Standardized Testing in Physics via the World Wide Web

by

Dr. Dan MacIsaac
SUNY—Buffalo State College

Dr. Rebecca Pollard Cole
Northern Arizona University

Dr. David M. Cole
Northern Arizona University

Dr. Laura McCullough
University of Wisconsin-Stout

and

Dr. Jim Maxka,
Northern Arizona University

Introduction

A Comparison of Paper-based and Web-based Assessment

Since the late 1970's, science educators have been experimenting with the use of microcomputers for the conceptual and attitudinal assessment of their students (Arons, 1984, 1986; Bork, 1981; Waugh, 1985). Since the late 1980's, multiple-choice, machine scored, standardized instruments have been developed to assess the conceptual and attitudinal state of introductory physics students. The Force Concept Inventory (FCI), perhaps the best known of these standardized instruments, assesses student's conceptual knowledge of physics (see Hestenes, Wells & Swackhamer, 1992). Recently, Redish, Saul, and Steinberg (1998) developed the Maryland Physics Expectations Survey (MPEX), a standardized instrument which assesses the attitudinal state of physics students. Both the FCI and the MPEX are widely used in the Physics Education Research (PER) community (Hake, 1998).

Although these instruments were initially used by experts for research only, more generalized interests in program evaluation, curriculum development, justifying and guiding interventions in physics teaching practices and comparing student learning and attitudinal outcomes have led to widespread desires to use these instruments. Anticipating this interest, the FCI was published with an author's statement that "[the FCI] is included here for teachers to use in any way they see fit" (Hestenes, Wells & Swackhamer, 1992. p142). As an example of such use, the FCI was recently adopted as one of a suite of instruments to be used for the regular and routine assessment of student learning in the physics course sequences at Northern Arizona University (MacIsaac, 1999).

There are administrative burdens associated with standard use of these instruments. For instance, completion of one of these instruments requires approximately thirty minutes of class, laboratory or recitation time. Since these instruments are typically administered both pre- and post-instruction, each instrument could therefore consume up to an hour of scarce and valuable instructional time. In addition, resources required to duplicate, administer, collect, collate, accurately code, score, record, and analyze the instrument data are sharply limited in many departments, strongly discouraging regular and routine paper-based administration of these instruments. Hake (1998) confirms that both the loss of instructional time and the administrative overhead may discourage the regular use of these instruments by many introductory physics instructors. Hence our interest in alternative, non-classroom administration of these instruments at NAU.

Web-based technologies provide students with an alternative to paper administration -- the opportunity to complete assessment instruments from personal computers via internet access (Titus, Martin & Beichner, 1998). Harvey and Mogey (1999) suggest economies of time, scale and student effort are possible by amortizing development of web coding infrastructure over many semesters, eliminating the need for expensive optical scan forms, reusing instrument data for multiple reasons and establishing uniform assessment administrations for future, continuing student use in following courses. Danson (1999) suggests further advantages to web testing such as improved response accuracy by reducing input response errors such as skipped rows of optically marked bubbles and assuring statistical software interpretability by input checking and appropriately constrained input selection. Cann & Pawley (1999) note that web pages can reduce coding errors and write student-provided data directly to computer files that can themselves be used as input files for computerized statistical analysis, removing any further need to code data for computer input. Web-based administration of standardized instruments can even allow simultaneous collection of new kinds of data for improving the instruments themselves (such as question latency data -- the length of time required for responses).

Security is another issue: web-administered instruments appear to trade security for flexibility (Harvey & Mogey, 1999). Authentication (verifying the identity of the person completing an instrument) is difficult or impossible to ensure outside of a monitored computer laboratory. Web test takers may be inappropriately collaborating with others, sharing questions with others, cheating or using reference materials.

Some student may also develop increased anxiety (Brosnan, 1999) associated with computer use that could lead to distorted data. Finally, all students may not have ready and appropriate access to computers and the web necessary to complete web administered instruments (Harvey & Mogey, 1999), which may become less of an issue for physics students as time progresses.

However, to be commensurate with the current collection of paper-administered FCI data, the equivalence or mapping for web-administered version of standardized physics instruments must be developed. As described by Brosnan (1999):

> The American Psychological Association's (1996) *Guidelines for Computer-based tests and interpretations* calls for equivalence to be established between the computerized and original versions of the assessments. This necessitates comparisons of means, distributions, ranking of scores and correlations with other variables. Tseng *et al* (1998) argue that for equivalence to be truly established, individual characteristics should not differentially affect a person's responses to a particular administration mode of an assessment. (p. 49)

To be widely used, the web-based administration of these instruments must be characterized in terms of reliability, and results from the web-based administration of these instruments must be statistically compared to results from standard paper administration. If measurements from web-based administrations are explored, they can be corrected or calibrated to paper-based administrations. Therefore, the purpose of this study is to begin this process by examining the differences in paper-based and web-based administrations of the Force Concept Inventory.

