JTML, Vol. 3, 15-32

STUDENTS' EXPERIENCES WITH INTERNET BASED TESTING IN AN INTRODUCTORY GENERAL MUSIC COURSE

Thomas Smialek
Lisa Swenson
The Pennsylvania State University, Hazleton

When the principal investigator's university shortened fall semester by one week, he implemented Internet based testing in his nonmajor Introduction to Western Music course as an alternative to expending already-limited class time on student evaluation. Students took two of the course's four required tests online in a secure, proctored environment at the campus Teaching and Learning Resource Center. This paper reports results of a two-year study of university general music students' experiences with taking online tests. At the end of fall semesters 2002 and 2003, 95 volunteer participants completed a 19-item survey. Students rated the online test registration system, practice test, test-taking environment, testing software, scoring/feedback, and preference for online versus in-class testing. Also included was the opportunity for open-ended responses and suggestions. Results indicated a polarized response to a preference for online over in-class testing by a 2-to-1 margin, with one third neutral. Scheduling and registration issues produced a similarly divergent outcome. Immediate scoring and the opportunity for repeated hearings of listening selections generated favorable comments. When subjects took tests online, they performed significantly worse than students who took tests exclusively in class, possibly due to the format of listening questions or from learning new material during the testing period.

Although computer administered testing has existed for over 30 years, since the early 1990s it has been employed in a variety of settings: state drivers' license tests, professional certification and licensure exams, military training exams, and college placement tests such as the GRE, GMAT, and Test of English as a Foreign Language (Bugbee, 1996; Russo, 2002). With the increasing availability of personal computers and educational software, educators have begun to use computer based testing (CBT) in the classroom (Mason, Patry & Bernstein, 2001). Compared to traditional paper-and-pencil tests, CBT reduces testing time (Bunderson, Inouye & Olsen, 1989) and saves class time required for test administration and review (DeAngelis, 2000). Other benefits of CBT include instant scoring (Bugbee & Bernt, 1990), accurate marking, and improved turnaround time for results to students. Commercial CBT products can also create reports on the performance of individual students or groups, and can track the percentage of individuals who answer a particular question correctly (Stephens, 2001).

Computer based tests can be categorized into two main types: fixed item tests (FIT, also known as non-adaptive) and computer adaptive tests (CAT). Fixed item tests are computerized versions of paper-and-pencil

tests that contain identical assessment items appearing individually in sequential order. "Adaptive CBT differs in that assessment items are selected based on the student's answers to previous items; so each test is tailored to the skills or abilities of the test taker" (Mason, et al. 2001, p. 30). Vispoel & Coffman (1992) found adaptive tests to be

> . . . more efficient than fixed item tests, because examinees respond only to items matched to their ability levels, bypassing those that are either too easy or too difficult. Adaptive tests are more reliable, because examinees receive a greater proportion of items at appropriate difficulty levels. And finally, adaptive tests are potentially more valid, because increases in reliability and reductions of boredom, fatigue, and guessing effects are likely to increase test validity. (pp. 30-31)

Vispoel & Coffman (1992) created a CAT that required 71% fewer items to produce scores with greater reliability and validity than those in the Wing Tonal Memory test (Wing, 1961). In three subsequent studies that compared CATs and FITs of tonal memory, adaptive tests needed 50% to 93% fewer items to match the reliability and concurrent validity of four standardized FITs of musical aptitude (Vispoel, Wang & Bleiler, 1997).

Despite their increased efficiency and validity, CATs have several logistical drawbacks: (a) They require a high level of measurement expertise to design, implement, and evaluate; (b) collection of large examinee calibration samples and construction of adequate item banks is difficult; (c) test development is expensive; and (d) commercially available CAT software could not produce music listening items (Vispoel, et al. 1997). Mason, et al. (2001) also felt it was not currently feasible to employ CAT for small scale classroom testing.

A number of researchers have investigated the equivalence and effectiveness of CBT compared to paper-and-pencil tests, and the results have been mixed. Half of the studies reviewed by Mazzeo and Harvey (1988) yielded higher scores on CBT and half favored traditional testing. Bunderson, et al. (1989) reported three studies where computer based test scores were higher, 13 that showed higher scores on paper-and pencil tests, and 11 that proved no difference in scores. In a meta-analysis of computer versus paper based cognitive ability tests, Mead and Drasgow (1993) found that paper based test scores were slightly greater.

Several recent studies of equivalence explored the effects of computer based test interfaces and effects of learner characteristics in nonadaptive tests. Mason, et al. (2001) cited numerous explanations for lower CBT scores reported by some researchers. Score distributions might be influenced by faster completion times for computer based tests. Earlier CBTs often did not allow students to skip items, to change answers, or to review previously answered items (Wise & Plake, 1990). Mason, et al. found that when their testing program incorporated these capabilities, students in an introductory psychology course were able to navigate effectively through

required computer based unit tests and obtained equivalent scores to paper-and-pencil versions.

DeAngelis (2000) found that senior dental hygiene students performed as well or better on two CBT exams than students using paper versions. Students taking the computer based exams could move backward and forward between test items and could review all responses before submission. Digitized images of clinical photographs, radiographs, and other illustrations allowed students taking the CBT exam to progress at their own speed, whereas students taking the paper test had to wait until all had satisfactory time to view each projected image or overhead transparency.

