bias is a confound when groups electing different levels of coursework are tested. The relevant observations are from Basic Lab students on pretest (O_1), Advanced Lab students on pretest (O_4), and Introductory Psychology students (O_6). Neither physics group had experienced more than 2 weeks of college physics instruction at the time of these tests, and all groups uniformly reported no prior college physics coursework. In other words, these comparisons should be relatively straightforward tests of the null hypothesis of no selection bias. Significant differences in performance

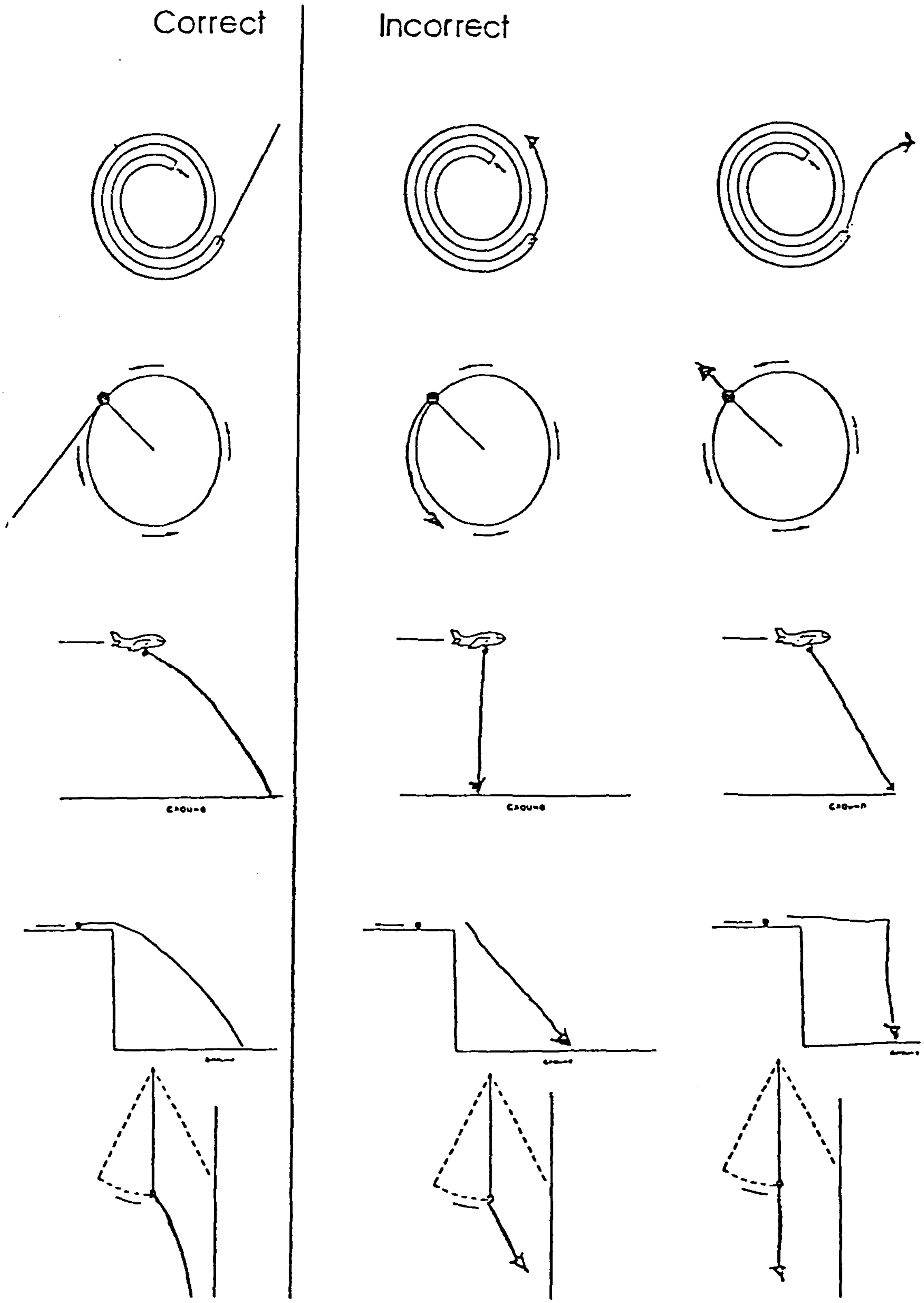

Figure 1. The correct and two most frequent incorrect pathways for the five motion problems.