Method

Participants

The participants in the study were students from three introductory physics courses taught at a medium sized university in the southwest during the Spring of 1998 and the Fall of 1999. The first two courses, General College Physics I (Physics 111) and General College Physics II (Physics 112) comprise the two semester algebra-based sequence for non-science majors. Students in these two courses were mostly pre-health professions, biology and education majors. The third course, University

Physics I (Physics 161) is part of the three semester calculus-based sequence for science majors. Students in this course were mostly science (e.g. physics, chemistry) and engineering majors.

The participants made up a sample of 376 students, 235 (62.5%) women and 141 (37.5%) men. As the majority of the students were Caucasian, in the age range of 18 to 22, age and ethnicity were not considered further.

Instruments

The Force Concept Inventory (FCI) is a 30 item multiple choice test which "requires a forced choice between Newtonian concepts and common-sense alternatives" (Hestenes, Wells, & Swackhamer, 1992, p. 142). The concepts tested include kinematics, Newton's First, Second and Third Laws, the superposition principle and forces. Student data from the FCI and related instruments have now been collected and published on thousands of students (Hake, 1998). The Maryland Physics Expectations Survey (MPEX) is a 34 item Likert instrument with 5 attitudinal subscales (Redish, Saul, and Steinberg, 1998) which was used as a filler task and is not analyzed further in this study.

Procedure

This study used a quasi-random, quasi-experimental design. During the Spring of 1998, one section of Physics 112 and one section of Physics 161 participated in the study. During the Fall of 1998, one section each of Physics 111, Physics 112, and Physics 161 participated. In total, 5 sections of three different courses participated. For simplicity, these will be referred to as classes. Each class section was divided into two equal (within one student) half-class groups by selecting every second name in alphabetical order from the roster. During the first week of each semester, thirty minutes was devoted to testing. In each class, one half-class group completed a paper-based FCI and was then asked to complete the web-based MPEX in the next seven days. The other half-class group completed a paper-based MPEX and was then asked to complete the web-based FCI in the next seven days.

Note the FCI is being applied as a pretest only in this study; no attempt was made to determine gains via comparisons of student pre-instruction (pretest) and post-instruction (posttest) scores.

Each student was supplied with the web address for the test appropriate to their assigned half-class group. No training was provided to the students for taking either the FCI or the MPEX on the web. Further, there was no attempt to authenticate the web users. Each student's work was accepted as their own. Overall completion times, submission times and dates were recorded. This information was used to ensure that students took no longer than 30 minutes to complete the test and that they took the test within the seven day period. It should be noted that the web-based format allowed students to retake the test after they received on-line feedback regarding their first submission. The date and time information ensured that the test data used as part of the study was their first submission.

All of the tests were graded for completeness and counted as the equivalent of one homework or quiz assignment (E.g. one instructor awarded grades of 0, 1 or 2 of two points. With respect to final class grades, students' participation comprised about 3 points out of one thousand total points, so that completion or non-completion had negligible impact on course grades, although completion of the instruments was rewarded.

Results

As a result of the paper-based and web-based administrations, 376 usable tests were collected. Tests that were turned in after the seven day period, or that were taken for longer than 30 minutes were deemed unusable. Student scores on the FCI were calculated by adding the total number of correct answers with a total possible FCI score being 30. For the entire data set ($\underline{N}$ = 376), the mean of the FCI was $\underline{M}$ = 13.71 ($\underline{SD}$ = 6.08). Table 1 presents the means and standard deviations of the Force Concept Inventory for all sections of all of the introductory physics classes tested.

Table 1

Means and Standard Deviations of FCI student scores in all sections of all physics classes.

	Spring 1998			Fall 1998		
Course	N	Mean	SD	N	Mean	SD
Physics 111	na	na	na	109	9.11	4.19
Physics 112	38	15.37	6.09	38	13.71	4.16
Physics 161	90	18.17	5.64	101	14.09	5.41

The purpose of the study was to examine differences in paper-based and web-based administrations of the Force Concept Inventory. Therefore, several different analyses were conducted. First, total FCI scores were calculated and differences between paper and web were examined. Second, differences in individual items between paper and web were explored. Third, patterns of responses in the individual items were examined to determine if differences existed between paper and web-based administrations. Finally, the predictive validity of the two different FCI administrations on students' course grades was examined. The results of these analyses are reported in the sections which follow.