Clariana and Wallace (2002) investigated the "test mode effect" of CBT versus paper based tests, considering individual learner characteristics of prior content familiarity, computer familiarity, competitiveness, and gender. Of these four characteristics, only content familiarity was related to the test mode effect. "Specifically, computer based tests especially helped the high-attaining students (relative to paper based testing)" (p. 598).

In the literature on computer based testing, surveys of students' attitudes toward this form of assessment generally were positive. Bugbee and Bernt (1990) reported that from 1982-1988, only 4% of students in a distance education institution devoted to financial services education disliked taking computerized exams. The problems identified by Bugbee and Bernt revolved around the 1980s era computer equipment and performance. System failure was also a serious concern. Benefits of CBT, such as immediate scoring and scheduling, received enthusiastic response from students at The American College; however, students disliked traveling long distances to the remote testing centers that were employed before the widespread use of the Internet. Vispoel and Coffman (1992) found that examinees overwhelmingly preferred the computer-adaptive test of tonal memory to a paper and pencil version.

In more recent studies, DeAngelis (2000) found student acceptance of a fixed item CBT to be mixed due to their limited exposure to this format. This study presented the first experience with CBT for over three-fourths of the participants. Students liked features such as immediate feedback and scoring, identification of wrong responses, not having to write as much, and self pacing. Anxiety over the combination of computers and testing caused the most distress. In assessing British students enrolled in a first year Information Science module, Stephens (2001) emphasized the importance of using pretests before actual computerized exams to familiarize students with the CBT software interface. He also suggested explaining the benefits and advantages of CBT to students to gain their acceptance and cooperation. Since students expressed fear of a possible computer or network crash during testing, Stephens recommended addressing the subject beforehand and describing contingency procedures.

Alexander, Truell and Bartlett (2002) surveyed students enrolled in a business information technology course for their perceptions of CBTs taken in a computer lab over the Internet. Although students' perceptions were

generally favorable on 34 survey items, this result may have been attributable to the course's emphasis on computer use. No significant differences existed based on age group, gender, or grade point average. Freshmen reported significantly higher perception levels of online testing than upperclassmen. However, the means for each year were within the "agree" range of perceptions. Students identified a drawback of online testing as the absence of an available instructor who could answer their questions during exams.

Although the literature on CBT indicates that adaptive tests provide greater reliability and validity while requiring fewer items, recent studies promote fixed item CBTs as a more practical alternative for everyday classroom use. The primary concern with FITs involves equivalence with paper based versions. Findings of earlier studies ranged from no difference to those where one form or the other produced superior results. Recent studies have established that the current generation of CBTs possesses the flexibility of navigation necessary to yield results similar to paper based tests. For the most part, students have responded favorably to computer based testing. Their concerns have included the possibilities of system failures and their inability to ask for clarification during online exams.

As an alternative to expending already-limited class time on student evaluation, the principal investigator began using Internet based testing in his nonmajor Introduction to Western Music course (MUSIC 5) in the fall of 2002. During the previous three years, cooperative learning was adopted for use in the course (Smialek, 2000). Although the use of group listening exercises proved to be pedagogically effective, it came at a cost: Four class meetings' worth of existing course content had to be eliminated from the syllabus in order to accommodate group work. The potential for further dilution of the syllabus arose when the university decided to shorten fall semester by one week. In addressing the implementation of the shortened semester, Pennsylvania State University President Graham Spanier (2002) remarked that there were many course activities that students could pursue just as effectively outside of class: "The [University] calendar should be forward-looking, toward new methods of course delivery and teaching, including greater use of online techniques and new approaches to partial or intermittent residency." The solution for restoring some of the lost content in the Western Music course was to have students take two of the course's four required tests online, outside of class.

After the initial investment of time needed to create, install, and pilot test online exams, computer based testing can offer music faculty many benefits of convenience. CBT would seem to be especially well suited for use in the large sections of introductory nonmajor courses that are commonplace in so many university music schools. Aside from Vispoel and Coffman's (1992) findings of students' preference for computer-adaptive tests, however, no research studies of fixed item CBTs could be found for music. Non-adaptive CBTs used in other academic disciplines (Alexander, et al., 2002; Clariana & Wallace, 2002; DeAngelis, 2000; Mason, et al., 2001)

have tested student knowledge of course content through multiple choice or short answer formats. In addition to this type of "fact-oriented" assessment, conventionally administered music appreciation tests typically contain some sort of listening component to evaluate student ability to perceive various musical elements or compositional styles. The Western Music course listening tests evaluate critical listening and thinking as well. They require students to apply their factual knowledge of musical style by using the elements of music they hear to determine a composition's style period, genre, and composer (see Method: Materials below).