Table 2
Percentage of Subjects in Each Group Responding Correctly, Experiment 1

Observation	Spiral	Softball	Bomb	Cliff	Pendulum	Pipes
1	54.0	39.5	14.5	64.5	11.0	48.0
2	76.0	56.0	26.5	74.5	15.5	40.5
3	84.5	66.5	27.5	68.5	20.5	47.0
4	55.0	64.0	42.0	65.0	23.0	56.0
5	74.0	74.0	43.0	78.0	62.0	70.0
6	75.0	47.0	5.0	47.0	27.0	47.0
7	86.0	100.0	100.0	100.0	100.0	100.0
8	100.0	75.0	25.0	50.0	50.0	88.0
9	77.0	62.0	69.0	77.0	30.0	92.0

Note—Observation (O_n) legend: (1) Basic Physics Lab, Physics 211, pretest (n = 37); (2) Basic Physics Lab, Physics 211, posttest (n = 34); (3) Basic Physics Lab, Physics 211, posttest only (n = 51); (4) Advanced Physics Lab, Physics 231 (n = 26); (5) Advanced Physics Lab, Physics 233 (n = 23); (6) Introductory Psychology, 121 (n = 19); (7) Physics faculty (n = 7); (8) Psychology faculty (n = 8); and (9) at-large faculty (n = 13).

Table 3
Observed Chi-Square Values for Four Analysis Families, Experiment 1

	Problem					
Observation	Spiral	Softball	Bomb	Cliff	Pendulum	Pipes
			1. Test-retest			
2, 3	0.68	1.11	0.10	0.62	0.29	0.33
			2. Selection bias			
1, 4, 6	1.68	4.10*	11.31†	2.07	2.72	0.68
1, 6	0.78	0.41	0.89	1.86*	2.19*	0.01
4, 6	1.28	1.18	7.70†	1.42*	0.04	0.32
1, 4	0.06	4.09†	6.68†	0.01	1.85*	0.63
			3. Basic Lab Instruction			
1, 3	7.55†	7.18†	2.46*	0.04	1.55*	0.01
			4. Advanced Lab Instruction			
4, 5	1.68*	0.55	0.01	0.99	6.22†	0.93

*.05 < p < .25. †p < .05.

would be attributable to factors such as different high school science and mathematics backgrounds, as well as others.

In the overall test, accuracy varied significantly on the Bomb problem, and marginally on the Softball problem. There was little evidence of serious selection bias in the Introductory Psychology versus Basic Lab comparisons; only two of the six problems achieved the p < .25 level. Advanced Lab students, however, were significantly superior to Introductory Psychology students on the Bomb problem, and marginally better on the Cliff problem. Finally, and most importantly, Advanced Lab students significantly outperformed Basic Lab students on the Softball and Bomb problems, and marginally on the Pendulum problem.

We conclude from these analyses that the evidence *for* selection bias effects in these groups is considerably stronger than the evidence *against* such effects. Of the 20 relevant tests (5 physics problems × 4 chi-square analyses), 4 were clearly significant, and 5 were marginally significant. In short, we reject the null hypothesis that selection bias is not a confounding variable in such research. Students electing different levels of coursework differed even prior to relevant instruction. Unfortunately, institutional records are not available to shed light on the most likely sources of the bias—differential aptitudes and backgrounds in high school physics, science, and mathematics.

The Effects of Instruction

Basic Physics Lab. We assessed the effects of relevant classroom instruction in Basic Lab 211, by means of 2×2 chi-square tests (pre- versus posttest × correct versus incorrect). As indicated earlier, this separate sample pre- versus posttest contrasted the pretest scores (O_1) with the posttest scores from the posttest only groups (O_3).

Classroom instruction on motion yielded a significant improvement on both the Spiral and Softball problems. We rule out the implausible explanation that some coincidental event or internal process (history or maturation in Campbell & Stanley, 1963) was responsible for the improvement, because both the fall and the winter sections of Physics 211 showed the same initial level of accuracy and the same improvement on posttest; the low drop-out rate argues against attrition as an explanation of the improvement. Thus, the improvements suggest that classroom instruction in Basic Physics Lab was effective, though not universally so, in counteracting the curved impetus belief (this effect was also confirmed in the examination of errors).

