4. REVIEW OF STUDIES CONCERNING PHYSICS, GRAPH CONCEPTS IN PHYSICS, AND COMPUTER AIDS IN GRAPHING

4.1. Hestenes, Wells, and Swackhamer

An important study was conducted by Hestenes, Wells, and Swackhamer [Hes92a] concerning a method to probe student beliefs on the concept of force and how they compare to the Newtonian concept. The Force Concept Inventory (FCI) is a multiple-choice test instrument that provided choices between correct and commonsense alternatives to Newtonian concept questions regarding aspects of force. This test has been given to many high school and college students and generated much literature. The primary uses for the FCI are as a diagnostic tool for student misconceptions and for evaluating instruction on Newtonian concepts.

The FCI is important because several of the questions in this well-researched test served as the basis for questions asked in the auditory graph tests. Not all of the questions could be utilized due to the nature of the display format, and the FCI covered only material relating to the concept of force whereas the conducted study had a more general basis of questioning. More will be mentioned of how this test was adapted in the section on Experimental Design. The Mechanics Baseline Test is similar instrument by Hestesnes and Wells [Hes92b]. Several questions from this test were adapted for use in the auditory graph study.

4.2. Trowbridge and McDermott

There are several studies that try to characterize how students conceptualize motion and the role that graphed information plays in their understanding. The first, by Trowbridge and McDermott [Tro80], looked only at how students understand velocity of simple observed motions. This paper described a guided interview process with over 300 subjects. The subjects were asked a series of questions relating to demonstrations about the motion of simple objects. In several of the questions, subjects were asked to compare

the speed of two objects. When responding to one of the questions, some students would spontaneously draw graphs to aid as a communication device. However, it was observed that students were unable to correctly incorporate their graphing skills into a successful understanding of velocity. From student responses to interview questions, it was stated that students have a disparity between what their graphs illustrate, and what they think their graphs illustrate. It is this disparity that provided an expanded study with additional research.

4.3. McDermott, Rosenquist, and van Zee

The expanded study by McDermott, Rosenquist, and van Zee [Mcd87], looked not only at velocity, but also at kinematics as a whole, and how students had trouble connecting physical concepts and graphical information. Their descriptive study with several hundred students involved identifying areas in which students have difficulty in their interpretation of graphical information. Their data were derived primarily from responses to questions given to the students, presumably as part of an exam. The results were mainly a categorization of the more prevalent difficulties observed.

There are two main areas of difficulty that were identified: connecting graphs to physical concepts, and connecting graphs to real world phenomena. In the first category, the identified problems were: difficulty discriminating between the slope and height, interpretations of changes in slope and height, relating graphs between position, velocity and acceleration coordinates, matching narrative information with relevant features of a graph, and interpretation of the integral, or area under the graph. In the relation of graphs to the real world, students drew graphs relating to the motion of a ball on various tracks. From these graphs common problems were: an inability to represent continuous motion with continuous lines, separating the shape of the graph from the path of the motion, representing negative velocity, representing constant acceleration, and distinguishing among different types of motion graphs (x, v, and a vs. t).

4.4. Beichner

A comprehensive study about students' interpretation of kinematics graphs was performed by Beichner [Bei94]. The primary purpose of this article was to report on a study aimed at uncovering student problems with interpreting kinematics graphs. A secondary purpose was the proposition of a model for creating research-oriented multiple-choice tests that could be used as diagnostic tools or as formative and summative evaluations of instruction. Parts of the multiple-choice test that were developed in the Beichner study were used as question templates for the current research

The test evolved in several parts. Draft versions of the test were administered to 134 community college students who had been taught kinematics. The results were used to modify several of the questions, and the revisions were given to 15 high school, community college, four-year college, and university science educators. These individuals completed, commented on the appropriateness of the objectives, criticized items, and matched items to objectives in an effort to establish content validity. The final tests were then given to 165 juniors and seniors from three high schools and 57 four-year college physics students.

The test instrument consisted of 21 multiple-choice questions divided into seven testing objectives. The objectives were chosen upon examination of commonly used test banks, introductory physics books and informal interviews with science teachers. The test was designed to focus on interpretation skills. Three test items were written for each objective, most of these being written by the author although some items were adapted from previously used tests. The test questions and results of student performance were appended at the end of the paper.

All of the statistical procedures indicated that the test was valid and reliable. Results of data analysis also indicated several other results. First, calculus-based physics students did significantly better on the test (mean of 9.8 vs. 7.4) than algebra/trigonometry-based physics students ($t = 4.87$, $p < 0.01$). Second, college students were not significantly better than their high school counterparts ($t = 1.50$, p 0.13). Third, the mean for males of 9.5 was significantly better than the 7.2 mean for females ($t = 5.66$, $p < 0.01$).

The developed instrument appeared to be generalizable to a wide range of students studying kinematics, from high school to university courses, across the country. The results allowed for objective grading and the ability to provide statistical analysis from large numbers of subjects.

In a later study, Beichner [Bei96] investigated the impact of students analyzing video motion on their ability to interpret kinematics graphs. In this study it was found that the greatest impact on student's ability to interpret graphical information comes from hands-on involvement in data acquisition. The study demonstrated a strong correlation between the amount of exposure to video graphing labs and students' scores on a multiple-choice test on graphs, indicating a better understanding of kinematics graphs.

4.5. Mokros and Tinker

A set of studies by Mokros and Tinker [Mok87] demonstrated that middle-school students could learn to communicate using graphs in the context of appropriate microcomputer-based laboratory (MBL) investigations. The first preliminary study attempted to locate graph-related misconceptions, the second investigated children's graphing skills, and the longitudinal study examined MBL intervention.

In the first study, 25 seventh and eight grade students in a suburban school participated. The students were given a carefully constructed set of graphing problems in an interview setting. The problems were developed from the results of a pilot test to ensure appropriateness in terms of language, difficulty level, and coverage of various problem types. The interviews consisted of six graphing items and lasted 20 to 40 minutes. A protocol summary was completed for each student's performance. The findings of this study were that students exhibited two major types of errors, which have also been observed in college populations: graph as picture confusion and a weaker indication of confusion with relating slope and height.

The second study investigated the ways in which students learn graphing skills through MBL. Data were collected by observing individual lab groups. Students' interactions were recording as narrative summaries and by an event sampling process that was subjected to quantitative analysis. The study utilized an MBL course unit consisting

of five days of activities on position and velocity plotting. The observations and scores from a nine-question quiz in the second preliminary study indicated that after five days, students had developed graph interpretation skills.

The longitudinal study was designed to provide more evidence about the impact of MBL on graphing skills. This study involved a pre-test, treatment, post-test design, with each test having two components: a multiple-choice test of graphing skills and an interview where the students talked through their thought process. In the longitudinal study, scores on the 16 graphing items showed a small ($\square$ = 15%), but significant, improvement. This research showed that students could learn graphing concepts over a long time frame when using MBL's.

4.6. Brasell

A study by Brasell [Bra87] not only extends Mokros and Tinker to high-school students but also assesses the effect of a very brief exposure to a kinematics unit on the ability to translate between a physical event and the corresponding graphical representation. The study also evaluated the effect of real-time graphing in comparison to delayed graphing of data on student learning.

The sample was drawn from entire physics classes (of seven to 17 students each) in seven rural schools in north Florida providing a total of 93 students. The students were mostly seniors and were familiar with the concepts included in the experimental activities. It is suspected that the choice of the students was a matter of convenience as the author is from the University of Florida.

The experiment was conducted over a three-day period, one day for the pre-test and orientation, one for the treatment, and one for post-testing and discussion. The treatment consisted of several groups: a Test only, a standard MBL display where data were displayed as it was acquired, a delayed MBL group where a 20-second delay was introduced between acquisition and display, and a pencil and paper graphing group that plotted their own graphs on paper. The MBL groups used curriculum units designed for the software. The paper and pencil group graphed complex motion described on a worksheet. Each class at each school had one group of students for each treatment to

provide a balanced design. Students were randomly assigned to each group on a class-wise basis.

Pre- and post-tests were described as consisting of content-specified, multiple-choice items requiring students to translate between a verbal description of a physical event and the graphic representation of it. The pre-test had been developed and used by a previous researcher for use with humanities college students. The post-test was conceptually similar to the previous study, but altered in format. Due to the format change, performance changes were utilized only as a covariant. SAT scores were recorded and used as a covariant. It was stated that neither the pre- nor post-tests were checked for reliability. Validity of the tests was not mentioned. Analysis of covariance was used to reduce error variance of post-test scores.

Factorial analysis of covariance was utilized. The pre- and post-tests were divided into two sub-tests, one for distance and another for velocity. It was found from *F* tests using 3 treatment degrees of freedom, and 68 degrees of freedom for the data, that overall scores for standard MBL treatment were significantly higher than scores from the other treatments (F (3, 68) = 6.59, $p < 0.001$). While it was shown that scores for both sections were higher, only the distance sub-test scores were significant with F (3, 68) = 6.47, $p < 0.001$. The velocity sub-test was not considered a significant difference, $F(3, 68) = 1.80$, $p = 0.156$. A table of the results as well as a graph of the mean error rates for the different groups were presented.

Brasell stated that 90% of the difference in the mean scores was due to the real-time nature of graphing provided by MBL. At no time was the performance of the delayed MBL graphing significantly superior to that of students in the control groups. It was found that even a short delay in displaying graphs dramatically reduced the effectiveness of the MBL on graphing skills. It was suggested that one of the effects of the delayed graphing was that students appeared less motivated, less actively engaged, less eager to experiment, and more concerned with the procedure, rather than the concepts.

4.7. Linn, Layman, and Nachmias

A study on the cognitive consequences of microcomputers on graphing skill development was attempted by Linn, Layman, and Nachmias [Lin87]. In their study, they explored how students' graphing skills changed after exposure to MBL intervention. Their study centered on an "ideal" chain of cognitive accomplishments. These were: graph features, graph templates or sequences of activities that are used repetitively to comprehend the graph, graph design skills which augment and consolidate the templates for new problems, and graph problem-solving skills. They found that the MBL intervention increased student's ability to identify trends and locate extrema, but did not compare their results to non-MBL methods. Exposure to the MBL graphs acted as a basis on which students built their graphing models.

4.8. Thornton and Sokoloff

A study that did attempt to compare the effectiveness of MBL techniques was conducted by Thornton and Sokoloff [Tho90]. The purpose of their study was to compare the effectiveness of curricula that take advantage of MBLs presenting data in immediately understandable graphical forms to the effectiveness of non-computer based courses. The ability to learn basic kinematics concepts was evaluated with pre- and post-testing as well as by observations.

The sample was drawn from more than 1500 college and university physics students taking non-calculus and calculus based General Physics courses at the University of Oregon and Tufts University over a three-year period. The research design consisted of testing students enrolled in a laboratory course involving microcomputers to display the graphical information and comparing the results from their post-test scores to those of students who were not enrolled in the lab. Data were collected by 50-item multiple-choice pre- and post-tests. It was not mentioned if the same test was given at both universities. The reported data showed dramatic reduction (up to 40%) in the error rates when compared to the non-MBL group.

4.9. Analysis and Discussion

The studies reviewed in this chapter concerned the interrelationship of how students learn physics when using graphs and the use of computers to display graphical information can affect student learning. Perhaps the most important studies with regards to the development of the auditory graph tests used in this work were those by Hestenes *et al*. and by Beichner.

The FCI questions were concerned with determining where students were having difficulties in physics and were more focused in their subject matter than those used in the current study. Beichner's study was of even greater aid in question development as it reported on a multiple-choice test involving kinematics graphs. While not all of the questions from these papers were compatible with the current research, they were valuable templates upon which to build the physics multiple-choice auditory graph questions. McDermott's studies were useful in their focus on understanding where students have difficulty, especially in the areas of connecting graphs to physical concepts and distinguishing among different types of motion.

Since the current research utilized computer portrayal of graphical information, some discussion of the research investigating how computers have played a role in graphing was included. These studies were valuable as they also provided a basis from which to draw material for questions used in the current studies.

While the MBL studies indicated that learning had taken place with the use of computer generated graphs, a major shortcoming of all these studies was the lack of comparison to equivalent non-MBLs. For example, in the study by Thornton and Sokoloff students who did not participate in the microcomputer lab did not participate in any lab experience, hence were not as practiced as the MBL group. In addition, for some of the subjects, the lab was a separate, and an optional course, so the students who took the MBL may have been self selected for better performance. Another explanation is that those students not taking the lab may not be as comfortable, practiced, or competent with physics as the MBL group, which would also cause a difference in scores between groups. These studies are useful however, as they demonstrate the prevalence of

computer use for graphing in current physics courses. In all these studies, the students were comfortable with computers as tools for displaying information.