Paper-based Versus Web-based FCI Student Scores

Data for this study were collected in different sections of 3 different physics courses (see Table 1). Another concern with the FCI is the gender gap on the test in which men tend to perform better than women (McCullough, 2001; Grim, 1999; Dancy, 2000). This gender gap could be affected by on-line administration of the test, especially in light of the differences in how men and women interact with computers. Women tend to use computers only as tools, and often show more anxiety about computer use than men do (Cheek & Agruso, 1995; Shasaani, 1997). Given the gender gap already present, a web-based version of the test may affect how men and women perform on the FCI.

Therefore, to examine differences in paper-based and web-based FCI student scores a 5 X 3 X 2 X 2 ANOVA was used (5 sections, 3 courses, 2 genders, 2 types of FCI administration). An alpha level of .01 was used for all statistical tests. Significant differences were found for the main effects of section, course, and gender. No significant differences were found for the main effect of FCI administration. For the first-order interactions, no significant differences were found due to type of FCI administration. Table 2 presents the results of the ANOVA.

Table 2

Four-Way ANOVA summary table for section, course, gender, and type of FCI administration

Source	df	MSe	F
course	2	1684.72	68.09*
section	2	421.75	17.05*
gender	1	499.79	20.20*
administration	1	29.06	1.17
course x administration	2	26.79	1.08
section x administration	2	41.45	1.68
gender x administration	1	.14	.01

$^*p < .01$

To further examine potential differences in the student scores, Cronbach's alpha was calculated separately for the paper and web administrations. For the entire sample = .86 (N = 376), for the paper-based administration = .86 (N = 212), and for the web-based administration = .85 (N = 164). These alpha levels appear to be comparable.

Paper-Based Versus Web-based Individual FCI Items

Differences in the paper-based and web-based administrations of the FCI for individual items was explored using t Tests. A probability level of .01 was used for all statistical tests. The F statistic was used to determine whether the variances of the paper- and web-based administrations of each item were equal. No significant differences were found for any of the 30 items. Table 3 presents the results of the t Tests. (df = [211.163] for all tests)

Table 3

Results of t Tests for paper-based and web-based administrations of FCI items

Item	F	prob<F	Item	F	prob<F
Item 1	1.29	.08	Item 16	1.00	.98
Item 2	1.08	.60	Item 17	1.04	.79
Item 3	1.02	.91	Item 18	1.12	.45
Item 4	1.05	.71	Item 19	1.07	.66
Item 5	1.04	.77	Item 20	1.08	.62
Item 6	1.05	.75	Item 21	1.04	.80
Item 7	1.13	.41	Item 22	1.01	.96
Item 8	1.03	.86	Item 23	1.00	.98
Item 9	1.01	.98	Item 24	1.00	.98
Item 10	1.04	.81	Item 25	1.06	.67
Item 11	1.12	.45	Item 26	1.01	.93
Item 12	1.02	.90	Item 27	1.00	.98
Item 13	1.04	.80	Item 28	1.10	.53
Item 14	1.03	.83	Item 29	1.07	.66
Item 15	1.20	.22	Item 30	1.06	.70

df = (211, 163) for all tests

Additional analysis was performed to explore potential differences in the administration for individual items. Chi Square tests of the paper- and web-based administrations of each item were conducted to determine whether the response patterns (patterns of A, B, C, D, and E distracters) of the two administrations differed. A probability level of .01 was used for all statistical tests. No significant differences in the response patterns were found for any of the 30 items. Table 4 presents the results of the Chi Square tests.

Table 4

Results of X^2 tests for paper-based and web-based administrations of FCI items

Item	X^2	p	Item	X^2	p
Item 1	8.89	.06	Item 16	1.95	.75
Item 2	3.33	.51	Item 17	1.23	.87
Item 3	2.12	.71	Item 18	4.69	.32
Item 4	11.85	.02	Item 19	2.78	.60
Item 5	.97	.91	Item 20	2.67	.67
Item 6	2.25	.69	Item 21	2.19	.70
Item 7	8.32	.08	Item 22	4.25	.37
Item 8	1.96	.74	Item 23	2.00	.74
Item 9	2.12	.70	Item 24	.91	.92
Item 10	1.14	.89	Item 25	4.99	.29
Item 11	2.40	.66	Item 26	11.47	.02
Item 12	3.27	.51	Item 27	2.55	.64
Item 13	8.12	.09	Item 28	4.62	.33
Item 14	8.08	.04	Item 29	8.10	.09
Item 15	9.42	.05	Item 30	4.60	.33

df = 4 for all tests

Predictive Validity of Paper-Based Versus Web-based FCI Scores

Finally, differences in the predictive validity of the paper-based and web-based administrations of the FCI were explored by examining the correlations between the student's score on the FCI and their subsequent letter grade in the course. Letter grades were changed to their numeric equivalents ("A" was given a value of 4, etc.). For the entire sample r = .26 (N = 376), for the paper-based administration r = .26 (N = 212), and for the web-based administration r = . 28 (N = 164). These correlations, indicating predictive validity, appear to be comparable. Table 5 presents these results.