Despite this favorable outcome, no significant improvement due to instruction was found on the three falling objects problems. Initial levels of accuracy varied considerably, with the Cliff problem eliciting fairly high accuracy (64.5%), the Bomb and Pendulum problems very low accuracy (14.5% and 11.0%, respectively). Only slight, nonsignificant changes in accuracy were observed at the posttest. We discuss below the consistency of this pattern with McCloskey et al.'s (1983) explanation of the *straight down* belief, and in particular the role students mistakenly give to active as opposed to passive motion. Finally, no improvement on the Pipes problem was observed, despite the emphasis on mathematics and computation in Basic Physics Lab.

Advanced Lab. We compared the pretest scores from Physics 231 during fall (O_4), prior to exposure to relevant lectures, with the posttest scores from Physics 233 (O_5) collected during the previous spring—that is, from the previous year's Advanced Lab students, who had never been pretested. Only the Bomb problem showed no trend toward improvement (42%-43%). The especially difficult Pendulum problem, however, showed a substantial and significant improvement attributable to classroom instruction, from 23% to 62% correct. The remaining Softball, Cliff, and Pipes problems showed only modest gains over already respectable pretest scores.

Errors

All student groups showed evidence of the typical naive beliefs to some degree (see Figure 1), although the frequency of errors depended on group membership and time of testing. Because the errors were of the same nature as those reported elsewhere (e.g., McCloskey, 1983), we present only a brief discussion of them here.[3]

Spiral and Softball problems. The most frequent error was a counterclockwise, curved trajectory, indicating naive belief in curved impetus. On the Spiral problem, this pathway declined from 24% on pretest to only 6% on posttest in Basic Lab. For the Softball problem, the percentages were 46% and 27%, for pretest and posttest, respectively. These changes, in combination with the accuracy effects, indicate that Basic Lab indeed corrected the curvilinear impetus belief to a significant degree.

Bomb problem. Two kinds of misbeliefs appeared frequently in this problem (see McCloskey, 1983; McCloskey et al., 1983). First, some subjects believed that a passively carried object, like a bomb, acquires no "forward impetus," and will thus fall *straight down* when released. In this study, belief in the straight down pathway was most common among Basic Lab students (43%) and declined only modestly at posttest (to 36%). Advanced Lab students, however, showed fewer straight down responses on the problem, even on the pretest observation (12%). According to McCloskey et al. (1983), a perceptual illusion is at least partly responsible for the straight down belief. If so, the present results suggest that this illusion is in some way overcome (or possibly never experienced) by those who self-select into more advanced study of physics. Alternatively, advanced students may have already rejected the erroneous passive motion belief, possibly because of explicit learning.

In contrast, the forward diagonal error appeared with similar frequencies in all groups (from 21% to 26%) and showed no decline after instruction. Thus, the straight down and diagonal forward misbeliefs may be quite separate. They were distributed differently across groups, and responded differentially to instruction (also note that the forward diagonal pattern is closer to the correct trajectory, so it may be harder to reject on the basis of environmental feedback).

Cliff problem. One of the three errors not predicted by physics faculty (see Note 3) involved the "straight out then straight down" pathway for the Cliff problem, as shown in Figure 1. This error accounted for 11% of the Basic Lab responses and showed no change with instruction. Half of these errors also included reference to a forward rolling pathway after impact, reminiscent of Clement's (1982) report in which a previous horizontal force resumes its influence after a second force is removed (one student added a forward bouncing pathway in the shape of an angular capital *M*).

Pendulum problem. Even though the pendulum ball drops at the bottom of the arc, a fairly constant proportion of all groups indicated that the ball would first rise and then begin its fall (from 9% to 19%). This proportion did not change on posttest.

Pipes problem. The most commonly observed error for students was a value less than 1.5 h, averaging 22% for all student groups. Because so few subjects showed formulas or computations (see the Appendix), we do not know the basis for this error type. In the second most common error (Reed, 1984), subjects apparently computed the average of 10 h and 2 h, yielding an answer of 6 h (some responded "5 h," probably because they merely divided 10 h by 2 h). The only undergraduate group that avoided the averaging error was the Advanced Lab on posttest, showing only 4% for this strategy, compared with a mean of 16% for this error among other undergraduates.