The results of a two-semester study of university general music students' experiences with taking online tests in a secure, proctored environment at the Penn State, Hazleton Teaching and Learning Resource Center (TLRC) are presented in this report. Alexander, et al. (2002) found mainly positive perceptions by business students who took fixed item CBTs online in a computer lab. Since the principal investigator's tests were administered in a similar fashion, we were interested in exploring the reactions of a more heterogeneous population of students. Since Introduction to Western Music is a general education, or "core," course, it is taken by students from a wide variety of majors. This also gave us the opportunity to work with a subject population that was more likely to have diverse experience and competency with computers, compared to business majors in an information technology course. Our survey sought to answer the following research questions:

1. Do nonmajor music appreciation students prefer online computer based tests to those administered on paper in class?

2. What are music appreciation students' perceptions of various aspects of online testing, such as test registration, test interface and instructions, testing environment, pacing, scoring, and feedback?

3. What are music appreciation students' best-liked aspects of Internet based testing?

4. What do music appreciation students like least about Internet based testing?

5. Do music appreciation students perceive online CBTs as equivalent to paper based tests?

Method

Subjects

The subject pool consisted of 202 undergraduate students (50% Freshmen, 41% Sophomores, 4% Juniors, 1% Seniors, and 4% Provisional/non-degree) who took an introductory Western music course at a branch campus of a large research university to fulfill a General Education requirement in the arts. Of the 104 students enrolled in MUSIC 5 during fall 2002 and

2003, 95 subjects were recruited to participate in an anonymous survey of their experiences with Internet based testing. Participation in the survey was voluntary. For their participation, subjects received two extra credit points on MUSIC 5 Test 4. Students who did not participate in the survey had the option of answering two extra credit questions online at the TLRC, each worth one additional percentage point on their Test 4 scores. The scores of 98 students who took MUSIC 5 tests exclusively in class during the 2001-2002 academic year were used for comparison to the performance of the 104 students who took two tests per semester online at the TLRC in fall 2002 and 2003.

Materials

MUSIC 5 tests, whether administered in class or given online in the TLRC, employed a multiple choice format and were in two parts. The first section included questions about key terms and concepts covered in the previous unit of study. The second part of each exam tested student perception and critical thinking skills on a series of 5-6 brief audio selections.

The format of in-class listening tests, administered on paper during the regularly scheduled class period, was based on Bennett Reimer's Style Perception Charts (1985, pp. 253-264). For each listening selection, students had to identify several musical elements (tone color, articulation, meter, texture) to use as data in determining the compositional genre, historical style period, and composer (see Figure 1).

Online tests were constructed in ClearLearning's TestPilot (2002), a web-based application for the creation of online assessments and surveys. The test interface allowed students to move forward or backward through the test at any time via "Previous" and "Next" buttons placed at the bottom of the web browser window (see Figure 2b). Examinees could also skip or review items, or change previous answers at any time before submission. Internet based listening tests asked for element identification and genre-period-composer choices through a series of individual questions, with possible answers listed vertically under each question (see Figures 2a and 2b)

At the end of fall semesters 2002 and 2003, 95 volunteer participants completed a 19-item survey (see Table 1). Survey items asked students to rate the online test registration system, practice test, test-taking environment, testing software, scoring/feedback, and preference for online versus in-class testing. Also included was the opportunity for open-ended responses and suggestions.

Procedure

In fall 2002 and 2003, 104 students visited the Teaching and Learning Resource Center on two separate occasions each semester to take MUSIC 5 Test 2 (Middle Ages, Renaissance, Baroque) and Test 3 (Baroque, Classic, Romantic). Tests 1 and 4 (Elements of Music, Twentieth Century) were administered on paper during regularly scheduled class meetings.

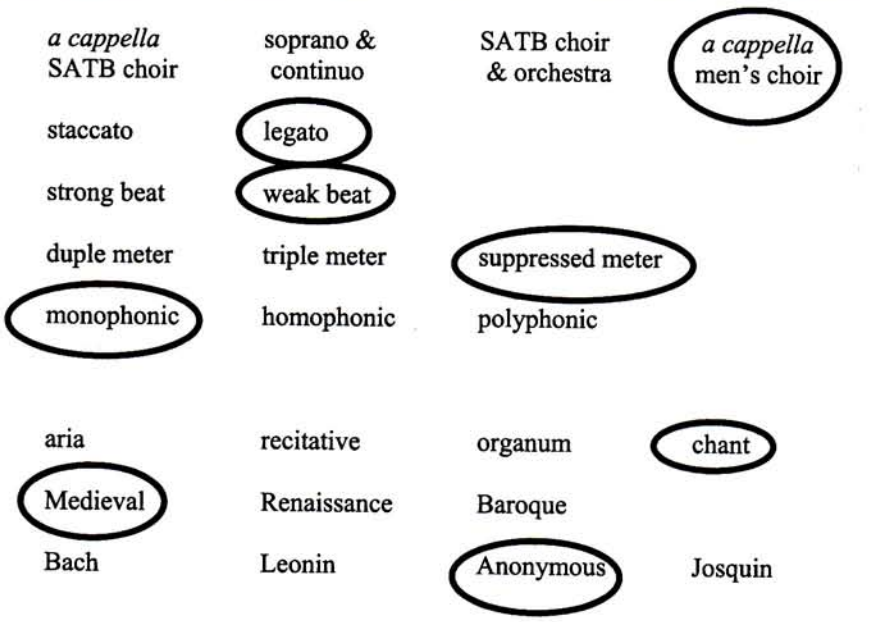

Figure 1. Typical in-class listening test selection (Gregorian chant, with correct answers)