5. REVIEW OF STUDIES ON AUDITORY GRAPHING TECHNIQUES

There is a large field devoted to the representation of data with sound. Generally, this field falls under the heading of Auditory Display and can encompass a wide range of sound representations such as the use of auditory cues ("earcons") as locators to more direct representations of data. The field is large enough for conferences such as the International Conference on Auditory Display (ICAD) with published proceedings [ICA94].

The quest to find a useful auditory data display has been approached from many fields such as mathematics, chemistry, computer science, as well as physics. From the diversity of auditory display techniques, it is readily apparent that no single display will suffice as a universal presentation method, just as no single visual graphing method works for all data. The following studies are those that directly relate to auditory techniques that would otherwise use two-dimensional plots.

5.1. Pollack and Ficks

One of the first studies concerning auditory display of information was performed by Pollack and Ficks [Pol54]. In their paper they investigated the relationship between auditory display stimuli in order to find a satisfactory procedure for increasing the information that can be transmitted from elementary auditory displays. The basic task of their subjects was to identify different qualities of sound stimuli. There were eight sound qualities tested using tones and noise: frequency ranges of noise and of tone, loudness of noise or of tone, rate of alternation between noise and tones, duration of tone display, the fraction of time tone was on, and direction of origination of the tone. Subjects were students and military personnel. The sounds were binary coded, in that the tones were either high or low, alternation rates were fast or slow, sound intensity levels were loud or soft, etc. In half of the tests subjects responded as they listened to the display, while in the other half, they responded after the sounds finished.

Pollack and Ficks reported that their subjects found the multidimensional displays easy to learn, especially the binary coded displays, and that subjects tended to associate the sounds with verbal symbols (e.g. “chirping birds”). They also reported that the multidimensional displays were able to effectively transmit more information than unidimensional displays. However, there was little improvement in information transmission when the dimensions were subdivided (degrees of loudness or alternation rates). The average error in correct identification of the auditory dimension was lowest for the binary comparison of frequency of the tone, at 0.08%. This rate was dramatically lower than for the other dimensions studied. The next lowest values were for sound duration (0.9%) and repetition rate (1.1%).

Their conclusion was that the use of multiple stimulus dimensions is a satisfactory method for increasing the transmission of information via auditory displays. Another conclusion was that it is more useful to have a greater number of binary coding dimensions rather than subdivision of only a few dimensions.

5.2. Mansur, Blattner, and Joy

Mansur, Blattner, and Joy [Man85] reported on a very significant study for representing data by sound. Their study, which generally provided the template for the current investigation, used sound patterns to represent two-dimensional line graphs. They were investigating a prototype system to provide the blind with a means of understanding line graphs similar to printed graphs for those with sight. This study used auditory graphs that had a three-second continuously-varying pitch to present the graphed data. The auditory graphs were also compared to engraved plastic tactile graphical representations of the same data. The authors cited research by Stevens, Volkmann, and Newman [Man85] on the pitch response of hearing that showed an exponential relationship between pitch and perceived height.

Mansur, Blattner, and Joy found in their study that there were difficulties in identifying secondary aspects of sound graphs such as the slope of the curves. They suggested that a full sound graph system should contain information for secondary aspects of the graph such as the first derivative. Their suggestion was to encode this

information by adding more overtones to the sound to change the timbre. They also suggested utilizing special signal tones to indicate a graph's maxima or minima, inflection points, or discontinuities.

Their main study consisted of several comparison tests to indicate the effectiveness of sound versus tactile graphing methods. These consisted of comparing the slope of lines, straight vs. exponential lines, monotonicity, convergence, and symmetry. There were fourteen subjects, half of whom were blind. The sighted subjects were blindfolded for the tests. The subjects were tested with one presentation method, and then re-tested with the other method. The type of graph subjects received first was by random assignment.

The results were that the tactile graphs had a small, but statistically significant, advantage to the sound graphs in overall accuracy (88.3% vs. 83.4%). This disparity appears to come mainly from the comparison of straight lines vs. exponential curves where there was a 12% difference in the accuracy of identification (96% vs. 84%). Also, a test of whether a graph was converging to some limiting value had a 9% difference in the scores (89% vs. 80%).

5.3. Lunney and Morrison

Lunney and Morrison [Lun90] describe an auditory alternative to visual graphs in order to provide access to instrumental measurements. Their system was to convert infrared chemical spectra into musical patterns. The translation method first converted the continuous spectral pattern into a "stick spectrum" in which absorption peaks are replaced with lines representing location and intensity. The spectrum was then mapped to a chromatic scale with the infrared frequency converted to pitch. The sound map was played in the form of two patterns. The first pattern was to play from highest pitch to lowest, with intensity represented by note duration. The second pattern was to play the spectrum in order of decreasing peak intensity, with equal note duration. The first pattern was played twice, and the second three times. The six strongest peaks were also played together as a chord at the end. The authors mentioned that this was an effective technique for chemical analysis of spectra.

5.4. Frysinger

A review paper by Frysinger [Fry90] details various research approaches to data sonification. The bulk of his review describes data sonification, the areas of psychoacoustics (the psychology of hearing), and sound perception issues. Several of the articles that were reviewed are summarized above. Frysinger provided some very general indications for research direction and methodology. Some of these suggestions are utilization of synthetic data generation to control parameters, the use of two sessions to compare the effectiveness of two display types, and using forced choice type questions.

5.5. Minghim and Forrest

For more complex sound mappings, Minghim and Forrest [Min95] presented a review of several studies and an analysis of data sonification development. They mentioned the following areas where sound can be a useful tool in aiding data visualization: adding further dimensionality to data, alternate perceptual properties, additional interactive processes, inherent time dimension for data, use of sound as a validation process, and increasing the ability to remember data due to additional modal encoding. They also described a sonification program called SSound which implements a number of sound functions for aiding surface-based data analysis. Various surface properties were mapped to sound qualities such as pitch for density, rhythm for change in a function, and timbre for data correlation. Sound was spatially located using quadraphonic speakers to indicate information depth. Users of this sonification system required training for interpretation of the complex sounds. The authors did not report any formal results as to the effectiveness of the system.

5.6. Wilson

A similar program to represent data by sound is the Listen data sonification toolkit described by Wilson [Wil96]. The primary goal of this program was to provide a flexible sound toolkit for use in sonification research. The Listen program is an object-

oriented modular system designed on Silicon Graphics (SGI) workstations incorporating MIDI sound libraries. Listen was designed to be a component for incorporation into other data visualization programs. The main modules of the Listen program are: Interface, Control, Data Manager, Sound Mapping, and the Sound Device modules. Only the Interface module interacts with the Control module, which then interacts with the other three. With this program, data fields can be mapped to four types of sound parameters: pitch, duration, volume, and location. Pitches used the semitone scale. Data could also be given timbres relating to various MIDI instruments for further diversification.

5.7. Flowers and Hauer

There are several important studies relating to the success of auditory graphs for display purposes. Flowers and Hauer produced a set of studies investigating the perceptual similarities between visual and auditory graphs.

The first paper [Flo92] described a single experiment to study how effective information about central tendency, variability, and the shape of data distributions could be portrayed with auditory graphs versus a visual graph. Data in this experiment were presented as auditory histograms, auditory quartile displays, and visual histograms. The auditory histograms presented the data distribution with the numeric value mapped to pitch, and the frequency of the data mapped to the number of times a note was repeated. The visual histogram was presented on a computer screen as text characters, with the numeric value mapped to the *x* axis, and the frequency of the data distribution was presented with vertical stacks of asterisk symbols. The auditory quartile displays were a musical analogue of the Tukey box and whiskers drawing that coded the minimum, first, second, and third quartile, and maximum data values as a set of five musical notes.

Twelve psychology graduate student subjects performed 132 comparison trials in each of three sessions, with only one presentation modality per session. The subjects gave a 1 to 10 similarity judgment rating for each of the graph comparisons. The judgments were based on differences in central tendency, variability, and the shape of the data distribution.

This study specifically investigated the perceptual structure of plots through dissimilarity judgments of a graph's slope or level when depicted by visual versus auditory displays. The study consisted of three tests, labeled Experiments 1, 2, and 3. Experiment 1 investigated student's ability to distinguish visual graphs, while Experiment 2 investigated auditory graphs. Experiment 3 was similar to 1 and 2 except that it provided a more sensitive evaluation between visual and auditory graphs. Results showed that the correlation between judgements and stimulus parameters for the auditory histogram (r=0.36) and quartile display (r= 0.40) graphing techniques produced a far greater dissimilarity rating than did the visual histogram (r= 0.06) graphs. However, the opposite was true for skew (r= 0.11, 0.06, and 0.39) and kurtosis (presence of long or short distribution tails, r= 0.07, 0.02, and 0.21). Judgments on the range of data values were similar for all three graph types. The authors commented that the surprisingly low correlations between the dissimilarity judgements may have been related to little variation in the standard deviations for the distributions used as stimuli.

The second study by Flowers and Hauer [Flo93] extended the first with two experiments investigating whether combined auditory and visual presentations enhanced discrimination of stimulus parameters, and whether the auditory quartile (Tukey box and whisker) plots provided an adequate distribution of information. In the first experiment, 25 paid student subjects, with normal hearing and vision, participated in a study similar to that conducted in their previous paper on comparative judgment analysis of visual and auditory histogram graphs. There were three display methods: visual presentation, auditory presentation, and a combined auditory and visual presentation, with the auditory histograms using the same method as previously described. Their results showed that visual graphs again had a greater reliability in the dissimilarity judgments than auditory graphs, and that there was no evidence that combined presentation led to a greater consistency of judgments than visual presentation alone. The second experiment consisted of the use of auditory quartile displays, slightly modified from the previous study in that these displays had an additional leading note, representing the median as a prefix to the five-note system. A comparison of dissimilarity judgments between the original quartile display method, and the leading note prefix method showed a greater attention to the median (r= 0.58 vs. 0.20) but reduced attention to skew and range

(r=0.20, 0.31 vs. 0.38, 0.41). Thus, focusing on the central tendency came at the expense of other characteristics.

In their third paper, Flowers and Hauer [Flo95] compared the perception equivalence between auditory and visual graphs and the ability to convey information regarding the profile of changes of an independent variable. At least two of the samples consisted of introductory psychology students at the University of Nebraska. In Experiment 1, there were 18 students (7 male, 11 female) who received credit for a research exposure requirement for their introductory course. Experiment 2 consisted of 14 student volunteers who were each paid $15. Experiment 3 consisted of two groups of students who were in a similar situation as those in Experiment 1. It was not stated if students in one experiment were also in another, or what size of a class population that these students were drawn from; thus the number of students involved could be from 19 to 51.

There was some discrepancy between the methods for comparing the graphs in Experiments 1 and 2. In the first, students were instructed to sort 68 graphs into no fewer than three and no more than 10 categories. In the second, students used a pair-wise numeric (1-10) dissimilarity rating procedure of all possible pairings of 34 of the 68 graphs. Half of the subjects (seven) received one pairing set, and the other half compared a second set. The auditory graphs used the same data sets as the visual plots.

In Experiment 3, 16 graphs were used for comparison purposes. In this trial, both the visual and auditory graphs were compared in a pair-wise fashion. The visual graphs were displayed at the same time while the auditory graphs were displayed sequentially.

The authors' conclusion was that the experiments illustrated a close correspondence between the perception of auditory and visual graphs with regards to gross differences in function shape, as well as slope and level (height) perception. The main result of this study was to demonstrate how to use auditory graphs to convey information about distribution central tendency, variability, and shape to observers who had not been previously exposed to auditory representations of data.

5.8. Turnage, Bonebright, Buhman, and Flowers

Turnage, Bonebright, Buhman, and Flowers [Tur96] reported on a study, comprising of two experiments, comparing the equivalence of visual and auditory representations of periodic numerical data. The first experiment investigated whether equivalence of auditory vs. visual presentations of wave form stimuli would parallel that reported for other graph types. Twenty-six undergraduate psychology student subjects participated to fulfill a course research requirement. The subjects were divided into two groups of 13. Graphs consisting of 100 data points were constructed with three shape patterns (sine, square, or combination), three frequencies (high - 8 cycles/100 data points, medium - 6/100, and low - 4/100), and two amplitudes (high, and low) for a total of 18 graphs. The visual graphs were constructed with Microsoft Excel and presented via overhead transparencies. The auditory graphs were played as a series of 100 musical notes with a two-octave range. The *y* axis was represented by pitch and the *x* axis as time. Each auditory graph had a 6-second duration.