Discussion

Students enrolling in the two different physics sequences, or in none at all (Introductory Psychology), demonstrate initial differences in accuracy and error patterns on naive physics diagrams. While this is not especially surprising, it is nonetheless evidence *for* a selection bias confound in several previous reports. Students with differing backgrounds, and students electing different levels of college instruction, cannot be treated as members of the same population. Reports in the existing literature that combined across different backgrounds should therefore be interpreted quite cautiously.

Further, conventional introductory college physics courses were found to have a significant impact on naive beliefs, at least as assessed by correct responding on the diagram task. A significant number of students enrolling in Basic Physics Lab apparently overcame their misbeliefs in curvilinear impetus. Advanced Lab students improved on a different subset of the problems. An alternative conclusion might interpret this as evidence that some misconceptions are more easily corrected than others, as opposed to viewing the improvement as due to classroom instruction. While it is surely true that the misconceptions differ in their ease of correction, denial of the involvement of instruction fails to explain why the Basic and Advanced Lab students improved on different problems. In any event, the results indicate that course-related improvements, as revealed by the diagram task, are specific to problem types and to either the level of instruction or the characteristics of students electing coursework at those levels. This conclusion is pursued further in Experiment 2.

EXPERIMENT 2

Halloun and Hestenes (1985b) reported one of the more careful "before and after" investigations of the effectiveness of college physics coursework. Although no untreated control group was included in the design, two small subgroups were tested to check for test-retest effects, and selection bias was minimized by keeping track of background and level of physics coursework. Most notable was their measurement device, an objective 36-item multiple-choice "diagnostic test," which was evaluated rather carefully for reliability and validity.

In general, Halloun and Hestenes's (1985b) data showed modest improvements from pretest to posttest. For the College Physics students, mean percent correct improved from 38% to 53%, respectively, for pre- and posttest, and from 52% to 64%, again respectively, for University Physics students. The authors' conclusion, however, was that "the small values (14%) for the gain indicate that conventional instruction has little effect on the student's basic knowledge state." They also noted that the correlations between pre- and posttests for the several groups "range between 0.60 and 0.76. These high values are statistical indicators of little change in basic knowledge" (see p. 1047, for both quotations).

We would dispute this conclusion for several reasons. First, there is the clear misinterpretation of the pre-/posttest correlations. More to the point, the conclusion rests on an a priori, rather than statistical, decision that a 14% improvement is negligible. As educators, we may agree with the sentiment that "they learned, but not enough to suit me." But in a methodologically acceptable design, significant improvement speaks for itself, even if it is less than universal in the classroom, and however modest it might be in absolute terms.[4]

Most relevant to the present study is that Halloun and Hestenes's (1985b) method of scoring the questionnaires may have been insensitive to the treatment effect. While the multiple choice test is obviously a more objective and conveniently scorable device than either diagrams or interviews, the test score was an unweighted composite of performance on six different problem types, with seriously disproportional representation of types (e.g., 19 speed questions, 2 curvilinear motion questions). Evidence from the present Experiment 1, however, suggests different improvement patterns for the different problems; for example, only on the curvilinear problems did performance improve significantly with Basic Lab instruction. Stated simply, it seems possible that significant instructional effects in Halloun and Hestenes's report may have been masked by the composite scoring method.

We therefore replicated the quasi-experimental design from Experiment 1, using the 36-item multiple choice test, and scored performance by problem type. Specifically, we tested the classes as we did in Groups 1, 3, 4, and 5 in Experiment 1: Basic Lab pretest, Basic Lab posttest only, Advanced Lab pretest, and Advanced Lab posttest only. Our purpose was to attempt a replication of the earlier results that would generalize to new subject samples and a different measurement instrument. Rather than alter questions on the test to equalize proportions, we used the published test to maintain comparability with previous reports. In our analyses, however, we examined both the composite scores and the six problem-type scores.