Table 5

Predictive validity of the paper-based and web-based administrations of the FCI.

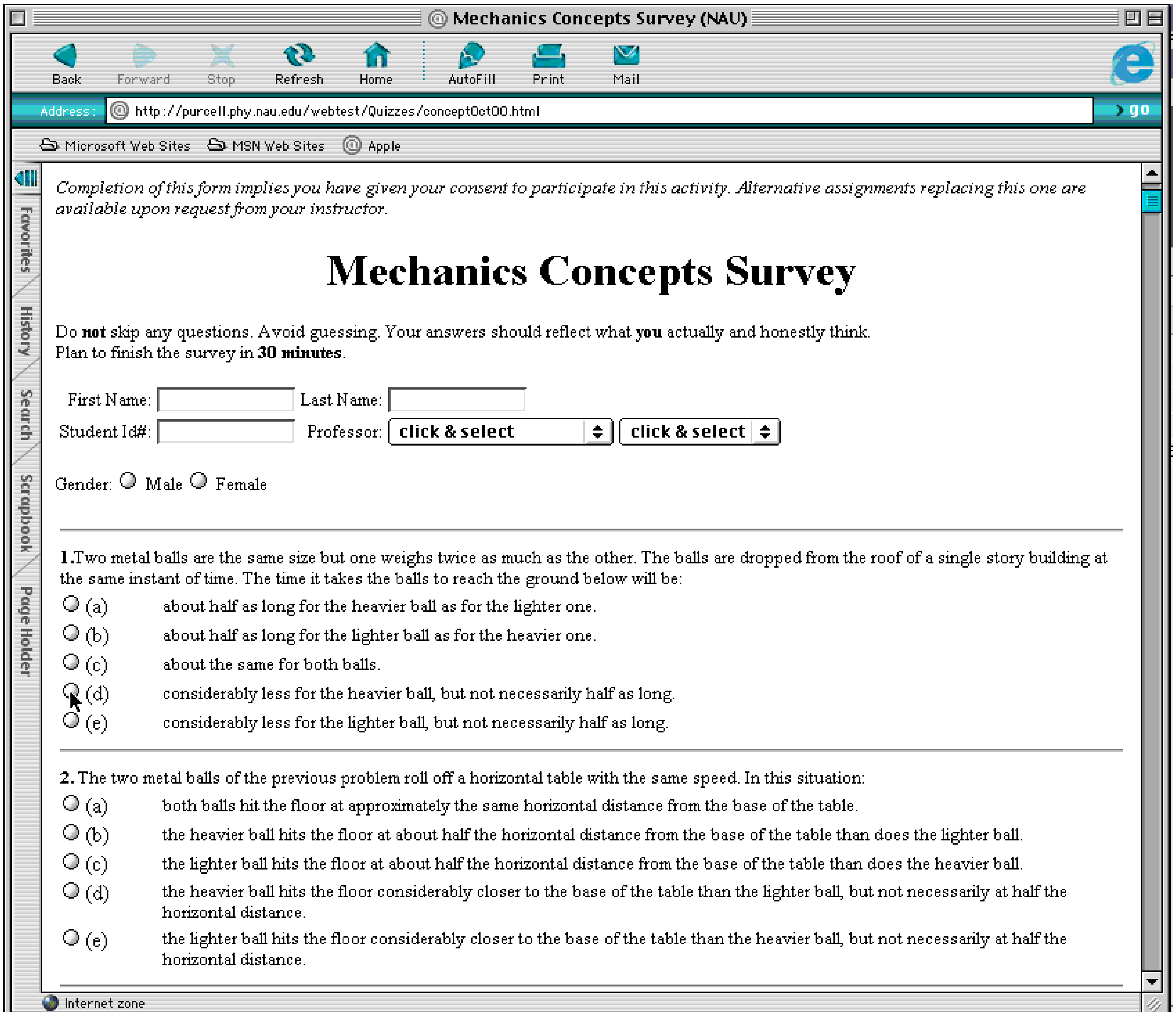
FCI Administration	N	M	SD	r
Paper-based				
FCI Score	212	13.29	6.11	.26
Grade	191	2.69	.96	
Web-based				
FCI Score	164	14.26	6.02	.28
Grade	153	2.69	.85	
Both Administrations				
FCI Score	376	13.71	6.08	.26
Grade	344	2.69	.91	