The subjects were presented with the task of providing similarity ratings for all independent pairs (153) of graphs. They were initially presented all 18 graphs in random order and had three practice tries for familiarization to the process of discriminating graphs. They then rated the graph pairs on a 9-point similarity scale for each of the three conditional dimensions (1: shape, 2: amplitude, and 3: frequency.) Coefficients of congruence (CC), interpreted like correlation coefficients, revealed that the visual and auditory graphs were very similar for all three condition dimensions (CC 1= 0.96, CC 2= 0.98, CC 3 = 0.94.) Thus, the two graphing methods have high similarity for difference discrimination. There was also some indication of slightly greater discrimination between sine and composite wave patterns with the auditory display than with the visual display.

The second experiment investigated the relative performance accuracy of visual and auditory graphs on a task involving discrimination between similar wave forms. Thirty-eight undergraduate psychology student subjects participated to fulfill a course research requirement. The subjects were divided into two groups of 19, for each graphing method. The graphs were constructed and presented as in the first experiment. Forty pairs of wave form graphs were selected for a comparison task. Subjects were sequentially

presented with two graphs, A and B, and then presented with a third graph, X, from which they determined whether *X* was the same as A, B, or neither. The subjects were given three practice trials for familiarization. Results showed a significant difference in the performance scores of the two groups with the Auditory graph group average of 81% correct, and the Visual graph average of 96% correct.

5.9. Flowers, Buhman, and Turnage

Most recently, a study relating to auditory graphs for display purposes was conducted by Flowers, Buhman, and Turnage [Flo97]. This study investigated the equivalence of visual and auditory scatter plots to explore bivariate data. Their study consisted of two experiments, the first examining the relationship between visual and auditory judgments for the direction and magnitude of correlation for 24 bivariate data samples.

The first experiment used 45 unpaid advanced undergraduate psychology student volunteers. Nineteen of the subjects, in groups of three to eight, judged visual scatter plots of data samples, while the remaining 26 were assigned in groups of five to 16 to judge auditory scatter plots of the same data. The graphed data samples consisted of 50 random numbers about a Gaussian distribution with a mean of 50. Some of the data samples were given transformations to produce various correlations between the resulting 24 sample plots. The standard deviation within data samples ranged from about 6.2 to 11.6. Sound generation was constructed using Microsoft Excel to compute parameters for use in the CSound program. Each auditory graphs had a five-second duration, with individual data points represented by 0.1 second guitar pluck note. The *x* axis was represented by time and the *y* axis by a pitch scale ranging one octave below to two octaves above middle C. The data was mapped to a chromatic scale.

Subjects rated the magnitude and sign of the correlation between the variables in the graphs. The judgment data were recorded as a distance from the zero point on the scale. The visual graphs were presented for 10 seconds, while each auditory graph was played twice for a total of 10 seconds of listening time. Pearson's correlation for the comparison between the actual correlation and the judged correlation was $r = 0.92$ for the

visual group and $r = 0.91$ for the auditory group. A t-test showed no significant difference between the auditory and visual groups.

The second experiment was a direct evaluation between visual and auditory perceptual sensitivity to data points lying outside the main data groupings. This was accomplished by examining changes in the perceived magnitudes and direction of the correlation for scatter plots that were altered with the addition of data points. In this experiment, 32 advanced undergraduate psychology student volunteers participated, 20 in a visual graph group, and 12 in an auditory graph group. Eight data sets from the first experiment were modified by moving one data point: in half of the sets the data point was moved to an outlying position in the center of the plot, and in the rest the data point was moved to an extreme end of the plot. The eight original plots, the eight modified plots, and eight additional plots were used so the number of test stimuli equaled that used in the first experiment.

Of the 24 plots, two of the modified plots showed significant differences in the judgment of correlation magnitudes. The two were plots where the outliers were for moderately correlated data samples rather than for weakly or strongly correlated data sets. Both auditory and visual conditions gave similar results. Thus, this study seems to indicate that judgments between correlation effects for both visual and auditory scatter plots are very similar. Both are effective in conveying sign and magnitude of correlation, and they are similarly influenced by error variances and by single outliers.

5.10. Analysis and Discussion

The studies reviewed in this chapter were found by the current author to be of great use when developing the auditory graph tests conducted for this work. The following discussion is a critique of how the reviewed studies helped define issues related to this work as well as some of their strengths and weaknesses.

The use of pitch to represent data was shown by Pollack and Ficks to have a lower error rate in comparison to other auditory dimensions such as sound duration, repetition rate, or loudness. The ability for pitch discrimination has been used by several researchers to create auditory graphs where the *y* axis data value is represented by pitch

and the *x* axis is represented with time. However, Pollack and Ficks also noted that greater information can be transmitted to the listener by increasing the number of binary coding dimensions rather than subdivision of the codings. Hence, if increased information in an auditory graph is desired, additional binary type sounds may be useful considerations.

Mansur *et al*. demonstrated the viability of auditory graphs by comparing auditory graphs to tactile graphs but noted some difficulty subjects had distinguishing between straight lines vs. exponential curves. Studies by Flowers *et al.* extended the study of auditory graphs in a series of comparisons to visual graphs. Their work included several graph types to histograms, scatterplots, and Tukey box and whiskers drawings.

The basic auditory graph served as a starting point for the auditory graphs used in the current research. The previous studies provided auditory graphing methods that had been found to be reasonably effective replacements for visual graphs. The studies concerning more complex methods for mapping data to sound were useful for gaining ideas of what had been investigated and for sound generation techniques and controls.

This chapter, along with chapters 3 and 4, have been an attempt to demonstrate that there is a wide range of literature related to the current research. Perhaps the most relevant studies are those concerning auditory graphing techniques, especially those by Mansur *et al*. and those by Flowers *et al*. The subject material for the questions on which to base the graphs came predominantly from those studies presented in the chapter on physics graphs and concepts. Those studies relating the use of computers in the graphing process demonstrated that the student subjects are familiar with the computer as a graphing tool, and that it need not be presented as an unfamiliar object.

While the studies concerning graph perception may seem the least relevant, they serve as an underlying basis for the foundation of this work. It is important to keep in mind the common structures that people are familiar with when creating new representations for data display.

6. THEORY

6.1. Hypothesis Development

Wavering noted that "graphing is a tool used in science to display data and aid in the analysis of relationships between variables. Also, graphs are part of our daily existence with their use in all media." [Wav89, p. 373] In spite of the prevalence of graphs, several studies have uncovered areas where students have difficulty interpreting graphical information that is used not only in physics, but also in mathematics and economics [Bei94, Mcd87, Lei90, Coh94]. Because of these difficulties, researchers have devoted considerable energy to the teaching and study of graphs.

While progress is being made in teaching graphical information, some attention to how the data is displayed is warranted. Tufte discusses the effectiveness of graphical display methods [Tuf90] and literature concerning optimal display methods was reviewed in chapter 3. Unfortunately, almost all of these studies concern visual display methods. There are several problems with focusing on only the visual data display aspect, most importantly: What happens when one cannot see the graph in question? Thus, it is important to explore other avenues of displaying the information contained in graphs. Auditory graphs are one method for presenting information in a non-visual format.

The current research is directed towards demonstrating not only that people can understand auditory graphs, but that they can also be used as effective displays for understanding and analyzing information. Previous research has focused on how people perceive graphs and how they use graphs to learn about physics. Several studies have also investigated how well people can make judgments about graphs. However, none of the previous studies have demonstrated whether auditory graphs can be practically implemented, and what sort of results could be drawn when students use auditory graphs to answer questions.

The ability to present data with an effective auditory format is one of the prime goals of this research. The working hypothesis for this study is that: in many cases, sound

graphs can be as effective as visual graphs for data representation and for making inferences about that data.

If graph types are highly equivalent, as suggested by the studies by Flowers [Flo92, Flo93, Flo97], then there should be little difference between a student's ability to identify and interpret information when given auditory or visual graphs. However, there is the possibility that there will be differences in performance due to unfamiliarity with the sound format. By asking questions based on graphical material, the effectiveness of auditory graphing methods can be measured.

To test the hypothesis, it is important to determine how well students are able to answer graph-based questions. One testing method is to have two equivalent groups of subjects answering questions. Each group receives either visual or auditory graphs with the questions. While identification of simple graphs is important, students' ability to interpret what those graphs mean is also significant. Thus, this study includes two types of questions: those that involve interpretation to identify a function, and those that require analysis of the data for interpretation of the physics concepts that the graphs represent.

A comparison of the performance of subjects using auditory graphs to that of subjects using visual graphs may indicate a difference between the two display methods. In addition, subjects may have better understanding of questions when both auditory and visual graphs are used. Subjects may find that the combination of formats is a helpful method to enhance the graph. Thus, three testing groups are reasonable to provide comparative data: visual graphs, auditory graphs, and both auditory and visual graphs. When the number of subjects is sufficiently large, a random assignment to one of the three groups should produce equivalent testing groups.

Comparing the performance of subjects' ability to answer graph-based questions with respect to which graph type they receive may yield several outcomes. The first, is that if student performance is equivalent among the auditory, visual, and the combination displays, then the display modalities are equivalent. They can answer and analyze questions equally well.

Studies by Flowers and Hauer demonstrated that there are several areas of perceptual equivalence between auditory and visual graphs. Thus, the possibility for equivalent performance when answering questions is a reasonable supposition. Turnage

et al. [Tur96] also reported rough equivalence between auditory and visual graphs when subjects were asked to identify properties of simple periodic wave patterns.

A second, albeit unlikely, outcome also exists: auditory graphs could outperform their visual counterpart. This outcome could be the result of an increased salience from auditory cues. Flowers and Hauer noticed this effect in some parts of their graph discrimination studies [Flo95, Flo97].

A more likely situation, however, is that there would be a performance difference due to greater familiarity of the visual graphs. This is understandable as students are trained to recognize and use visual graphs for many years by the time they take university level courses. Auditory graphs, on the other hand, are a completely new experience, and the amount of training they receive may strongly influence their performance. A study investigating an upper limit of the use of auditory graphs to convey information would require subjects with extensive auditory graph training. Comparable, but not equivalent, performances for discriminating differences between data sets when using auditory or visual graphs have been shown in the aforementioned studies [Flo97, Tur96].

If subjects completely fail to understand data presented with auditory graphs, currently reported research would be called into question. A finite limit on the practicality of auditory displays may exist. Also, such a result may demonstrate that the understanding of auditory graphs is not intuitive. Even simple data comparisons and analysis would require that subjects have intensive training and alternate auditory methods would need to be investigated.

6.2. Further Justifications for the Research

At the most fundamental level, this research provides a method for portraying graphical information to people who are unable to interpret a visual graph. While haptic (pertaining to the sense of touch) methods for creating graphical information have been used in the past, there are several difficulties including interactiveness, resolution, production, portability, and storage issues. Haptic graphs require a significant amount of time for identifying contained elements and often a tutor is necessary for explanation of the information. The auditory format can remove many of these limitations.

The basic auditory display used throughout this study, centers on mapping the *y* axis data value to pitch and the *x* axis to time. The exact relationship for the *y* axis pitch varies between experiments. However, there is always the association that high pitch (higher frequency values on the order of a couple of kilohertz) represents high data values, and low pitch (around 200 Hz) represents low values. This method provides a direct one-to-one mapping between pitch and data. In the Triangle Pilot, Web Pilot, and the Main Auditory Graph tests all of the graphs had zero or positive *y* axis data values. Thus, the lowest magnitude value had the lowest note, and the highest magnitude value had the highest note. There is a strong similarity between this mapping method and music notation.

The association of pitch to the magnitude of data values is common practice and has been widely used in research and in other data sonification programs. There are different sound mapping methods, but pitch is the most common, has been applied in many cases, and appears to be intuitive for most people. Investigation of other auditory mapping techniques was outside the scope of this work.

Previous studies have focused on general similarities, or the ability of subjects to identify trends or differences in comparative data sets. The next logical stage after the identification of parts of graphs is the interpretation of the graph as a whole and the analysis of the graph's meaning. However, previous studies have not investigated the ability of subjects to interpret auditory graphs. Interpretation of a graph includes identifying the trends of a graph, making conclusions based on displayed data, or using a graph to infer properties about the system used to produce that graph.