Method

The subjects were drawn from the physics lab courses used in Experiment 1, this time from the 1990-1991 school year. Of the total 64 subjects, 40 were enrolled in Basic Physics Lab 211, 12 were enrolled in the first quarter of the Advanced Physics Lab 231, and 12 in the third quarter of Advanced Lab 233. As before, the subjects were assured that their performance would not affect their grades, and they gave informed consent. They completed the 36-item test anonymously during normal lab sessions.

Materials

The questionnaire was the 36-item Mechanics Diagnostic Test (Halloun & Hestenes, 1985b), with 35 items presenting a five-choice set of answers, and one presenting a four-choice set. Chance performance was 20.1% (7.25 correct). Each item gave the situation being tested, which in most cases was accompanied by a diagram.

Procedure

Procedures were the same as those followed in Experiment 1. Most subjects completed the questionnaire in 20-25 min.

Response Scoring

In consultation with two physics faculty members, the 36 questions were classified into the following categories, with number of items per category noted in parentheses: speed (19), force (5), active motion (3), passive motion (2), curvilinear motion (2), and miscellaneous (5).[5] The terms *active* and *passive* refer to questions in which the falling object may be considered to be moving actively, as in the Cliff problem, or passively, as in the Bomb problem. As noted above, this distinction is itself a manifestation of a naive misconception. Nonetheless, it seemed important to maintain this (false) distinction because of the far higher accuracy in Experiment 1 on

Table 4
Mean Percentage Correct in Each Group, Experiment 2

Observation	Curvilinear	Passive	Active	Speed	Force	Misc	Composite
1.	30.7	40.4	35.8	38.5	28.5	49.2	38.0
2.	57.1	35.7	38.1	44.4	38.6	51.4	44.2
3.	45.8	41.7	41.7	46.9	35.0	43.3	44.0
4.	83.3	62.5	50.0	48.7	71.7	76.7	58.6
5.							38.0
6.							53.0
7.							51.2
8.							64.2

Note—Observation (O_n) legend: (1) Basic Physics Lab, Physics 211, pretest (n = 26); (2) Basic Physics Lab, Physics 211, posttest only (n = 14); (3) Advanced Physics Lab, Physics 231, pretest (n = 12); (4) Advanced Physics Lab, Physics 233, posttest only (n = 12); (5) College Physics pretest; (6) College Physics posttest; (7) University Physics pretest; and (8) University Physics posttest. Observations 5–8 are from Halloun and Hestenes (1985b).

the active motion Cliff problem. Thus, each subject's performance was scored on the six categories separately, as well as on the overall composite score.

Results

Table 4 shows the accuracy scores for all four groups, separately for the six problem categories, as well as the overall composite score. Because of the different numbers of items per category, all scores are reported as percentages. Halloun and Hestenes's (1985b) composite scores for college and university physics courses are also presented for comparison.

Selection Bias

Pretest scores for the Basic and Advanced Labs were compared to assess selection bias effects. Unlike in Experiment 1, none of the t tests approached significance; the closest was on curvilinear motion problems (t = 1.274, $p < .23$). Thus, the several apparent differences between Observations 1 and 3 in Table 4 are nonetheless not reliably different.

Effects of Instruction

Basic Physics Lab. If the educational effectiveness of Basic Physics Lab is evaluated on the basis of overall composite scores, the improvement in accuracy from 38% to 44.2% is nonsignificant (t = 1.004; all Basic Lab tests were conducted on 38 *df*). But improvement on the specific test on curvilinear motion, the problem type that showed a significant improvement in Experiment 1, was in fact significant here as well (t = 2.357, $p < .05$). As can be seen in Table 4, scores improved from 33.1% to 57.2% correct. Although the majority of subjects (57.1% of the group) on posttest got only one of the two questions correct, the percentage that scored zero correct was only 14.2% on postttest, compared with 50% of subjects on the pretest. In no other problem category did performance show a significant improvement in the Basic Lab data (all remaining ts < 1.238), although all but one showed modest trends toward higher accuracy. Interestingly, the posttest composite score mean (44.2%) was marginally lower than the 53% reported by Halloun and Hestenes (1985b) [$t(94)$ = 1.91, $p < .10$].