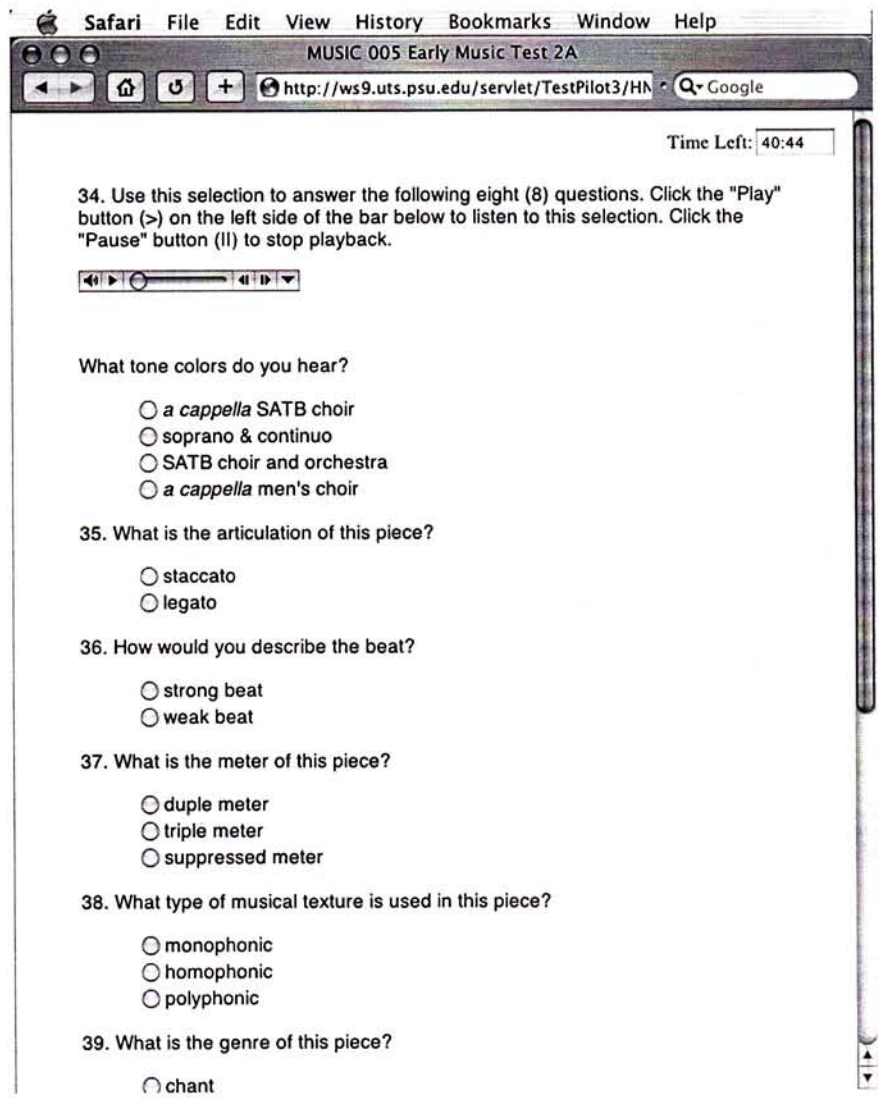

Figure 2a. Online listening test for Gregorian chant selection in TestPilot (top of web browser window).