The current research addresses this issue by investigating how well students are able to answer physics and math questions based on graphed data. Many systems studied in physics use graphs to display data for analysis. Ideally, the data can be represented by mathematical equations. Physics is an ideal topic for the study of auditory graphs since there can be a separation between the identification of mathematical functions representing the graph, and the inferred properties of a system that the graph represents.

6.3. Implementation of the Research

6.3.1. Genesis of the Testing Process.

Personal computers have been used in many auditory display studies because of their ability to generate a wide variety of sounds. The TRIANGLE program developed by Oregon State University's Science Access Project takes advantage of this sound capability to generate an auditory graph as a complement to, and substitution for, the visual graph display.

TRIANGLE's primary purpose is to provide a workspace for students and scientists to read, write, and manipulate mathematics. TRIANGLE contains a calculator that permits evaluation of most standard math expressions. The calculator also evaluates *y* versus *x* functions and displays the results in a plot window. An auditory graph of the function or of data provides a blind or visually impaired user with a quick semi-quantitative overview of the graph. The auditory graph contains a number of display options. In addition, there is a moving icon on the screen to provide information about the graph to users who are both blind and deaf [Gar96].

The auditory graphs produced by the TRIANGLE program created the question of: How useful is this type of display to the intended user? To answer this question, it was necessary to develop an unbiased testing method between auditory graphs and visual graphs in the context for which they would be used in the program. The context is the investigation of properties of mathematical functions and the display of scientific data.

Because the TRIANGLE program was the genesis of the research, the initial investigation centered on using this program as a testing medium. TRIANGLE displayed both visual and auditory graph formats, as well as a text region that could be used to display questions about the graphs. Hence, in the initial stages, it was a good candidate for implementation of the testing process. Later, a testing process based on the World Wide Web proved to be a more flexible alternative with many advantages and is discussed in chapter 8.

6.3.2. General Test Design

The first stages of the testing process required several assumptions. The first was that the auditory display method implemented with the TRIANGLE program would be sufficient to the task. This was not an unreasonable assumption given that previous research employed similar auditory mapping methods. Also, TRIANGLE had been tested by several people for usability and stability.

The TRIANGLE program was used as the initial basis for the study. It was necessary to formulate an unbiased testing process that would demonstrate the ability of subjects to answer and evaluate questions based on graph types. A standard testing method is the causal-comparative design. This method consists of a pre-test, treatment, and post-test. Subjects are given a pre-test to measure their initial state, some form of teaching or learning treatment, and a post-test to measure their final state. Comparing the pre- and post-test scores provides a judgment on the effectiveness of treatment methods.

Although the causal-comparative method is convenient, it only demonstrates the ability of subjects to learn and to use auditory graphs. It provides information about the type of training that the subjects receive. Auditory graphs have been shown in previous research [Man86, Flo92] to be useful for identification of basic graph types, such as linear or curved, and for dissimilarity judgments. These types of investigations are not the focus of this study.

A comparison between two or more groups of subjects can be combined with the pre-test, treatment, post-test method. The pre-test verifies that the groups have equivalent abilities, or can be used to give a basis for correction if the groups are found to be non-equivalent. In the current study, the treatment was the visual or auditory display of a graph. The post-test was a series of questions and their associated graphs. The results of the post-test were compared to judge the effectiveness of the graphing treatment methods.

An assumption of this study was that knowledge of the subject matter used in the questions would be an important effect. A subject's understanding of the material could affect his or her overall performance. Subjects were randomly assigned to different groups so that student performance in each group would ideally be similar. By comparing the performance of two groups on identical questions, any difference was thus focused on

the ability of subjects to utilize the graphs, and not necessarily on the knowledge of the material in the questions. The questions acted as a basis for different reasoning structures that are important to physics and math such as identification of functions, discontinuity, implications of the slope, maxima, or other prominent features.

The level of difficulty of the graph questions was gauged to the target population for which the graphing method was used. As the TRIANGLE program was designed for college level use, appropriate questions centered on introductory college level math and physics. The population for the study was drawn from subjects who had taken, or were in the process of taking college level physics courses.

One difficulty of this study is gauging subject involvement when answering the questions. Since the subjects recruited in this study were all volunteers, and no incentive for their performance level could be applied, there is no guarantee that the subjects performed at their best level when answering the questions. However, since subjects were randomly chosen from the same population, on average, any performance issues should be the same for each group of subjects. In addition, the results of the test can be adjusted for random guessing which should reduce the effect of any student apathy towards the test.

6.3.3. General Data Collection Procedure

The specific methods used to collect data varied between the Triangle Pilot, the Web Pilot, the Main Auditory Graph, and the Auditory Preference Pilot tests. The methods are fully developed in the chapters relating to each test. The first two pilot test studies investigated the test environment and development of test questions used in the Main Auditory Graph test. The general process of data collection in the first three studies consisted of giving each subject a statement of informed consent to read and agree to, a survey questionnaire (Survey) for demographic purposes, a pre-test to assess equivalence among the three groups (Pre-test), and a number of questions consisting of one or more randomly assigned graph types (Main test). Recruitment of subjects involved soliciting various instructors to volunteer their classes. The Auditory Preference Pilot test differed

as subjects taking the test were not assigned into graph type groups, there was no Pre-test, and subject recruitment was based on convenience.

The informed consent page consisted of a statement of the test procedure that was involved, the names of the principal researchers and contact numbers, and an agreement clause. This page was required by the Institutional Review Board as human subjects were involved. The Survey questionnaire was used to gather data such as gender, age, and the number of physics, math, or other courses relating to graphical information that subjects had taken. This page also queried whether the subject had musical training or any vision or hearing difficulties.

The Pre-test consisted of a total of five questions about two graphs, four questions for the first graph and one question for the second graph. The first four questions asked for the number of local maxima, the location of maximum slope, etc. and were used to determine whether the subject could properly read a graph. The last question was similar to the questions used in the Main test and concerned the interpretation of the physics described by a graph.

The Main test was presented in different manners depending on the study. For the Triangle Pilot test, the subjects were presented with multiple-choice questions on a computer screen, and either listened to and/or looked at a graph that the question related to. Subjects' answers were recorded in a written format. Assignment of the graph presentation method was random, with the subject receiving a single method (visual, auditory, or both visual and auditory) for all questions. For the Web Pilot test, the subjects accessed a series of Web pages that presented the graph and multiple-choice question, with one question per Web page. Answers were transmitted by selection of multiple-choice "radio buttons" and the answer was recorded by a scripting program. For the Main Auditory Graph test, the same presentation and recording method was utilized as for the Web Pilot, although the number and type of questions were modified and extended due to reliability and validity issues.

6.3.4. Testing Considerations

There is the possibility that a difference in performance levels between visual and auditory graph groups does not necessarily demonstrate an inability of subjects to understand and interpret the presented material. Instead, the difference could be attributable to training and familiarity effects. It was assumed that since the subjects were drawn from standard physics courses, they had been exposed to many visual graphs in the course of their studies. It was also assumed that the subjects had been exposed to virtually no auditory graphs as this is a new representational method. Thus, it was assumed that subjects had much more experience with visual graphs than with the auditory graph representation. Some explanation and training for the auditory graph representation was necessary, but the amount of required training remains undetermined.

The issue of performance effects due to the familiarity of graphs and subject material can be addressed by comparing the subjects' results with those from more experienced graph users. By looking at how well a group of experts (physics graduate students) perform on the questions when given auditory graphs, it can be demonstrated whether the questions are answerable, and what would be the best expected outcome for the groups. Any issues of unfamiliarity with the testing material can be eliminated, thus focusing only on the difference in the graph styles.

There are several reasons why expert subjects were not solely used for these experiments. First was the issue of the audience that the auditory graph representation is trying to target. This graphing method is envisioned to be used as a common tool to help students understand basic data graphs. As such, it is important to discover whether beginning students can understand these graphs with little training and experience.

Another issue was the number of subjects required for meaningful statistical results. Given a normal population distribution, for the 95% probability level, the approximate size of a group required for a 95% chance that the average measurement, $\bar{X}$, is within the limits of $\mu \pm \sigma$ is given by:

$$n = \left(1.96\frac{\sigma}{L}\right)^2 \tag{6.1}$$

where σ is the target population mean and n is the sample size [Sne89, p. 52]. Now, assuming a standard deviation of 20%, since there are 5 possible answers to each question, and an error limit of 5%,

$$n = \left(1.96\frac{0.2}{0.05}\right)^2 = 62. \tag{6.2}$$

Thus, each graph test group should have a minimum of 62 subjects.

The prime target audience of the auditory graph representation is blind users. While it would be desirable to use 62 blind first-year physics students to gauge their ability to answer the questions using the auditory graphing technique and compare their results to sighted users of equivalent background, this is not possible. There are extremely few blind students meeting the conditions of ever having had physics at the college level, even on a national scale. Blind people who have completed physics courses were solicited for their participation via requests on electronic mailing lists. However, only a very small number of people (five) participated. This will be discussed in more depth in chapter 9.

In any test, there is a question of whether the test is reliable and valid. Validity of these tests was determined by review of the questions with experts, and by comparing test results between first-year and graduate students. By designing the tests so that questions can be divided for split-half analysis, a statement about the test's reliability can be made. If there is a high degree of correlation between the scores in the two halves, then there is a greater probability that similar questions will have similar results. This helps to indicate how well subjects can reliably use the auditory graphs to answer questions.

One difficulty with the testing process used in this study that should be noted was the high reliance on technology. While this posed certain challenges, the technological problems affected all subjects equally in the Web-based tests. In the preliminary Triangle Pilot, an instrument method that was not technologically dependent was utilized for initial comparative purposes.

It is possible that a better scheme for testing and producing auditory graphs can be developed. The Main Auditory Graph test was an evolution of the processes used in the

Triangle Pilot and Web Pilot tests. As was previously stated, the auditory methods used in this study were chosen to a large extent by results from previous research, similarity to musical representations, and prior device development.

Research into better graphing techniques is necessary and is the issue of further studies. Some indications of possible graph questions as well as alternate auditory display methods are demonstrated in the Auditory Preference Pilot test discussed in chapter 10.

7. TRIANGLE PILOT

7.1. Overview

The first experiment conducted was a pilot test to investigate the advantages, difficulties, and question layout of a study involving auditory graphs. This experiment, named the Triangle Pilot, used the TRIANGLE program to display the questions and the visual and auditory graphs to a majority of the subjects. The results from this experiment not only helped elucidate several inadequacies in the production and testing of the auditory graphs but also showed that there were no insurmountable difficulties with the auditory technique. The Triangle Pilot provided the basis material that was used in later studies.

This experiment consisted of three instruments: an initial Survey questionnaire, a Pre-test, and a Main test. The purpose of the questionnaire was to provide basic demographic and other relevant information to aid in analysis of the responses. The Pre-test consisted of five questions to check subject understanding of basic graph concepts. The Pre-test was given in a printed form, and consisted of labeled graphs that subjects could easily identify. The Main test consisted of 14 multiple-choice questions. Additionally, there were fill-in-the-blank supplements for two of the multiple-choice questions. The questions were designed to be equally valid for either visual or sonified displays. Appendix A contains a copy of the Survey (A.3), Pre-test (A.4), and Main test (A.5).

Subject matter for the questions centered on previously published research involving graphs and physics. Most notably, questions from the Force Concept Inventory (FCI) [Hes92a], Mechanics Baseline [Hes92b] test, and the Beichner [Bei94] study were used after some modifications. Other questions were developed after an analysis of subject matter presented in several introductory physics text books. The final questions were reviewed for content validity by several physics and science education faculty known for their interest and excellence in teaching at Oregon State University.

There were four treatment methods for this study: graphs visually presented on paper, graphs visually presented on the computer, auditory graphs presented on the computer, and both auditory and visual graphs presented on the computer. The paper presentation method was to check for any novelty effects that the TRIANGLE program might introduce. The presentation method with both sound and picture graphs was to check for any increase in students' ability to answer questions due to multi-modal presentation.

All materials used in this study were submitted to the OSU Institutional Review Board (IRB) for review and approval. After receiving endorsement by the IRB, subjects were solicited for participation in the study. Two graduate students participated using auditory graphs for purposes of testing validity, and for estimation of time allotment for scheduling purposes.

7.2. Sample

Ideally, a random sampling from a wide variety of first-year physics students would be desirable. However, this was not possible due the scope of the pilot study. Since the Triangle Pilot was intended to determine the feasibility of a study on auditory graphs, it was decided to limit participation to local students. The use of local students was a matter of convenience and limited the generalizablity of the study. The Main Auditory Graph test included subjects from several educational institutions to allow for greater generalization of the studies' results. As the testing process was designed for first-year physics students, instructors of these courses were solicited for the possibility of letting their students participate in this study.