Summary of Results

This study sought to examine potential differences in paper-based and web-based administrations of the Force Concept Inventory. The results of these analyses demonstrated no appreciable differences on FCI scores or items based on the type of administration. While the results of a 4 way ANOVA did demonstrate differences in FCI student scores due to different sections, courses, and gender*, none of these differences were influenced by the type of test administration*. FCI student scores were comparable with respect to both reliability and predictive validity. For individual FCI items, paper- and web-based comparisons were made by examining potential differences in item means and by examining potential differences in response patterns. *Again, no differences in item means (as demonstrated by t Tests) and no differences in response patterns (as demonstrated by Chi Squares) were found*. In summary, the web-based administration of the Force Concept Inventory appears to be as efficacious as the paper-based administration.

Although this study reports no differences between web and paper-administrations of the FCI, there are a number of issues related to web-administered testing of concern to students, instructors and researchers. The first of these is academic dishonesty. In our study, students were awarded only a small grade (0-2 points maximum from 1000 total for the course) for completing the survey. We wanted to encourage students to participate and to be conscientious in their responses, yet minimize the incentive to cheat. We did not prevent students from copying or printing out the test, nor did we authenticate that the students were who they claimed to be. There is no practical way of doing these things without requiring students to take the test in a proctored computer lab; a solution which has been used at other institutions (e.g. Harvard). In earlier research, we developed the expertise to reduce the likelihood of inappropriate printing or sharing of the instrument by restricting access to the online tests with a changing login and password that was only functional for limited times at the start and end of the semester. Originally, our software reported the number of correct responses for the instrument back to the student; we removed this feedback after having an experience where a student repeatedly submitted the survey while varying answers trying to maximize their score. Now the instrument simply thanks the student upon submission.

Another issue related to web-administered tests is the resolution of the student's computer video monitor. Computer video monitors have a much lower resolution than paper printouts (typically 72 dots per inch vs. 600 dots per inch). In the present study, the paper-administered FCI was a direct printout of the web pages (Fig 1). However, the finer resolution of the laser printer made it easier to read both the text and graphics, particularly the vectors and dotted lines which indicated trajectories. While Clausing and Schmitt (1989, 1990a, 1990b) found that with reasonable diligence, there was no a difference in reading errors between computer video monitors and paper-printed tests, the finer paper resolution may still be more comfortable to work with.



Mechanics Concepts Survey (NAU)

Back Forward Stop Refresh Home AutoFill Print Mail

Address: http://purcell.phy.nau.edu/webtest/Quizzes/conceptOct00.html

Microsoft Web Sites MSN Web Sites Apple

Completion of this form implies you have given your consent to participate in this activity. Alternative assignments replacing this one are available upon request from your instructor.

Mechanics Concepts Survey

Do **not** skip any questions. Avoid guessing. Your answers should reflect what **you** actually and honestly think.
Plan to finish the survey in **30 minutes**.

First Name: Last Name:
Student Id#: Professor: click & select click & select

Gender: Male Female

1. Two metal balls are the same size but one weighs twice as much as the other. The balls are dropped from the roof of a single story building at the same instant of time. The time it takes the balls to reach the ground below will be:

(a) about half as long for the heavier ball as for the lighter one.
(b) about half as long for the lighter ball as for the heavier one.
(c) about the same for both balls.
(d) considerably less for the heavier ball, but not necessarily half as long.
(e) considerably less for the lighter ball, but not necessarily half as long.

2. The two metal balls of the previous problem roll off a horizontal table with the same speed. In this situation:

(a) both balls hit the floor at approximately the same horizontal distance from the base of the table.
(b) the heavier ball hits the floor at about half the horizontal distance from the base of the table than does the lighter ball.
(c) the lighter ball hits the floor at about half the horizontal distance from the base of the table than does the heavier ball.
(d) the heavier ball hits the floor considerably closer to the base of the table than the lighter ball, but not necessarily at half the horizontal distance.
(e) the lighter ball hits the floor considerably closer to the base of the table than the heavier ball, but not necessarily at half the horizontal distance.

Internet zone

Figure 1: The FCI in scrolling format, matched to standard paper instrument.