Advanced Lab. Although Advanced Lab students were more accurate overall than those in Basic Lab, the improvement in composite scores for Advanced Lab only approached conventional significance (t = 1.897, $p <$ .08, for the pre-/posttest comparison of 44.0% to 58.6%; all Advanced Lab tests conducted on 22 *df*). Thus, the obtained 14% improvement on the objective questionnaire, the same effect size found by Halloun and Hestenes (1985b), was at best marginally significant. Again, and for undetermined reasons, final performance here was lower than the Halloun and Hestenes values in Table 4, though not significantly so [$t(107)$ = 1.42, $p < .20$]. In marked contrast, however, three of the six problem types did show significant improvements due to instruction: curvilinear problems (t = 3.129), force problems (t = 2.786), and the miscellaneous category (t = 2.334). Thus, the concern that a composite score across several problem types may mask significant improvements appears to be well founded.

GENERAL DISCUSSION

We wish to highlight three major points about the present studies and results, and then conclude with a general observation. First, the data that we have presented both replicate and extend the research reported by many others on naive beliefs in physics. Many undergraduates, some even in their third academic quarter of instruction, still show evidence of various mistaken beliefs about physical objects in motion. In fact, it is possibly the case that, at the group level, only upper-level or even graduate training in physics will correct these naive beliefs (or, alternatively, that those seeking graduate training are more sophisticated to begin with). Interesting questions abound: How general is our insensitivity to environmental feedback of this nature? Is this feedback inherently weak, or is it relatively unimportant to everyday interactions with moving objects? How does the cognitive system maintain these mistaken beliefs even as the kinesthetic system takes advantage of the feedback (e.g., by learning to throw a ball)? How much transfer of understanding can be expected among the formal, common-sense, and kinesthetic systems? Answers to such questions might be of interest in a variety of ways.

Second, the results suggest some important differences among the various naive beliefs, especially in their susceptibility to correction in the physics classroom. Curved impetus beliefs, as assessed with the Spiral and Softball problems for instance, appear to be less common and somewhat more easy to correct than the naive belief that a falling object will follow a straight trajectory. The resistance of *straight down* and *diagonal forward* beliefs to correction, furthermore, seems related in some fashion to the mistaken notion that passive movement differs substantially from active movement. These patterns are at least similar to those reported by Proffitt et al. (1990), in that the events under consideration in curvilinear motion problems are probably less complex cognitively and hence easier to alter in the classroom. That is, one need only learn Newton's second law of motion—that in the absence of an external force an object continues its motion and direction—to overcome curved impetus beliefs. Falling object problems of the sort studied here, on the other hand, are solved by learning how two vectors jointly determine trajectory. As such, the latter would appear to be substantially more complex, especially since the external force, gravity, is an acceleration force; that is, it does not represent a constant *speed* force.

In Proffitt et al.'s (1990) terminology, the unidimensional curved motion problems would be less cognitively complex than the two-dimensional falling object problems. Note further that the two misbeliefs in falling object situations probably differ on several grounds. For instance, it may be possible for an individual to overcome the passive movement misconception through everyday experience (but cf. McCloskey et al., 1983). It seems patently unlikely, however, that everyday experience would reveal that gravity is an acceleration force, as opposed to one imparting constant speed, given a viewer's perspective on such events and the brevity of the events themselves. We note in passing that the diagonal forward pathway is consistent with the notion of "gravity as a constant speed" force, and we speculate that this may be a component of the misconception. In any event, it is surely valuable, from the standpoint of physics education, to understand students' naive beliefs at this more detailed level.

Finally, it is important to demonstrate that the effectiveness of classroom instruction can be determined in a "before and after" design that maintains adequate internal validity. As noted at the outset, it is interesting to know that physics students still harbor various misbeliefs, but this is not compelling evidence of a widespread ineffectiveness of physics education. Instead, it may just confirm what every educator knows, that some students do not master the entire course content. By using a quasi-experimental design that includes controls for selection bias and test-retest effects, we have found evidence of improvement due to relevant physics instruction. Furthermore, the data indicate that the level of physics instruction, represented here by the Basic Lab as opposed to the Advanced Lab groups, is also quite important. Students electing courses at different levels not only differ in their initial performance, but also respond differently to instruction. Clearly, future research cannot merely contrast "college physics students" with students who lack a background in physics, nor can it combine performance across different problem types without losing important precision.