The sample was drawn from an introductory, algebra-based physics course at OSU during the 1997 summer-session. It was arranged with the professor of this course that the investigator would ask for student volunteers from the course's laboratory component. Subjects participated in the study during the same time as their normally assigned laboratory section. For their participation, students received full credit for the missed class. Names of volunteers were taken from each of three laboratory sections that met on the same day, with 43 of approximately 60 students volunteering. While some of

the volunteers may have chosen to participate due to a higher motivational level, the offered incentive attracted many of the volunteers.

Due to time constraints and resources, the study limited the focus to twelve students. The number of subjects was chosen due to the number of students who could be tested by one researcher during the three two-hour laboratory sections on a single day. The time for completion of the questions for the test was estimated to be half an hour as this was approximately the time taken by graduate student volunteers on a previous day. A list was formed of the volunteers from each laboratory section and a computer randomly selected four subjects from each laboratory section for a total of 12 subjects. The chosen student volunteers were taken from their next lab session and came to a designated room at assigned 1/2-hour intervals.

7.3. Data Collection

Data were collected through a guided interview process. The interviewer served as a guide to answer general questions, such as those arising from ambiguous wording or instructions, and set up the questions and graphs on the computer. This last step was necessary as the TRIANGLE program was not designed as a testing environment. Subject volunteers selected from a random computer-generated list met with the interviewer at an appointed time and place. Subjects were randomly assigned to one of the four graph category groups (three subjects per group): visual graphs printed on paper (Print group), visual graphs displayed on the computer (Visual group), auditory graphs produced by the computer (Sound group), or both visual and auditory computer graphs (Both group). Subjects were shown each of the questions on the computer, except for the print group which had questions on paper, and given as much time as they wanted to answer the question before proceeding to the next question.

The testing area consisted of a room with a large table upon which a computer was placed, and several chairs at the table. A video camera recorded the sessions, and all subjects had been questioned and gave their consent to being videotaped; the subjects where shown where the camera was located. Upon entering the room each subject recorded his or her name on a log page and was given a Document of Informed Consent

(Appendix A.2) to read and agree to. They were next presented with the Survey and the Pre-Test. After completing the initial questions, the subjects were given an answer sheet (Appendix A.6) to record their responses. For each subject, the Survey and answer sheet were marked with a unique code for identification purposes and for recording the type of graphs on the test.

Each subject was given one of the four graph formats for the Main test. The order of the type of test given was changed between groups of students. The listing is given in Table 7.1 where the representation is P for the test given on paper (Print), V for the test displayed on the computer (Visual), S for the test presented on the computer with auditory graphs (Sound), and B for the test presented on the computer with both auditory and visual graphs (Both).

Table 7.1 Test Type per Interview Time.

Student :	1	2	3	4
Group 1: 10-12 am	S	V	B	P
Group 2: 1-3 pm	B	V	P	S
Group 3: 6-8 pm	P	S	V	B

In cases where the computer was used (S, V, and B groups) the investigator changed the displayed question and graph after the subject had finished with the previous question. In the studies with auditory graphs, subjects had control of the graph playback via the computer's keyboard. Subjects were allowed as much time as they wanted to study and listen to the graphs and to answer the questions. Subjects were also allowed to return to previous questions if they wanted to change their answers. The interview process was videotaped for later study, most notably to check for leading by the interviewer.

7.4. Instrument Development

The test questions and graphs were displayed with the TRIANGLE program. This is a DOS-based program developed by the Science Access Project at OSU. This program has a text region where the questions can be displayed, as well as a display for visual and auditory graphs that can be generated from a table of data points and then plotted on the screen. While viewing the graph, a user can also listen to a sound representation (sonification) of that graph. In the case of the TRIANGLE program, sonification of the data was represented with a linear relationship of pitch to the *y* axis data values. The *x* axis values were converted to time, so that the graph was played from left to right. In addition, data points were located in space by stereo speakers so that the sound panned from left to right.

The resulting auditory graphs could be played either continuously, or by stepping through data points with keys on the computer's keyboard. Subjects in the test groups that used sound graphs were given a brief description of the auditory graphs but no specific training was performed. Screen images of the TRIANGLE display can be found in Appendix A.7.

Questions for the study's Main test proved to be a challenge to create. From a review of previous research, it was decided that a multiple choice question format would provide useful information for determining the effectiveness of auditory graphs. The primary difficulty in question development was creating multiple-choice questions that would reference a single graph. In addition, the graphs needed to be comprehensible displays both as visual pictures as well as sonified data sets. The TRIANGLE program imposed a further limitation on the auditory graphs because at the time of the study, there was no method for describing negative values in an auditory format. Therefore, graphs used for the questions could reference only positive *y* axis values.

To provide testing situations that were as nearly identical to each other as possible, there was no difference in the information contained in each of the graph types or in the question wording. For example, title and axis representations were mentioned explicitly in the question text, rather than on the graph's axes, as the auditory graphs had

no labeling method. Each graph was displayed separately from the question text, although this was partially an aspect of the program used to display the graphs.

Only simple graphical information was portrayed in the questions because previous studies had not determined the effectiveness of interpretation of auditory graphs. The restrictions placed on the development of questions and investigation of introductory texts and previous studies provided material for 14 questions. These questions underwent review by graduate physics students and professors in physics and science education at OSU.

The wording of the individual questions was designed to provide correct and clear distinction between choices with an emphasis on drawing conclusions from the information displayed in the graphs and not primarily on their background knowledge of physics. Two short-answer questions were included to probe their understanding of more complex physical issues (11b and 12b) but these were not the primary focus of the test. Answers to the multiple-choice questions were evenly distributed among five choices (A, B, C, D, and E). The answer sheet was developed to provide a consistent method for the subjects to record their answers. Extra space on the answer sheet allowed subjects to write any additional comments or questions about the wording of the test questions. A copy of the answer sheet is located in Appendix A.6.

To check reliability, the test questions were constructed to be applicable for split-half analysis. Table 7.2 displays the correspondence between the graph type and the question number. Test splitting was for similar graph type rather than for questions concerning similar physical phenomenon.

Table 7.2: Distribution of Graph Types.

Graph Type	Question Numbers
Linear: Constant	1, 4
Linear: Increasing	2, 7
Linear: Decreasing	5, 8
Segmented: Linear	3, 6
Segmented: $1/x^2$	11, 12
Nonlinear: x^2	9, 13
Nonlinear: Root, $1/x$	10, 14

The graphs used in the Pre-test and the Main test were created in a multi-step process. First the physics principle investigated was modeled by an equation or segmented graph. A small program was developed to aid in creating a two column table of numbers that represented the desired graphs. Each graph had 100 data points, as the TRIANGLE program could create an auditory graph lasting approximately three seconds with that many data points. Each table was imported into Microsoft Excel for collation into a larger table so that each question was represented by one column of data. The resulting table was converted into a format for use by the TRIANGLE program. Each column of numbers was plotted at the time that the corresponding question was asked.

The necessity of using an interview process arose from the difficulty in learning and using the TRIANGLE program to display questions and graphs. While the auditory display was straightforward to use once the data had been loaded, the process of loading and manipulating the data could have interfered with the interpretation of the graph. Thus, the interviewer was responsible for displaying the data so that the subjects needed only to be concerned with interpretations derived from the display methods.

7.5. Data Results

Table 7.3 is a summary of the results contained in Appendix A.8. The table is divided by results from the different test groups.

Table 7.3. Percent of Subjects Answering Given Questions Correctly.

	% Correct				
Question	Print	Visual	Both	Sound	Grad
Pre-test:					
P1	100%	100%	100%	100%	100%
P2	100%	100%	100%	100%	100%
P3	100%	100%	100%	67%	100%
P4	67%	100%	100%	67%	100%
P5	0%	33%	0%	100%	100%
Average	73%	87%	80%	87%	
Main Test					
Q 1	67%	67%	0%	0%	100%
Q 2	33%	33%	33%	0%	100%
Q 3	0%	33%	0%	0%	100%
Q 4	0%	67%	100%	67%	100%
Q 5	0%	33%	0%	33%	100%
Q 6	33%	67%	67%	67%	100%
Q 7	67%	100%	100%	33%	100%
Q 8	67%	100%	100%	100%	100%
Q 9	100%	100%	100%	33%	50%
Q 10	67%	67%	33%	0%	100%
Q 11	100%	67%	67%	67%	100%
Q 12	67%	100%	67%	67%	100%
Q 13	100%	33%	67%	0%	100%
Q 14	100%	67%	67%	0%	50%
Average:	57%	67%	57%	33%	
Standard Dev.	38%	27%	38%	34%	

While the summary table provides an accurate listing of the data, it is helpful to view the same data as a bar chart to recognize patterns in the data and to easily see where any difficulties may lie. The following chart displays the percent correct scores of each test group vs. the individual test questions.

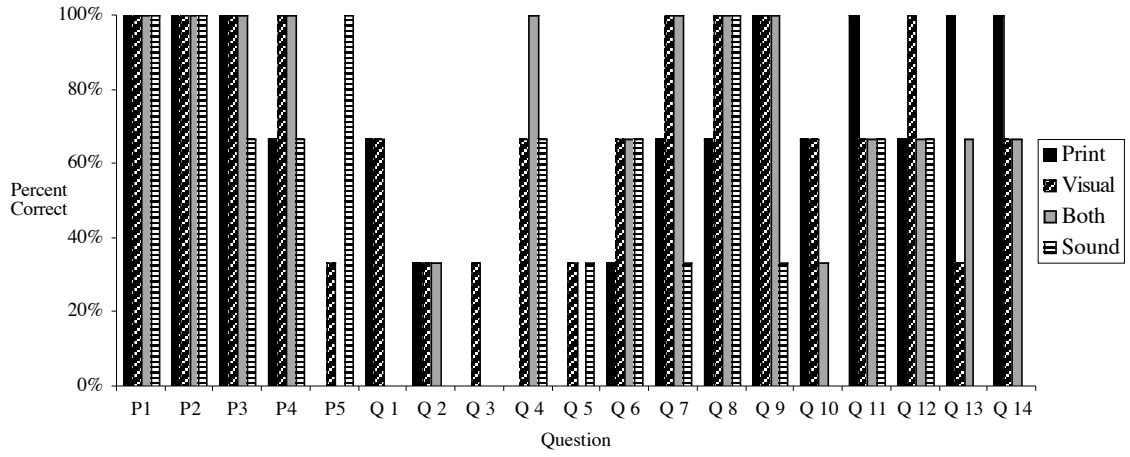

Figure 7.1 Comparison of Triangle Pilot Test Results for Different Presentation Methods.

Three subjects in each of the four groups is far below the required number to produce valid statistical analysis. However, keeping in mind that errors are greatly exaggerated, the pilot test leads to a number of insights. The first point to be noted is the striking difference between the Sound and Visual groups for Main test questions 1, 2, 3, 10, 13, and 14. These questions are reviewed in section 7.6.

The poor performance of the Sound group on a large number of questions gives an early indication that there may have been an oversight in the method of auditory graph production that was used. A dramatically lower score from the Sound group indicates that either the auditory graphs were not properly explained and understood, or that there was an fundamental display problem that prevented subjects from fully understanding the graphs. The two graduate students who participated using auditory graphs had perfect scores except for questions 9 and 14. These questions may not be valid, or the auditory representation is not adequate even for experts.

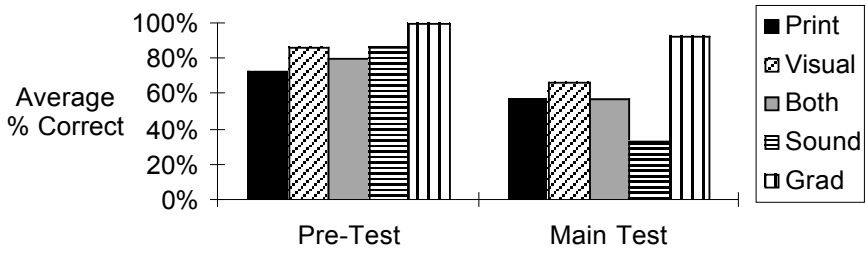

Figure 7.2. Average Scores of Groups on the Triangle Pilot Test.

The mean score on the Main test for all the undergraduate respondents was 53% answering correctly. While this may seem like a satisfactory value as it is in the center of the distribution, the average standard deviation for the student groups was a very large 34%. The large standard deviation is in part attributable to the small sample size, but may also be attributable to poor wording of the questions or lack of knowledge by the respondents. Analysis of individual questions is useful for determining any specific problems and is covered in section 7.6.