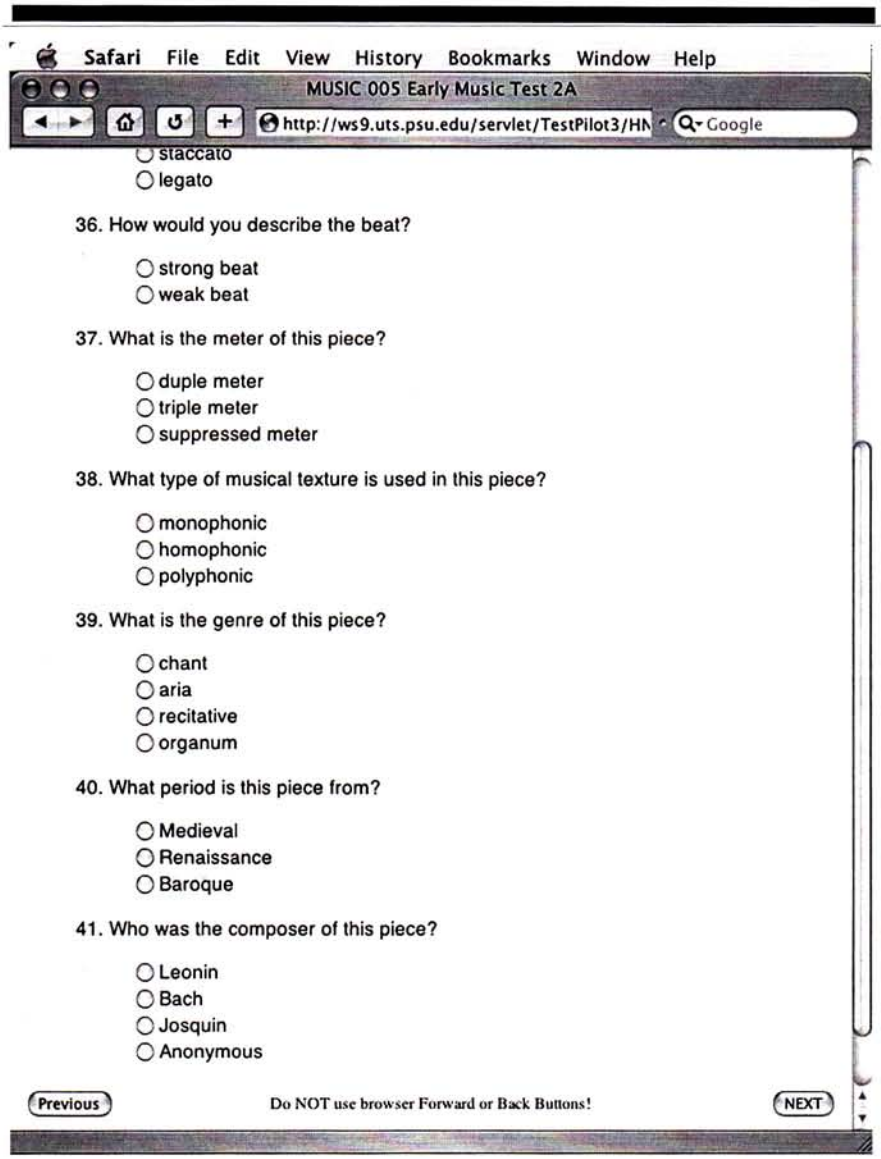

Figure 2b. Listening test for Gregorian chant selection in TestPilot (scrolled to bottom of web browser window).

Four computer stations with Internet access in the TLRC were designated by the University Testing Services as the only locations from which students could access MUSIC 5 online tests. The computers were equipped with high-fidelity stereo headphones for use in listening portions of tests. Students could take each online test over a four-day period by reserving in advance a one-hour time slot during TLRC hours (Monday-Friday, 9 a.m. through 5 p.m.). A TLRC staff member was available during the testing period to make certain that students did not confer with each other while taking a test and that they did not open other documents, web browsers, or e-mail on their computers. To take an online test, students logged on to University Testing Services' Computer Based Testing Page on the Internet. Students had 45 minutes to take each test. Time remaining was indicated by a countdown timer in their browser window.

Before the first test of the semester was given, students could take a practice quiz—at home or in a computer lab—to get familiar with the look and feel of the TestPilot interface. They also had the opportunity to experience the various question formats used in each test.

During the second-to-last class meeting of each semester, survey subjects were recruited immediately following the in-class administration of Test 4. Volunteer participants completed the survey of their experiences with Internet based testing during the 15 minutes that remained in the class period.

Results and Discussion

Preference for Online Versus In-Class Tests

Survey results indicate a polarized response to a preference of online over in-class testing (see Table 1, Item 19). Although moderately to strongly positive responses to Internet based testing outweighed negative ones by nearly a 2 to 1 margin, an overall negative response of 25% indicates substantive dissent. Nearly one-third of the respondents were somewhat neutral toward online testing. Alexander, et al. (2002) also found students' responses to CBT to be generally favorable, but did not report the dispersion of ratings in their results. Computer use was a major component of their business information technology course. Subjects in the present study represented a more heterogeneous group, as Introduction to Western Music fulfills a University general education requirement. Hence, our results would seem more likely to show a mixed acceptance of CBT, as did those of DeAngelis (2000). Three-fourths of her students were new to CBT.

Perceptions of various aspects of online testing

Scheduling and registration issues generated a similarly divergent outcome: 80% of the respondents (moderately-to-strongly) agreed that the online test registration system was easy to use and 73% indicated that they had little to no difficulty registering for a test time. However, 19% had at least some difficulty. In the comments section of the survey, only three students mentioned having actual technical difficulty in registering. The

Table 1

Responses to Online Test survey (N = 95)