In addition, it was difficult for students using a smaller computer monitor to see several test questions together with the accompanying diagrams. Conversely, printed pages afford students the opportunity to easily flip back and forth or lay successive pages side by side. For the web-administrations, this can only be accomplished by the unwieldy process of scrolling back and forth. A new version of our software for administering instruments works around this by allowing flipping back-and-forth style access to other items on the instrument while simultaneously collecting latency data by the individual item (see Fig 2). This software was developed as a result of this study, and will be evaluated in the future.

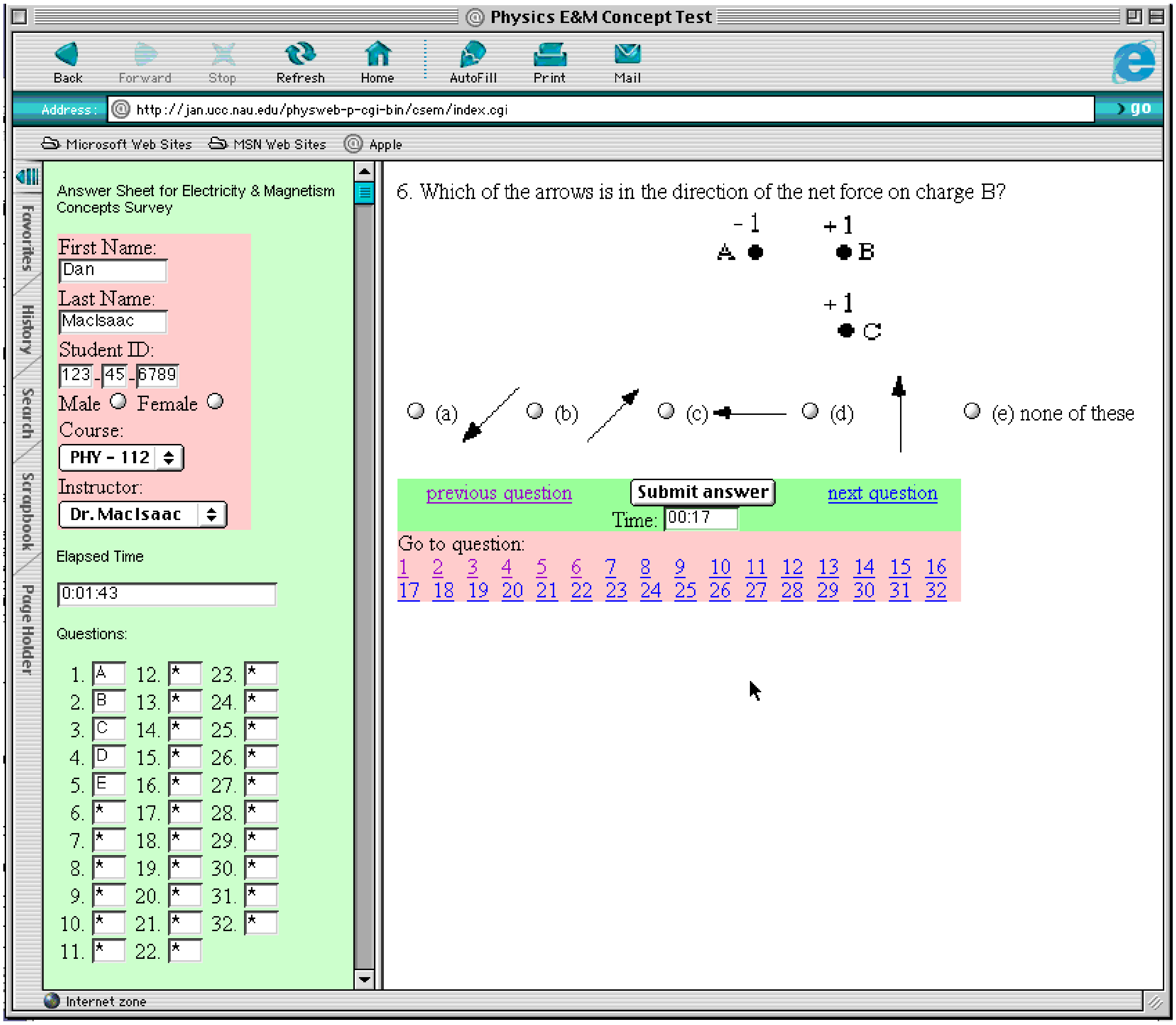


Figure 2: A web-administered standard instrument collecting item latency data.