The average score for the Visual group was twice that of the Sound group. Although this difference may be attributable to random variation, it is more likely that it is due to subject difficulty with the auditory graphs. A revised auditory graphing method was used for later tests.

Table 7.4 Split-Half Analysis of Test Questions.

Question	% Correct	Split Question	% Correct
1	33	4	58
2	25	7	75
3	8	6	58
5	17	8	92
9	75	13	50
10	42	14	58
11	75	12	75

To comment further on the validity of the test questions in general, analysis of the question types is necessary. The percent of correct answers in question 11 is within one standard deviation of the score for its split question (12). The same is true for question pairs 1-4, 10-14, and 9-13. Thus, 4/7 of the pairs have differences within one standard deviation. When looking at two standard deviations, question pairs 2-7 and 3-6 are also included bringing the total to 6/7. These are reasonably close to the ideal limits of 0.66 for one standard deviation and 0.95 for two.

However, split-half analysis to check the reliability of problems of similar graph types generally shows a disappointing relationship. The correlation coefficient between the two groupings yields $r = -0.28$ which is very poor in comparison to the ideal of 1. The low r value indicates that either the question wording is ambiguous and needs to be rewritten, or that similar graph types do not lead to similar response rates. If the latter is the case, then the question's material plays a greater role than the graph in the subject's understanding of the graph.

7.6. Detailed Analysis of Selected Questions

The following is a commentary on the answers arising from the Survey portion of the pilot test. The Survey consisted of eight items that were a combination of multiple-choice and short answer questions. While the results of this Survey were not used for analysis purposes due to the small number of subjects, it was instructive to find where ambiguities lay. Refined versions of the Survey were used in later experiments such as the Web Pilot and the Main Auditory Graph tests.

Survey questions 1, 2, and 3 concerned general demographics of gender, age and high school physics experience. There was some confusion with question 4: *Number of years of college-level physics?* by the first-year students who took the test. They often asked if they should circle the 0 (because they hadn't completed a year yet) or 1 (as they were currently taking a first-year course.) This difficulty was corrected in later tests by restating the question as: *How many courses of college-level physics have you completed?*

Survey questions 5 and 6 asked subjects to state courses that they felt had been helpful in understanding graphical information. Responses often only stated course numbers which were difficult to decipher. These questions were rewritten for later studies.

Survey question 7: *Have you learned graphing techniques other than from academic settings?* was confusing to a large number of subjects. During the Triangle Pilot test, subjects were given examples as to what type of answer was being sought which tended to ease the confusion. This question was rewritten in later studies. There did not seem to be any difficulties with questions 8 or 9 which related to musical training or physical difficulties.

The Pre-test questions were interesting. It was expected that questions 1 to 4 would be answered correctly by everyone. They were designed to test the ability of the subjects to simply read a graph and locate points. Three of 14 subjects missed one of the questions and one subject missed two questions. This leads one to conclude that there may be a 9% error rate on the test due to students misreading, or simply not being careful with, the questions.

Question 5 on the Pre-test was designed to be similar to one of the moderately difficult graphs. It was a nonlinear graph referring to motion of an object. The number of correct responses was dramatically lower than the previous questions at 33% correct. This question has content validity as the graduate students had a 100% correct response. It should be noted that all but one of the subjects who had correctly responded to this question were randomly assigned to the auditory graph test group.

The text and answer choices for the questions in the Main test tended to be more complex. A full listing can be found in Appendix A.5, so only a brief description of the graph or changes to be made to the questions is described below. A listing of the percentage of correct responses per group is provided for each question. The categories are: the total results for the undergraduate test subjects (Total), the printed test group (Print), the group that only saw the test and graph displayed on the computer screen (Vision), the group that both saw and heard the graphs (Both), the group that only heard the sound graphs (Sound), and the graduate student subjects (Grad). The graduate students are used for validity comparison.

An important point to be stressed with these results is that the undergraduate subjects were divided into four groups, so that each group only had three test subjects, and that only two graduate student subjects participated in the test. Statistical fluctuations could account for many irregularities in the results as one incorrect response would manifest in a change of 33% for any given group's correct response rate for a question. A larger number of test subjects reduces the ambiguity and is demonstrated in the Web Pilot test described in chapter 8.

Main test question 1 was one of the more surprising results. The graph was designed to be the easiest to recognize, a straight flat line, and have common axes (distance, time) yet most students were not able to answer this question correctly. The averages for the Print and Visual group were equal whereas the subjects in the Sound and Both groups all missed this question. The difference, especially for the Both group could be attributable to a novelty effect of hearing the sound display and unfamiliarity with what the display was representing.

The auditory graph in this case was a constant tone. A training period for the sound graphs seems to be necessary for less experienced students, especially when considering the results from the next two questions. It was noted that answer *A) The object is moving with a constant non-zero acceleration* and *C) The object is moving with a uniformly increasing velocity* are the same. The second answer was changed to *C) The object is moving with a uniformly decreasing velocity* for use in the Web Pilot test.

Question 2 was again intended to be fairly straightforward. The Grad response shows that experienced graph readers using the auditory display can understand this question. The equality between Print, Vision, and Both groups shows that displaying the graphs on the computer may not be a significant effect. The low score for Sound, may be from random variations, or as a result of unfamiliarity with the auditory graph representation. The same change to answer C was later made on this question for the same reason as with question 1.

The results of question 3 were particularly interesting as this question was modeled on a similar question used in the study by Beichner [Bei94]. In that study, there was a 33% correct response rate whereas this pilot test had an 8% rate. Again, the Sound group completely missed this question. The results may not be as significant as the other

groups also did very poorly. It is possible that the entire question should be rewritten. However, the Grad subjects found the question legitimate. Answer *B) The object doesn't move at first. Then it rolls down a hill and finally stops.* may have caused some confusion as it could possibly be construed as correct. A change *to B) The object doesn't move at first, then it moves away from the reference point, and finally stops.* may correct any misunderstanding.

Question 4 gives some indication that auditory graphs may be a valid form of data representation as the Sound group score was as high as the Vision group and the Both was higher still. However, the dramatic difference between Print and Vision scores may indicate that random fluctuations are greater than the one standard deviation previously mentioned. This question was designed to be a complement to question 1, and the Vision group did have the same score. Groups using sound had improved scores, perhaps indicating that a training period had taken place. Answer C should probably be modified as in Question 1.

The generally poor results on this question 5 can be attributed to the wording of the answers, which while correct, were meant to distinguish amoung subjects who had a good understanding of the concept of acceleration. Subjects chose only one of two answers: A) *The object is moving with a constant acceleration.* and B) *The object is moving with a decreasing acceleration.* The graph showed a linearly decreasing velocity, hence a constant, albeit negative, acceleration. Experienced subjects did not have difficulty with this question.

Subjects performed generally well on question 6 which had one of the most complex graphs. This question paralleled one of the Pre-test questions. As can be seen, the Sound group did at least as well as the other groups.

It is not understood why the Sound group did considerably worse on question 7 than the other groups, except that this question is the same graph type as question 2 and the results were also poor on that question. There could be a difficulty in recognizing the pitch, and hence the y value, as representing a linear increase.

Questions 7 and 8 were related by the physics principles that they questioned. Hence the similarity in the majority of the scores is not unexpected. The Sound group shows a dramatic change. This difference could be attributed to statistical fluctuations.

There is also a possibility that auditory graphs that start with a high pitch and end with a low pitch may be more easily distinguished than in those that start low and end high.

Question 9 gives a clear indication that curved graphs are not well perceived in this form of data sonification. Not only did the Sound group do poorly, but the graduate students also noted that they could not tell the correct answer from the sound graph. The graduate students mentioned that they knew what the one correct answer was but answered according to what they heard.

The results of question 10 are interesting in that subjects only chose one of three answers: the correct response of 1/wavelength, the incorrect response of decreasing linearly with wavelength, or the answer of: not related to wavelength. The second answer may have been the result of confusion with the wording; this answer should be rephrased to *B) The frequency has a constant, linear decrease as the wavelength increases.* While the question appears to be generally valid, the low scores of both groups utilizing sound indicates that a better sonification technique is necessary for less experienced students.

Subjects did well on the similar questions of 11 and 12. These questions involved graphs that were more complex. Answers tended to be between those that were most similarly related to the graphs, perhaps indicating that better distracters need to be constructed. The Sound group performed as well as the other groups on these questions, indicating that more complex auditory graphs can be used.

There were second parts to these questions, *11b* and *12b. What does the peak on this graph represent?* The answers tended to be statements of the graph, rather than its physical interpretation, indicating poor question wording. This type of questioning was dropped in later experiments.

The large difference in scores between the Print and Vision groups in question 13 may be due to random fluctuations. Again, the Sound group shows that this sonification method was not adequate for curved graphs, and the Grad results also back this statement.

For question 14 the problem seems to be not in the question, but in the sonification of the graph. These results are perhaps the clearest indication that the method of sonification used in the Triangle Pilot was not useful for simple curved graphs.

7.7. Conclusions From the Triangle Pilot Test

The guided interview process for the test was necessary so that the interviewer could answer questions about the test, set up questions and graphs on the computer display, and observe if there were any particular difficulties with the test. One subject noted that the proximity of the interviewer was uncomfortable in a performance-type setting. The proximity issue was immediately resolved, but question set-up became more time consuming. The Web Pilot test removed this issue by providing a self-running testing environment that could be accessed from remote locations and did not require the intervention of an experienced user for setting up test questions.

Subjects were able to answer all the test questions in the allotted time. The assumed time to complete all test parts was about 30 minutes, or about one question per minute. This was a reasonable guess as many of the problems were conceptual, multiple-choice questions that did not involve calculations. Subjects in the Print group tended to finish the Main test in less time than did the others due to not having to wait for the investigator set-up the computer display for the each problem.

While a cursory review of the data shows that the Sound group performed substantially below the level of the other groups, this problem may be able to be overcome. First, looking at the questions where the Sound group performed as well as the others and comparing the question types to those where they did not perform well provides important clues as to better sonification methods. It appears that subjects were able to distinguish absolute values by sound (i.e. pitch being higher or lower). Also, the subjects were able to identify the first derivative of the function (is it increasing or decreasing) but not the second derivative (the rate at which a function is increasing or decreasing). These conclusions are shown by the poor results on any of the graphs portraying curved functions. The more experienced subjects (graduate students) were able to interpret the shape of the graph from the more limited information. This came about because they immediately converted the sound into a picture. Even with this experience, they still had difficulty interpreting graphs that had a positive curvature.

A solution to the problem of identifying curved graphs is to enhance the derivative information. One method is to add sonic indicators for the first and second

derivatives such as with a series of clicking noises, where the rate (tempo of the clicks) represents the first derivative. The tempo is set by how often the curve crosses some *y* axis interval. The pitch of the clicks can indicate the second derivative. This method was used in the auditory graphs for the Web Pilot, Main Auditory Graph, and Auditory Preference Pilot tests.

While the Print group was useful for comparison, and a slight difference in the scores was noticed, the difference did not seem to be significant. This result indicates that this fourth grouping is not necessary. Remote computer administration of the test, and greater subject participation, was accomplished more easily without this unnecessary group. With fewer testing sections, the group sizes were larger for a given number of subjects resulting in greater confidence limits.

Lastly, it was evident that the original hypothesis of the equivalence between simple graphs produced with TRIANGLE's basic sonification technique and visual graphs was not realized for a significant part of this test. One factor for these results was the lack of training that subjects in the Sound group encountered before the test. While some training is evidently necessary, the goal of this auditory display is to have a method that is reasonably intuitive. Subject performance for the Sound group generally seemed to increase up to question 7, thus providing an initial estimate for the number of graphs for training. This pilot test demonstrated the need for first and second derivative information to be incorporated into auditory displays in order to increase the distinction between curved and linear graphs. The modifications to the questions, initial information, and display formats formed the basis for the second pilot test called the Web Pilot.

8. WEB PILOT

8.1. Overview

The second experiment conducted was a pilot test to investigate the advantages, difficulties, and question layout of a study involving auditory graphs using the World Wide Web (Web) as a testing environment. The Triangle Pilot test suggested that there would be logistical difficulties when having a large number of subjects take a test with the computer generated auditory graphs used in the pilot. Also, a more flexible testing environment was necessary than that provided in the Triangle Pilot. It was suggested to the author that a Web-based test could overcome these difficulties and provide many advantages. Such a test would allow access by many student subjects as well as provide a flexible testing environment. Pictures, sounds, and text could be easily configured and changed, and multiple graphs could be displayed with little effort. In addition, it would be possible to record subjects' responses with Web-based forms [Ceb97].