This final conclusion can be applied to any research setting in which the effects of instruction are examined, including areas much closer to our own home than physics. For example, Kahneman and Tversky's (1973) graduate-level statistics students failed to notice a probable effect of statistical regression, despite relevant knowledge, and Tversky and Kahneman's (1971) PhD-level psychologists showed evidence of a mistaken belief in "the law of small numbers." We suggest that the quasi-experimental approach could be profitably extended to such areas of investigation. And, in a society increasingly concerned with the effectiveness of mathematics and science education, we must devise procedures and research designs that yield useful evidence.

REFERENCES

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition*, **9**, 117-123.

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, **50**, 66-71.

Donley, R.D. (1989, April). *Naive physics*. Paper presented at the meeting of the Northeast Ohio Undergraduate Psychology Research Conference, Cleveland.

Gardner, M., Greeno, J. G., Reif, F., Schoenfeld, A. H., DiSessa, A., & Stage, E. (Eds.) (1990). *Toward a scientific practice of science education*. Hillsdale, NJ: Erlbaum.

Gunstone, R. F. (1987). Student understanding in mechanics: A large population survey. *American Journal of Physics*, **55**, 691-696.

Hake, R. R. (1987). Promoting student crossover to the Newtonian world. *American Journal of Physics*, **55**, 878-884.

Halloun, I. A., & Hestenes, D. (1985a). Common sense concepts about motion. *American Journal of Physics*, **53**, 1056-1065.

Halloun, I. A., & Hestenes, D. (1985b). The initial knowledge state of college physics students. *American Journal of Physics*, **53**, 1043-1055.

Halloun, I. A., & Hestenes, D. (1987). Modeling instruction in mechanics. *American Journal of Physics*, **55**, 455-461.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237-251.

McCloskey, M. (1983). Naive theories of motion. In D. Genter & A. L. Stevens (Eds.), *Mental models* (pp. 299-324). Hillsdale, NJ: Erlbaum.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, **210**, 1137-1141.

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **9**, 636-649.

McDermott, L. C. (1990). A view from physics. In M. Gardner, J. G. Greeno, F. Reif, A. H. Schoenfeld, A. DiSessa, & E. Stage (Eds.), *Toward a scientific practice of science education* (pp. 3-30). Hillsdale, NJ: Erlbaum.

Proffitt, D. R., Kaiser, M. K., & Whelan, S. M. (1990). Understanding wheel dynamics. *Cognitive Psychology*, **22**, 342-373.

Reed, S. K. (1984). Estimating answers to algebra word problems. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 778-790.

Reif, F. (1990). Transcending prevailing approaches to science education. In M. Gardner, J. G. Greeno, F. Reif, A. H. Schoenfeld, A. DiSessa, & E. Stage (Eds.), *Toward a scientific practice of science education* (pp. 91-110). Hillsdale, NJ: Erlbaum.

Trowbridge, D. E., & McDermott, L. C. (1981). Investigation of student understanding of the concept of acceleration in one dimension. *American Journal of Physics*, **49**, 242-253.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, **76**, 105-110.

Whitaker, R. J. (1983). Aristotle is not dead: Student understanding of trajectory motion. *American Journal of Physics*, **51**, 352-357.

NOTES

1. Three faculty groups also completed the questionnaire, 7 of 14 faculty members in physics, 8 of 16 in psychology (3 others were not sampled, due to familiarity with the topic), and 13 of 25 randomly selected faculty in other arts and sciences departments. The physicists showed nearly perfect performance (one problem was missed by only 1 professor, and this most likely due to misunderstanding of the instructions), but performance in psychology and other departments was on the same par as Advanced Lab students' pretests. The only exception to this pattern was faculty performance on the algebraic Pipes problem: 100% correct for Physics faculty, 90% correct for all other faculty, but no better than 70% for any student group.

2. Because we scored all questionnaires from the school year 1988-1989 prior to testing the fall 1989 sample of Physics 231 students, we were not blind to the latter group's identity.

3. In addition to testing physics faculty members on the questionnaire, we asked these individuals to complete the questions a second time, predicting the typical errors that undergraduates would make. Despite speculation that professors are insensitive to the misbeliefs, our sample of seven physics faculty members was quite accurate; they failed to mention only 3 of the 19 distinguishable error patterns that we observed across the five problems.