	Survey item	strongly disagree 1	2	3	neutral 4	5	6	strongly agree 7	Mean
1.	The online test registration system was easy to use.	0	1	7	1	10	27	49	6.13
2.	The instructions on the test registration web page were clear.	0	1	1	2	12	28	51	6.30
3.	The instructions on the test registration web page were accurate.	1	1	2	2	9	26	53	6.27
4.	I had problems registering for a test time.	44	25	3	3	5	6	9	2.52
5.	The practice test was effective in helping me to master the online test interface.	6	4	7	34	9	19	16	4.65
6.	The online test interface was easy to learn.	0	2	3	11	11	28	40	5.90
7.	The test registration handout (or e-mail) was clear.	0	0	1	7	10	29	48	6.22
8.	The Teaching and Learning Center is a convenient place to take tests online.	5	2	14	11	11	29	23	5.11
9.	The Teaching and Learning Center provides an atmosphere suitable for online test taking.	7	12	15	9	15	16	21	4.52
10.	The online test instructions were adequate.	0	0	1	5	7	40	41	6.22
11.	I had a hard time navigating through the online test.	54	24	5	4	3	2	3	1.91
12.	The computers loaded each test page and audio file quickly.	1	1	0	3	13	35	41	6.14
13.	Audio quality of listening examples is satisfactory.	0	1	1	3	9	40	41	6.20
14.	I went back over my answers and checked them before I submitted my test for grading.	8	5	2	5	12	16	47	5.57
15.	I feel that 45 minutes was an adequate amount of time to take my online test.	0	0	1	3	6	24	61	6.48
16.	I like getting immediate feedback on my test score.	1	2	0	2	3	16	71	6.54
17.	After the testing period concluded, I reviewed my test for right/wrong answers.	25	10	7	19	9	8	17	3.73
18.	I feel that taking a test online negatively affected my score.	31	18	6	9	14	7	10	3.19
19.	I prefer taking MUSIC 5 tests online instead of in class.	15	9	3	20	7	14	27	4.53

most likely cause of difficulty in registration concerned students who were slow to register and did not get a preferred time slot. On the "least-liked" comments section of the survey, 12% complained that too few registration slots were available. (Typically, 80 spots were available over a four-day period for 60 students.) Four students said they resented taking a test outside of regular class meeting times. On the "best-liked" comments, however, 21% of students said that they *liked* the ability to schedule the online test at a time convenient to them. Convenience of scheduling online tests resulted in a tie for the third highest number of responses to the "best-liked" comments. Alexander, et al. (2002) similarly found ease of scheduling and increased study time to be highly rated aspects of online testing.

Best-liked aspects of Internet based testing

Items that generated the most favorable responses or comments included immediate scoring of tests (over 92% favorable response on the survey and the leading item on students' "best-liked" comments, mentioned by 35% of students), the opportunity for repeated hearings of listening selections (mentioned by 30% of students as "best-liked"), and the opportunity to work at one's own pace (mentioned by 21%). These findings concur with those of Bugbee and Bernt (1990), DeAngelis (2000), and Stephens (2001).

Least-liked aspects of Internet based testing

The strongest negative response concerned the Teaching and Learning Resource Center. Although 55% of students found the Learning Center to have a suitable atmosphere for testing, 36% disagreed. Thirty students cited excessive background noise as distracting or annoying when taking online tests. (In Alexander, et al.'s 2002 study, the item "Proctors were talking during the test" received a mean score of 3.69 on a five-point Likert scale.) Before fall 2003, this issue was discussed with the Director of the Learning Center. We agreed that the Center, whose mission primarily involves tutoring and supporting the writing lab, needed to maintain its "relaxed" atmosphere in order to attract and retain clients. The staff did pledge to keep noise levels lower during testing periods and complaints declined during the second semester of the online testing's implementation. (A new Learning Center will open at Pennsylvania State University, Hazleton, in fall semester 2005, featuring a dedicated facility for proctored online testing.)

Another negative response to online testing involved the opinion voiced by one-third of the respondents: Taking a test online somehow negatively affected their score. Regarding "least-liked" comments, only four students mentioned that online tests were more difficult. Thirteen students observed that they just felt more comfortable with in-class testing. Despite the generally positive perceptions of online testing, Alexander, et al. (2002) reported a mean of 3.48, on a five-point scale, to the statement that students felt they would have done better on a paper-and-pencil test. Vispoel and Coffman (1992) stated that, despite a clear preference for computer adaptive tests, students do not necessarily believe that one testing mode is inherently more

reliable or valid than the other. The literature on the equivalence of CBTs and paper tests has shown mixed results, so it is not surprising that student perceptions of the validity of these two modes of test administration is at odds as well. To follow up on these concerns, we supplemented our investigation of students' attitudes toward Internet based testing with an examination of their test performance.

Equivalence of online CBTs to paper based tests

When comparing the average score for each test between the class sections that took two exams online (online group, $N = 104$) and those sections that took exams exclusively in class the year prior to the present study (in-class group, $N = 98$), it appears that concerns about the possible negative effects of online testing have some validity. The group that took Tests 2 and 3 of the semester online performed significantly better on Test 1 than did the group that took tests exclusively in the classroom (see Table 2). Similarly, the online group performed significantly better than the in-class group on Test 4 (not including the two extra credit points given as compensation to study participants in the online group). However, when the online group took Tests 2 and 3 in the Learning Center, it performed worse than the in-class group. The differences in the scores between the two groups for Tests 2 and 3 were not significantly different, but as Figure 3 shows, it was the pattern of change from test to test that was of interest. Both groups performed worse on Test 2 than they had on Test 1. There was a significant difference, however, in how much their scores changed, $F(1, 198) = 8.82$, $p < .01$, with the online group's scores dropping by 7.63 points on average ($SD = 9.82$) as compared to the in-class group's scores dropping by only 3.16 points on average. Another significant difference occurred in the pattern of change in test scores when examining the change from Test 3 to Test 4, $F(1, 195) = 11.62$, $p < .01$. In this case, the online group increased its scores by an average of 6.82 points ($SD = 10.86$), whereas the in-class group increased its scores by an average of only 1.43 points ($SD = 11.33$).