This experiment, named the Web Pilot, used a standard browser program, such as Netscape or Microsoft's Internet Explorer, to display the introductory materials, questions, and visual and auditory graphs in a series of Web pages. The results from this experiment helped show where revisions were needed when creating a test with this new medium. The Triangle Pilot provided the basis material for this experiment; the Web Pilot provided the testing technique used in the Main Auditory Graph and Auditory Preference Pilot studies.

8.2. Sample

As the testing process was designed for introductory physics students, instructors of these courses were solicited during the Fall 1997 quarter for the possibility of letting their students participate in this study. It was arranged with one instructor of an introductory algebra-based physics course at OSU to provide extra-credit homework

points to students taking Web Pilot test. The instructor announced the location of the Web Pilot test's introductory Web page in class and posted a link on the course's information Web page. Student volunteers were given one week after the initial announcement to complete the test.

From this single course, 221 out of about 400 enrolled students completed the Web Pilot test. At most, six students who logged into the test, due to technical difficulties or lack of interest, did not complete all of the questions. Only subjects completing all questions had their data recorded. Of the 221 recorded subjects, 74 subjects received the Main test with auditory graphs, 75 received visual graphs, and 72 received both auditory and visual graphs. These numbers allow for statistically significant results at the $p = 0.05$ level since $n \geq 62$ (from equation 6.2) for each group.

8.3. Data Collection

Subject volunteers accessed the Web Pilot test site from remote computers at various locations at Oregon State University. Several PERL scripting programs recorded data from subject responses to questions presented on the Web pages. The data were written to secure files. The PERL scripting programs can be found in Appendix F.

When subjects accessed the Web address announced in their class, they were presented with a welcoming page stating the purpose of the test. The welcoming page also contained a brief description of auditory graphs, a link to a Web page containing further descriptions and examples of auditory graphs, a copy of the informed consent document, and a link to the test. A copy of this page is located in Appendix B.2.

After the introductory Web page, subjects were presented with a Web page to record their name and a school class code into text entry form fields. The names, class code, and an identification (ID) code number were appended by a PERL script program called "namepage" to a secure file that contained previous subject's names. This program also randomly assigned the subject to one of the graph test groupings, labeled by b, s, or v, and then passed the ID and graph codes to the next Web page.

The use of the ID code number provided anonymity of the final results (so that names would not appear with the test scores), yet allowed the investigator to identify

students so that multiple attempts of the test could be eliminated. Also, the code provided a record of which students from a given class completed the test, and allowed for comparison of results between tests. Due to security, anonymity, and coding issues, subject's names were not written to the test's Web pages.

The Survey Web page contained text entry fields as well as radio-button type choice fields. A second PERL program called "surveyrecord" appended the subjects ID code number and any long text answers to a separate file when subjects chose the "Next page" button. The ID code, graph code, and several of the Survey answers were passed as a text string to the Pre-test page. The graph code was also passed as a variable to later pages.

A third PERL program called "prerecord" added the Pre-test page answers to the text string when subjects chose the "Next page" button. This program passed the answer string, along with variables for the graph code and the "start time" of the test, to the next Web page. The time that subjects took to answer the test was measured to provide some insight as to how long students took with the different presentation methods. The prerecord program generated the first question page for the Main test.

A fourth PERL program called "temprecord" generated subsequent Web pages for the Main test. The questions were read from individual files, and contained multiple-choice, radio-button style answer selections. Graph codes, previous answers, and starting time information were written to the generated Web pages. The graph presentation was determined from the graph code value and incorporated into the pages as well. When subjects chose the "Next" button, their answer for the question was added to the answer string and the next question was read from a predetermined file. When the subjects had completed the last question, the temprecord program calculated the total time, added this information to the answer string, and appended the string to a secure file of previous subjects' answers.

8.4. Instrument Development

As in the prior experiment, there were three instruments: an initial survey questionnaire (Survey), a Pre-test, and a Main test. The content and subject matter of the

Survey, Pre-test, and Main test were similar to those of the Triangle Pilot, but with the revisions as noted in the previous experiment. The presentation was through a linear series of Web pages. A copies of the Web Pilot Survey is located in Appendix B.3, the Pre-test is in Appendix B.4. A screen image of a typical question for the Main test is located in Appendix B.5. The questions for the Main test were the same as those in the Triangle Pilot test but with the revisions noted in chapter 7. There were three display methods for this study: visual graphs, auditory graphs, and both auditory and visual graphs. There was no paper presentation method as was performed in the Triangle Pilot due to that test's similarity in scores between the visual and paper presentation methods, and due to logistical difficulties.

The challenge of this experiment was to convert the testing process of the Triangle Pilot to a Web-based format. Methods to display the questions' text and visual graphs were not difficult as the standard Web browser has this ability built into the display. The method of producing auditory graphs that could be played from the browser window was more problematic. The difficulty lies in the ability of computers to produce sound from various audio file formats. One of the most common and useful formats, the Microsoft .wav format, leads to large file sizes (on the order of 100 Kb). Transferring large files on the Web resulted in long delays when displaying auditory graphs, especially if the subject was using lower speed modems to access the test. The use of MIDI reduced the auditory graph file sizes to about 2 Kb, which produced a page that would download and display more rapidly.

The MIDI protocol uses data streams to trigger stored sound-wave patterns on the host computer. Each sound wave represents an instrument, or voice, whose pitch, onset, duration, and decay are triggered by the data. Thus, complex sounds can be reduced to small data files. For the Web Pilot, *y* axis data values were represented with a piano voice that varied in pitch.

One disadvantage of the MIDI format is that sounds are not completely consistent from one computer to another, as each computer may have different stored wave-pattern representations corresponding to a given MIDI voice code. Also, there is an inherent limitation on the resolution of sounds since MIDI uses a chromatic scale as a basis for the

divisions between notes. Producing sounds between given notes greatly expands the file size.

As was shown in the Triangle Pilot, auditory test subjects had great difficulty distinguishing between linear and curved graphs. One of the Triangle Pilot subjects made the suggestion to add tick marks to represent the *y* axis values. This suggestion was incorporated into the new auditory graphs.

When data values passed certain intervals, a tick mark was sounded. The tick mark sound was represented by a drum instrument voice. The resulting frequency, or tempo, of tick marks represented the magnitude of the slope, or first derivative, of the graph at a given point. A small magnitude slope resulted in a slow tempo of the sounding of the tick marks, while a large magnitude slope resulted in a fast tempo. The sign of the slope was easily determined by listening to whether the data value pitch was increasing or decreasing.

While the tempo of the drum beat to indicate slope provided much needed information, the second derivative was also easily incorporated by modifying the pitch of the drum voice. To reduce the auditory load, it was decided to only use three pitches to represent the second derivative, one for negative values, one for positive values, and a third for 0. The optimal choice of pitches is a matter some debate and is the subject of future research.

For this study, the investigator chose to represent negative values of the second derivative with a high drum pitch, positive second derivative values with a low pitch, and 0 was represented with a pitch in between the two. Thus, the graph of $y = x^2$, from 0 to 1, had an increasing piano tone and a low pitch drum that would increase in tempo, and the graph of $y = 1 - x^2$, from 0 to 1, had a decreasing piano tone and a high pitch drum that also increased in tempo. The graph of $y = x$ had an increasing piano tone with a constant tempo drum beat whose pitch was between the high and low drum pitches.

The reasoning for this choice of the tick mark pitches was that, aside from areas with inflection points, negative curvature occurs at local maxima, while positive curvature occurs at local minima. Thus, the tick mark pitch would reinforce the data pitch in those areas.

The auditory graphs used in this study were produced in a multi-step process. The *x-y* data sets used to create the graphs in the Triangle Pilot were converted by the DataSonify program into an SLG formatted text file. SLG is a text file format and is an instruction set that the MIDIGraphy program [Ton99] can import to convert text data to MIDI sound files. DataSonify set the instrument, time duration of the notes, length of the play time of the data set, and calculated and set the drum tick mark derivative information. The MIDI file was converted into the .wav format with SoundMachine [Sou99], and was converted into Apple Computer's QuickTime format with the MoviePlayer program.

The QuickTime sound file format was chosen because it allowed for a Web browser plug-in module that could display embedded play and pause controls in the Web page display. Also, since this module was available for several computer platforms, subjects would have little difficulty locating a computer from which to take the test. To provide alternate access to the sound files, links were included so that subjects could download the MIDI formatted file or the much larger .wav formatted file. The visually presented graphs were produced with the KaleidaGraph program from Synergy Software. These graphs were converted to a .gif file format. For display on the Web pages.

8.5. Data Results

Table 8.1 is a summary of the full results contained in Appendix B.6. The table is divided by results from the different test groups.

Table 8.1 Percent Correct per Question for Each Group.

	% correct per Group		
Question	Visual	Both	Sound
Pre-test			
P1	77%	71%	78%
P2	97%	97%	96%
P3	99%	96%	95%
P4	83%	78%	81%
P5	71%	58%	73%
Avg.	85%	80%	85%
Main Test			
M1	68%	58%	30%
M2	67%	65%	36%
M3	59%	53%	22%
M4	71%	75%	57%
M5	11%	10%	16%
M6	68%	65%	23%
M7	81%	71%	64%
M8	84%	89%	73%
M9	69%	74%	54%
M10	69%	71%	41%
M11	71%	81%	41%
M12	71%	81%	41%
M13	67%	69%	42%
M14	67%	69%	31%
Average	66%	66%	41%

While the summary table provides an accurate listing of the data, it is helpful to view the same data as a bar chart to recognize patterns in the data and to easily see where any difficulties may lie. Figure 8.1 displays the percent correct scores of each test group vs. the individual test questions.

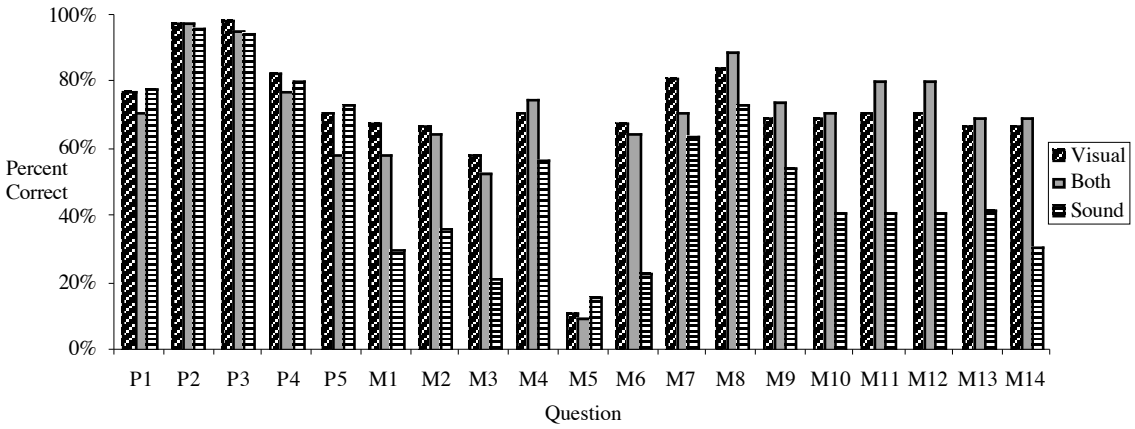

Figure 8.1 Histogram of the Results of Table 8.1: A Comparison of Correct Answers per Group.

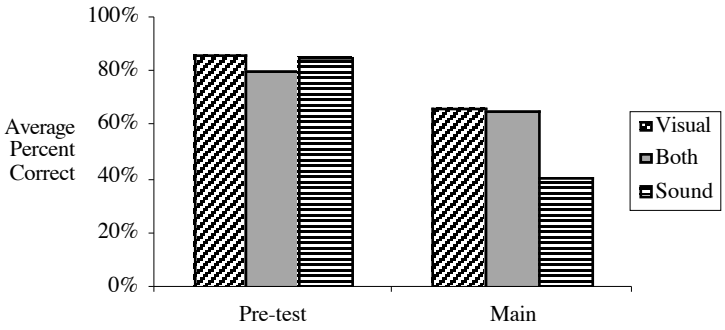

Figure 8.2 Average Scores for the Groups.

It is evident from the displayed results in Figure 8.2 that the Sound group performed at a lower level than did the Visual and Both groups. However, it should be noted that the Sound group for the Web Pilot had a greater percentage correct (40%) than did the Sound group of the Triangle Pilot test (33%).