Finally, the paper-administered FCI coding sheets demonstrated problems. In our study, the optically-encoded scanned bubble sheets produced errors due to skipped rows of questions and incomplete erasures. We eliminated such errors from our data set by rigorously proofreading and screening bubble sheets prior to scanning, and by comparing scanner output files to the original bubble sheets. Such proofing is unlikely to occur with typical paper-administrations, as it poses a significant additional burden on the instructor. Eliminating the use of bubble sheets and allowing students to mark directly on the test might alleviate this problem, but would complicate the grading process. In comparison, the web administered FCI used "radio buttons" for responses. These buttons accurately code only one solution per question, allowed students to cleanly change responses (i.e. no erasing), and aligned each and every response with the question text and graphics on the screen.

Conclusions and Implications

This study demonstrated no differences between the paper-based and web-based pre-instruction administration of a major standardized physics test, the Force Concept Inventory. The main implication of this finding is that, at least for the FCI, web-based administrations could be used in place of paper- administrations, thus saving precious instructional time, reducing the administrative overhead associated with testing, grading, and photocopying thus cutting the costs associated with large scale data collection. Further, web-based administrations offer information that paper-based administrations do not. For example, item latency and completion data can be collected.

We are extending this research in several directions: obviously we must confirm these findings for the case of post-instruction testing and student FCI gains. We are intrigued by and will characterize the discovered FCI gender gap by collecting a larger set on a single class of students for gender-specific analysis by item and distracter. The FCI gender gap seen in our data has been alluded to at research conferences, but is unreported and unexamined in the peer-reviewed literature. We intend to collect and analyze item completion data by electronic vs. paper administration and electronically collect and analyze and latency data by item.

We are investigating the possibility of creating a web-based "Online Introductory Science Testing Center" that could administer tests and feed resulting measurements directly into a modern database. Such a testing center would allow for the routine collection of conceptual and attitudinal data and be available for longitudinal studies of student learning and instruction. This would enhance our understanding of programs and pedagogy both inside and outside our university. Another use of an Online Testing Center would be the opportunity for researchers to pilot and standardize new instruments by providing access to large numbers of student participants. Faculty from other departments have seen our efforts and have started the design and develop of 'screening' instruments intended for student guidance and placement in the gatekeeper science courses at NAU.

Along these lines, the authors have begun to collaborate with other researchers and institutions in an attempt to create such a centralized web-based testing center and common database. In addition, we are expanding our on-line standardized testing effort to include other instruments. Specifically, we are readying the Conceptual Survey in Electricity and Magnetism (Hieggelke, Maloney, O'Kuma, & van Heuvelen, 1996) for web-based administration.

Author Notes

Financial Support:

This research was partially supported by Northern Arizona University Organized Research Grant funds and by the NAU Department of Physics & Astronomy. Additional support was provided by National Science Foundation funding through the Arizona Collaborative for Excellence in the Preparation of Teachers (ACEPT) and the US Department of Education funded AriZona Teacher's Excellence Coalition (AzTEC).

Acknowledgments:

The authors wish to acknowledge the helpful contributions of Nate Davis, Brian Nance and Eric Tse who assisted with HTML and data coding and were funded by an NAU Hooper Undergraduate Research Fellowship. Dr. Jim Maxka developed the new format for the online administration of these standardized instruments. Professors Gary Layton, Gary Kanner, Gary Bowman, William Delinger, Dan MacIsaac and David Cole of the Northern Arizona University Department of Physics and Astronomy participated in and supported this study. Valuable comments and suggestions regarding the statistical comparisons of item response patterns was provided by Professor Philip Sadler of Harvard University.

References

Arons, A.B. (1984). Computer-based instructional dialogs in science courses. Science, 224, 1051.

Arons, A.B. (1986). Overcoming conceptual difficulties in physical science through computer-based Socratic dialogs. In A. Bork & H. Weinstock (Eds.), Designing computer-based learning materials: NATO ASI Series Vol F23. Berlin: Springer-Verlag.

Bork, A. (1981). Learning with computers. Bedford MA: Digital Press.

Brosnan, M. (1999). Computer anxiety in students: Should computer-based assessment be used at all? In Brown, S., Race, P. & Bull, J. (1999) (Eds), Computer-assisted assessment in higher education. London: Kogan-Page.

Cann, A.J. & Pawley, E.L. (1999). Automated online tutorials: New formats for assessment on the WWW. In Brown, S., Race, P. & Bull, J. (1999) (Eds), Computer-assisted assessment in higher education. London: Kogan-Page.

Champagne, A.B., Klopfer, L.E., & Anderson, J.H. (1980). Factors influencing the learning of classical physics. American Journal of Physics, 48, 1074-1079.

Clausing, C.S., & Schmitt, D. R. (1989). The effects of computer usage on computer screen reading rate. Paper presented at the annual meeting of the Mid-South Educational Research Association. Little Rock, AK.