One possible explanation for a dip in performance when taking online tests, aside from anxiety, could be that students tended to work too fast (Mason, et al., 2001). Although students felt they had adequate time (the second-highest rated item on the survey, $M = 6.48$), 13% of students commented that they liked how quickly online tests could be taken. Alexander, et al. (2002) found similar high ratings for both adequate time and speed. One of the students confided that he took online tests in as little as eight minutes (15-20 minutes was more the norm). However, 79% of the students said they checked over their answers before submitting their tests.

A more likely cause for a decrease in online test scores involves the layout of the listening portion of these tests. On the paper exam (see Figure 1), students could view the choices of various elements they perceived all at once, in chart form, *before* selecting the piece's genre, style period, and composer. When creating the listening portion of an online test with TestPilot, answer choices must be displayed as vertical lists, requiring students to

Table 2

Comparison of Average Test Scores

	Online group	In-class group
Test 1		
M	87.00[a]	84.12[a]
SD	9.04	9.86
Test 2		
M	79.47	81.10
SD	11.88	12.30
Test 3		
M	75.57	77.91
SD	13.29	13.30
Test 4		
M	82.32[a]	79.08[a]
SD	10.15	12.65

[a]Significant difference between groups in the mean test score ($p < .05$)

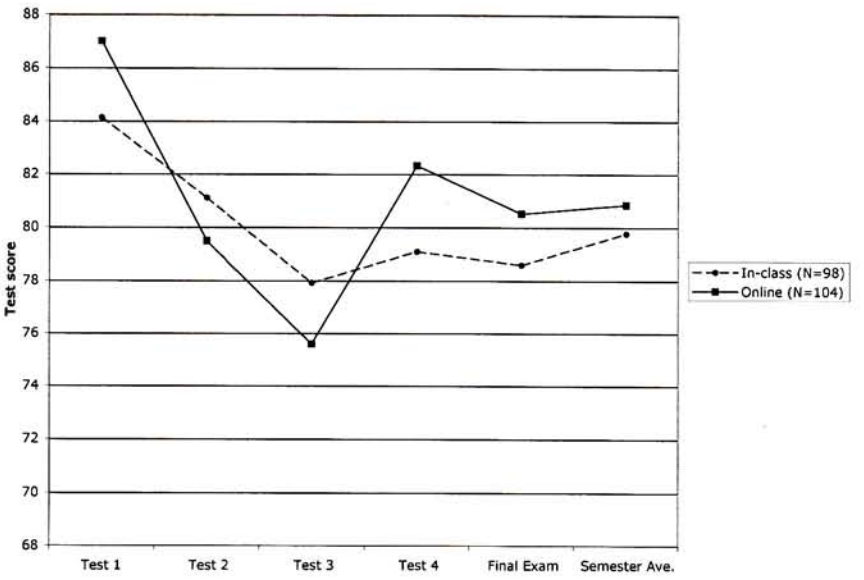

Figure 3. Online versus in-class groups' test scores.

scroll up and down the test window when comparing their answers (see Figures 2a and 2b). If students who are taking the online tests looked at each listening question as a discrete, unconnected item, they may not have been as likely to consider how their musical element choices could form a pattern that led to the correct genre, period, and composer. Mason, et al. (2001) warned that instructors must account for computer screen capacity and question response forms in CBTs, as these factors may influence equivalence with paper based tests. Stephens (2001) also mentioned students' concerns of not all questions being visible on the screen.

Another factor to consider is cognitive dissonance: formerly, when a unit had been completed, an entire class period was dedicated to assessment. In the following class meeting, tests were returned and new content was introduced. With online testing, one to two class meetings' worth of new content was covered during the four-day testing period. The introduction of new material may have competed for students' attention with that being tested from the previous unit. In a number of recent studies of equivalence between computer and paper based tests (Clariana & Wallace, 2002; DeAngelis, 2000; Mason, et al., 2001; Stephens, 2001), students took CBTs during regular class meeting times. Students in Alexander, et al.'s (2002) preference study took Internet based tests outside of class, in a proctored computer lab; however, the researchers made no mention of any effect of competition between old and new material.

To compensate for this effect, new musical styles subsequently have been introduced in the Western Music class using cooperative listening exercises in which small groups of students compare the way a musical element was used in the previously covered style to how it sounds in a composition that represents the new style period. Thanks to increased testing capacity that will be provided by the new campus learning center, in future semesters a shorter testing period may also be implemented to prevent competition between familiar and new course content.

One of TestPilot's most powerful features enables students to recall their tests over the Internet after the conclusion of the testing period. In addition to displaying correct and incorrect answers, TestPilot offers students the ability to replay audio selections when reviewing their corrected listening tests. Unfortunately, only a third of students chose to recall and review their exams after the conclusion of the four-day testing period. Despite e-mail reminders and class announcements, only a slight improvement occurred in this number in the second semester that online testing was implemented. This trend is a matter of concern, as students were not taking the time to investigate where problems occurred and to consider how to improve their future performance. (One needs to bear in mind that these students were primarily freshmen and often had not yet developed effective study habits.) More work must be done to encourage students to take advantage of this opportunity for feedback and improvement.