4. Although Halloun and Hestenes reported group means and *SDs* on pre- and posttest accuracy, for groups ranging in size from 70 to 196, apparently no statistical tests were conducted to evaluate the gains. It is quite likely that the 14% gain would have been statistically significant with such large groups. Nonetheless, they argued that "the posttest scores . . . are unacceptably low considering the elementary nature of the test. . . . We think that one should not be satisfied with any instruction which fails to bring all students who pass the course above the 75% level. Conventional instruction is far from meeting this standard" (1985b, p. 1048). For several reasons, among them our earlier remarks about absolute levels of performance, statistical significance, and differences of opinion about the "elementary nature of the test," this argument concerning lack of instructional effectiveness is less than persuasive.

5. As an example, one miscellaneous problem showed a second ball, B, being dropped at the same time as Ball A rolled off a horizontal surface; subjects were asked where B would be, relative to A's position at a point prior to impact. Because the question involved both a passively and an actively moving object, it was classified in the miscellaneous category.

APPENDIX
Motion Problems and Scoring Criteria

1. *Spiral*. You are looking down on a curved metal tube resting on a flat surface. A marble is placed inside the tube at the point indicated by the arrow, and is shot through the tube at high speed. Draw the path of the marble upon emerging from the other end of the tube.

Scoring: A pathway exiting the tube in a straight line (technically, a straight line tangent to the exit point of the tube) was scored as correct. A total of 12% of the subjects apparently misunderstood the problem, and drew the pathway *in* the tube rather than "emerging from" the tube; these were scored as omitted.

2. *Softball*. Suppose that you have a softball attached to a string and you are circling it at high speeds above your head. In the illustrated diagram, you are looking down on the softball. The circle shows the path the softball is following, and the arrows indicate the direction of motion. The string is the line from the center of the circle to the softball. If the string disconnects from the softball at the point indicated in the diagram, draw the path of the softball, ignoring all air resistance.

Scoring: A straight pathway, drawn at an angle reasonably close to the tangent of the circle, was scored as correct. Curved pathways and pathways clearly deviating from the tangent were incorrect.

3. *Bomb*. In this diagram, an airplane is flying at a constant speed and altitude, with a flight path parallel to the ground. An arrow shows the direction of the flight path. When the airplane is in the position shown, a bomb is going to disconnect from the airplane and fall to the ground. Keeping in mind that the airplane is moving at a constant speed and altitude, draw the path of the bomb as it falls to the ground.

Scoring: An arc-shaped trajectory (technically, parabolic), maintaining forward motion while gaining vertical motion, was correct. Any pathway that either completely lost or never showed forward motion was scored as incorrect, as were all straight pathways.

4. *Cliff*. In this diagram, a ball is moving at a constant speed of 55 mph in the indicated direction. Draw the path of the ball after it crosses the edge of the cliff. Ignore all friction and air resistance.

Scoring: Same as for Bomb.

5. *Pendulum*. In this diagram, a pendulum swings through an arc. The pendulum breaks at the point shown in the diagram. Draw the path of the ball as it falls to the ground.

Scoring: Same as for Bomb, except that any pattern demonstrating an initial *increase* in height was also incorrect, regardless of the subsequent pathway. Some subjects interpreted the vertical line as a wall, and showed the ball bouncing off the wall; these were scored as omitted, since the pathway before hitting the wall was too short to determine its shape.

6. *Pipes*. Pipe 1 can fill the tank in 2 h by itself. Pipe 2 can fill the tank in 10 h by itself. How long will it take to fill the tank with both pipes running?

Scoring: The correct answer is 1 h 40 min. Any duration deviating 10 min or less was scored as correct, since our intention was to detect responses showing the common averaging error (Reed, 1984; i.e., 10 h plus 2 h equals 12 h, divided by 2 equals 6 h). Responses stating "a little less than 2 h" were counted as correct; those merely stating "less than 2 h" were scored as incorrect. Relatively few subjects showed formulas or computations, so we made no attempt to score the solution methods.

(Manuscript received April 15, 1991;
revision accepted for publication December 31, 1991.)