Clausing, C.S., & Schmitt, D. R. (1990a). Paper versus CRT: Are reading rate and comprehension affected? Proceedings of Selected Paper Presentations at the Convention of the Association for Educational Communications and Technology.

Clausing, C.S., & Schmitt, D. R. (1990b). Does computer screen color affect reading rate? Paper presented at the annual meeting of the Mid-South Educational Research Association. New Orleans, LA.

Danson, M. (1999). Optimal mark reading as CAA: Experiences at Loughborough University. In Brown, S., Race, P. & Bull, J. (1999) (Eds), Computer-assisted assessment in higher education. London: Kogan-Page.

Grim, N. (1999). A force concept correlation study with instructional methods, anxiety, perceptions of difficulty and student background variables. Chicago, IL: Chicago State University. ERIC document ED 438 164.

Hake, R.R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics, 64-74.

Hallouin, H. (1998). Views about science and physics achievement: The VASS story. Proceedings of the Third International Conference on Undergraduate Physics Education. College Park, MD.

Hallouin, H., & Hestenes, D. (1996). The initial knowledge state of college physics students. American Journal of Physics, 53, 1043-1055.

Harvey, J. & Mogey, N. (1999). Pragmatic issues when integrating technology into the assessment of students. In Brown, S., Race, P. & Bull, J. (1999) (Eds), Computer-assisted assessment in higher education. London: Kogan-Page.

Hestenes D., & Hallouin, I. (1995). Interpreting the Force Concept Inventory: A response to March 1995 critiques by Huffman and Heller. The Physics Teacher, 33, 502.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. The Physics Teacher, 30, 141-158.

Hieggelke, C.J., Maloney, D., O'Kuma, T., & Van Heuvelen, A. (1996). Electric and magnetic concept inventory. AAPT Announcer, 26, 80.

Huffman D., & Heller, P. (1995). What does the Force Concept Inventory actually measure? The Physics Teacher, 33, 138.

Linn, R.L. (Ed.). (1978) Educational Measurement (3rd ed.). New York: Macmillan.

MacIsaac, D.L. (1999a). Department of physics and astronomy assessment plan and preliminary data report. Unpublished manuscript available from the author.

MacIsaac, D.L. (1999b). Comparing web and paper-based administration of standardized introductory physics conceptual and attitudinal instruments. Final Program and Abstracts: 72nd NARST Annual Meeting 1999, 204.

McCullough, L. (2001, January). A gender context for the Force Concept Inventory. Paper presented at the meeting of the American Association of Physics Teachers, San Diego, CA.

Redish, E.F., Saul, J.M., & Steinberg, R.N. (1998). Students expectations in introductory physics. American Journal of Physics, 6, 212-224.

Shasaani, L. (1997). Gender differences in computer attitudes and use among college students. Journal of Educational Computing Research, 16(1), 37-51.

Straetmans, G., & Eggen, T. (1998). Computerized adaptive testing: What it is and how it works. Educational Technology, JanFeb98, 45.

Titus, A.P., Martin, L.W., & Beichner, R.J. (1998). Web-based testing in physics education: Methods and opportunities. Computers in Physics, 12, 117.

Waugh, M.L. (1985). Effects of microcomputer-administered diagnostic testing on immediate and continuing science achievement and attitudes. Journal of Research in Science Teaching, 23, 793.

Wise, S.L., & Plake, B.S. (1990). Computer-based testing in higher education. Measurement and Evaluation in Counseling and Development, 23, 3.

Zandvliet, D., & Farrager, P. (1997). A comparison of computer-administered and written tests. Journal of Research on Computing in Education, 29, 423.

About the authors…

Dr. Dan MacIsaac is an Assistant Professor of Physics at SUNY—Buffalo State College specializing in teacher preparation whose research examines how non-majors learn physical science and physics. He is active in the American Association of Physics Teachers (AAPT). Correspondence: danmac@att.net

Dr. Rebecca Pollard Cole is an Associate Professor of Educational Psychology at NAU specializing in teacher preparation whose research examines self-regulated learning, undergraduate under-preparedness and academic success. She is active in the American Education Research Association.

Dr. David M. Cole is a Lecturer of Physics, and the Assistant Director of the NAU/NASA Spacegrant Program. He is the AAPT Arizona section representative.

Dr. Laura McCullough is an Assistant Professor of Physics at the University of Wisconsin-Stout in Menomonie. She researches gender issues in physics instruction, and is active in the AAPT.

Dr. Jim Maxka is a Senior Lecturer in the NAU Department of Chemistry. His research includes web-based assessment of both student readiness and conceptual learning.