Conclusions

The success of Internet based testing requires an organized and flexible instructor (Granger & McGarry, 2002). Systems must be well designed and thoroughly tested prior to implementation. Contingency plans need to be formulated to account for the inevitable breakdowns that will occur. Students' perceptions of Internet based testing must also be taken into account. Students expect a clear and easy-to-use test interface, ease and flexibility of scheduling, reliable systems and security, a relaxed setting free of distractions, and timely, useful feedback. While the majority of students surveyed seemed satisfied with, or at least unfazed by, online testing, not everyone was pleased. In the first semester that online testing was implemented, course evaluation ratings for quality of instructor dropped over half a point (on a 1-7 scale). As Stephens (2001) recommended, when told in advance the reasons why this approach to assessment was being used and of its potential benefits, students seemed more accepting of computer based testing. For the second semester in which online testing was used, the quality of instructor ratings for Introduction to Western Music were as good, or better, than ever.

The significant drop in scores that students experienced when taking tests online is troubling, especially when compared to lower-achieving students who took tests of the same difficulty in class. The effect of test anxiety caused by taking tests online needs to be examined further. Many students already experience stress because the tests involve music listening.

As the use of Internet based testing becomes more commonplace, student apprehension may lessen over time (DeAngelis, 2000). Future research should be conducted to investigate whether the disparity in students' performance between paper based and computer based music tests occurs in the terms/definitions or in the listening portions of tests. Another equivalence study could explore the effectiveness of computer based listening test formats that enable students to see all items associated with a musical selection at the same time, as they can with paper tests. If TestPilot's limitations in formatting listening questions do, in fact, have an adverse effect on student performance, music professors will need to weigh this drawback against other potential advantages that Internet based testing offers.

For the principal investigator/teacher, the decision to employ online testing was driven by the need to reclaim time to cover course content that had been displaced by in-class assessment and by the use of cooperative learning. Negative factors the students may have perceived or experienced—resentment over scheduling and taking a test outside of class time, studying for and taking a test while covering new material in class, or shortcomings of the testing software—have at least been offset by positive benefits to the students and the investigator/teacher, such as exposure to more course content; the opportunity for immediate, detailed feedback; and better pacing during test taking.

References

Alexander, M. W., Truell, A. D., & Bartlett, J. E., II. (2002). Students' perceptions of online testing. *The Delta Pi Epsilon Journal, 44*(1), 59-68.

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer based testing. *Journal of Research on Computing in Education, 28*(3), 282-299.

Bugbee, A. C., & Bernt, F. M. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education, 23*(1), 87-100.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 367-407). New York: American Council on Education—Macmillan.

Clariana, R., & Wallace, P. (2002). Paper based versus computer based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593-602.

DeAngelis, S. (2000). Equivalency of computer based and paper-and-pencil testing. *Journal of Allied Health, 29*(3), 161-164.

Granger, M. J., & McGarry, N. (2002). Incorporating on-line testing into face-to-face traditional Information Systems courses. *Proceedings of the 17th Annual Conference of the International Academy for Information Management: International Conference on Informatics Education Research*, 220-226.

Mason, B. J., Patry, M., & Berstein, D. J. (2001). An examination of the equivalence between non-adaptive computer based and traditional testing. *Journal of Educational Computing Research, 24*(1), 29-39.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests* (College Board Report No. 88-8). New York: College Entrance Examination Board.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-58.

Reimer, B. (1985). *Developing the experience of music* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Russo, A. (2002). Mixing technology and testing [Electronic version]. *The School Administrator, 4*(59), 6-12.

Smialek, T. (2000, November). *Active and collaborative learning in an introductory music course for non-majors.* Paper presented at the Forty-third Annual Meeting of the College Music Society, Toronto, Canada.

Spanier, G. (2002, April 18). Statement by Penn State President Graham B. Spanier on the Penn State Calendar. *Penn State Intercom.* Retrieved July 8, 2004, from http://www.psu.edu/ur/archives/intercom 2002/April18/spanier.html/.

Stephens, D. (2001). Use of computer assisted assessment: Benefits to students and staff. *Education for Information, 19*(4), 265-275.

TestPilot (Version 3.2.2p5) [Computer software]. (2002). Battle Ground, IN: ClearLearning. http://www.clearlearning.com/.

Vispoel, W. P., & Coffman, D. D. (1992). Computerized adaptive testing of music-related skills. *Bulletin of the Council for Research in Music Education, 112*, 29-49.

Vispoel, W. P., Wang, T., & Bleiler, T. (1997). Computerized adaptive and fixed item testing of music listening skill: A comparison of efficiency, precision, and concurrent validity. *Journal of Educational Measurement, 34*(1), 43-63.

Wing, H. D. (1961). *Wing Standardized Tests of Musical Intelligence.* Windsor, England: NFER Publishing Company.

Wise, S. L., & Plake, B. S. (1990). Computer based testing in higher education. *Measurement and Evaluation in Counseling and Development, 23,* 3-10.