8.6. Analysis

Analysis of the Pre-test was performed using Microsoft Excel on the data and is displayed in Table 8.2. The results are from a single valued analysis of variance (ANOVA) test at the $\alpha = 0.05$ probability level comparing the Sound, Visual and Both groups.

Table 8.2 Pre-test ANOVA Analysis.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Both	72	288	4.00	0.96
Sound	74	313	4.23	0.92
Vision	75	320	4.27	0.79

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	$F_{critical}$
Between Groups	3.04	2	1.52	1.71	0.18	3.04
Within Groups	193.76	218	0.89			
Total	196.81	220				

Since $F < F_{critical}$, H_0 is valid. No significant difference between groups.

The numbers in the *SS* column represent the sum of the squares of deviations for all data. There are two degrees of freedom (*df*). The degrees of freedom value are one less than the number of items being compared. Since there are three groups to compare, there are two degrees of freedom between the groups. The number of degrees of freedom when looking for variations within all of the groups reduces the number of subjects in each group by one and then sums the resulting values. *MS* represents the mean squares and is simply the sum of squares divided by the degrees of freedom. The *F* value is calculated by dividing the between groups *MS* by the within groups *MS* [Huc96 p. 277]. The *P-value* indicates the probability of obtaining sample data that deviate as much or more from the hypothetical difference (in this case 0) than the observed data.

ANOVA of all three groups showed no significant differences. Since $F = 1.71 < F_{critical} = 3.04$, the hypothesis that the three groups are equivalent is accepted. Thus, the three groups can be considered identical.

Table 8.3 Web Pilot Main Test ANOVA Analysis.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Both	72	661	9.18	10.63
Sound	74	418	5.65	8.48
Vision	75	697	9.29	10.32

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	634.65	2	317.32	32.37	0.00	3.04
Within Groups	2137.06	218	9.80			
Total	2771.71	220				

Since $F > F_{critical}$, H_0 is rejected. Groups are not equivalent.

Analysis of the Main test produces a different result however. In this case, as shown in Table 8.3, ANOVA comparison at the $\alpha = 0.05$ level produces $F = 32.37 > F_{critical} = 3.04$, and the equality hypothesis is rejected. Since there is a significant difference between the groups, and since the experiment was designed to make such comparisons, analysis of the data can be made with a series of *t* -tests.

Table 8.4 Two Tailed *t* -test Between Groups for Web Pilot Main Test.

Group	Mean	Variance	*df*	*t*	$t_{critical}$	*P* (*T* <=*t*)
Sound	5.65	8.48	73	-8.13	1.99	0.00
Visual	9.28	10.45				
Both	9.18	10.63	145	-0.21	1.98	0.83
Visual	9.29	10.32				
Both	9.18	10.63	144	6.91	1.98	0.00
Sound	5.65	8.48				

In the Sound vs. Visual comparisons $|t| = 8.13 > t_{critical} = 1.99$. In the Both vs. Sound, comparisons $|t| = 6.91 > t_{critical} = 1.99$. These results indicate that there is a significant difference between the Sound group and the other two groups. In the Both vs. Visual comparisons $|t| = 0.21 < t_{critical} = 1.98$. This result indicates that there is no significant difference between these two groups. Thus, the difference in performance between the Sound group (at 40% correct) and the others (at 66% correct) is a significant effect.

It is interesting to note that the Sound group took 1.8 minutes longer to answer the 14 questions of the test than did the Both group. This averages out to about 8 seconds more per question. Since both of these groups had similar times for download and display of the graphs, the extra time may indicate the extra time required for understanding the graph when there is no visual cue. However, more likely explanations are that on average, the Both group did not play the graphs, or that the Sound group replayed the graphs an extra time.

There were eight questions where the Sound group fared particularly poorly in comparison to the other groups. The difference in average percentage of correct responses for questions 1, 2, and 3 was 38, 30, and 37% respectively. This might be attributable to unfamiliarity and lack of training with the display format. If subjects did not follow the optional links on the introductory page, the auditory format may have caused some confusion. The experiment had not been designed to record whether or not subjects had reviewed the supplementary material.

It is clear from Figure 8.1 that all subject groups had difficulty with question 5 of the Main test since the correct response rate was very low. Question 5 displayed a linearly decreasing graph plotted on axes of velocity and time. The question asked about an objects' motion. Analysis of the answers showed that a majority of the subjects (B – 81%, S – 68%, V – 72%) choose the incorrect answer of a linearly decreasing acceleration rather than the correct response of constant acceleration. This question demonstrates subjects' difficulty with concept of acceleration.

Question 6 had the largest difference in the percent correct between the groups (45%). The question displayed a complicated, segmented graph of an object's motion and asked subjects to find where the acceleration was greatest. The large difference in performance may have been due to a lack of training and a misunderstanding of the derivative indicators, or the results could indicate that this was a particularly poor question for auditory graphs.

Questions 11 and 12 involved composite graphs with linear and curved sections and each had a difference of 30%. The sound group may have had difficulty with these questions due to the difficulty representing the value $y = 0$ with sound. Questions 10 (29%) and 14 (36%) involved curved graphs where the Sound group may have again been hindered by lack of training and thus found these graphs confusing.

Several questions show that the auditory format has at least some promise, even when subjects have had virtually no training. Questions 4 and 7 involved linear graphs and had differences of only 14% and 18%. While not perfect, this may still indicate that the auditory format can be used even with very limited examples and training.

Questions 9 and 13 were both graphs of x^2, but had differences of 15% and 25%. The sound group tended to perform somewhat better with these curved graphs than with the others, but the 10% range is troublesome.

Table 8.5 Split-half Analysis for Web Pilot.

Question	% Correct	Split Question	% Correct
1	52	4	67
2	56	7	72
3	44	6	52
5	12	8	52
9	66	13	59
10	60	14	56
11	64	12	61

Split half analysis on the difference between the Sound and Visual groups gives a correlation $r = 0.47$. Using the Spearman-Brown fromula of $r' = 2r / (1+r)$ yields a correlation of $r' = 0.64$. This result shows some consistency between the questions but also shows the effects of the wildly varying performances.

8.7. Conclusion From the Web Pilot Test

It was strikingly apparent from this pilot study that using the World Wide Web as a testing environment had enormous advantages. An automated display and recording system was able to provide results that would otherwise have required over 100 hours of guided interviews. The Web-based test also eliminated scheduling conflicts and provided reasonable participation. Even though only about a quarter of the class was in attendance the day the test was announced, over half of the enrolled students participated.

Several subjects e-mailed comments about how interesting and enjoyable the auditory test was. The Web-based testing method eliminated any effects of pressure due to the proximity of an investigator as well as allowed for an unlimited time to complete the test. While this method produced many good results with relatively few problems, the method was not perfect. Approximately 10% of the subjects attempting the test either were not able to complete it, or had to try multiple times due to technical difficulties.

The results showed a difference between the Sound and Visual groups' average correct response rates of 26%. It is evident that the auditory graphs used in this test were

not as effective as visually displayed graphs. One possibility is that difference was caused by the lack of a proper introduction to the new graphing technique. Since subjects were not forced to understand the auditory graphs before starting the test, they may have found the graphs confusing.

This problem was addressed in the Main Auditory Graph test discussed next. It should be noted that had the auditory graph group been simply guessing, the correct response rate would have been about 20% instead of 41%. Thus, subjects were able to use these graphs to a limited extent even without training.

It was also evident from this pilot test that there were too few questions to provide a useful comparison between the test groups performances on linear, curved, and more complex graph patterns. Also, it was not clear from these questions how well the subjects were able to understand the shape of the graph versus their ability to draw conclusions from the graphs. Therefore, the Main Auditory Graph test used an expanded set of questions, including separate sections devoted to math or physics based graphs.

9. MAIN AUDITORY GRAPH STUDY

9.1. Overview

The Main Auditory Graph test was the culmination of the techniques used in the pilot tests. Web-based testing techniques and instruments were developed in the Web Pilot test. However, from that pilot test it was evident that a better introduction to auditory graphs and more complex and complete test questions were needed for the Main test section. The Main test questions were rewritten to produce better data for analysis. The number of questions was expanded to include questions concerning mathematical functions, as well as to provide a wider range of questions that would probe subjects' understanding of physics concepts.

The original goal of the auditory graphing method was to provide visually disabled people with a method to quickly access information that is usually portrayed by picture graphs. The Main Auditory Graph test therefore included a small group of blind subject volunteers to evaluate the effectiveness of these graphs for the intended user. This group of subjects was not used in the pilot tests due to the extreme scarcity of subjects fitting the testing requirements. The subject population for this experiment included: undergraduate students from several institutions, graduate students to check the reliability of test questions, and blind volunteers.

9.2. Sample

As the testing process was designed for first-year physics students, instructors of these courses at several educational institutions were solicited during the Spring and Fall 1998 terms for the possibility of letting their students participate in this study. It was arranged with one instructor at Oregon State University (OSU) and one instructor at Pacific University (PU) of introductory, algebra-based, physics courses to provide extra credit homework points to students taking Web Pilot test. An instructor of a calculus-

based introductory course at Pacific University also had her students participate for credit. An instructor of an algebra-based physics course at Linn-Benton Community College (LBCC), and a professor of a calculus-based course at OSU mentioned the study and Web address in class but did not offer credit for participation.

OSU physics graduate student subjects were informally solicited throughout 1998 for their participation. Six graduate students took the test using the auditory graph presentation. Graduate students were used as experts in order to provide data about the test's validity. Two additional graduate subjects used the wrong class code and received the test with visual graphs: their scores are not reported with the rest of the data. Two other graduate students attempted the test but due to technical difficulties their scores were not recorded.

Student subject participation from each physics course was not uniform. There were two factors for this. The most important factor for participation was the willingness of the instructor to issue extra credit for participation. When extra-credit was given for the test, participation was generally over 50%. The credit that subjects received played virtually no part in their overall grade. When extra-credit was not given, participation was greatly reduced. The second most important factor for participation in the Main Auditory Graph test was course size. However, credit was by far the dominant factor.

Blind subject volunteers who had experience with college-level physics and who where willing to participate in a Web-based test were solicited by posts to e-mail lists, and through personal contact at conferences. Interested subjects were sent Braille formatted information packets containing tactile graphs, introductory information, and the Web address location. A computer diskette containing the same text as the Braille information was also included in the packet. Blind subjects participated throughout 1998.

A blind physics professor was consulted during test development, and had acted as a critical evaluator of this study. Five blind subjects participated as subjects. Although this is a small number, the level of participation is a significant achievement as none of the test subjects participated locally. One of the subjects participated internationally from Europe while the other four were domestic. From solicitations, 15 interested volunteers provided mailing addresses for the information packets. Of this number, six subjects

decided to participate and were able to access the Web page test. One participant was unable to complete the test due to technical difficulties.

The Table 9.1 shows the distribution of the subjects among courses, schools, and approximate course sizes from which they were drawn.

Table 9.1 Distribution of Subjects per Course

Course	# subjects	Approx. Course Total	Date
OSU 203, algebra	189	350	Spring 98
OSU 213, calculus	2	200	Spring 98
LBCC 203	4	20	Spring 98
PU, algebra	28	44	Fall 98
PU, calculus	8	30	Fall 98
Graduate	6	N/A	98
Blind	5	N/A	98

Although there did not seem to be any effect on the test results, of the 189 subjects in the OSU 203 course, 85 had taken the Web Pilot test. Subjects from physics classes were randomly assigned to one of three test groupings. Of the 231 subjects, 74 subjects received auditory graph, 76 received visual graph, and 81 received both auditory and visual graph presentation methods. These numbers allow for statistically significant results at the $p = 0.05$ level since for three test groups, the number of subjects in each group should be greater than 62 (from Equation 6.2).

9.3. Data Collection

Data were collected in a similar manner as in the Web Pilot test. After an initial welcoming and informed consent page, all subjects were given a short tutorial on the auditory graph presentation method with several examples for them to try. The tutorial consisted of a series of graph descriptions, images, and sound files of increasingly complex auditory graphs for them to experience. After the introductory page, there was a log-in page to record the subject's name and class code. PERL script programs recorded subjects' answers and presented them with subsequent Web question pages in an identical

fashion to the Web Pilot test. Material pertaining to the Main Auditory Graph test can be found in Appendix C.

At the end of the test, subjects were presented with a page that thanked them for their participation. This page also contained links to a page of correct answers, an e-mail response form for any comments, informational pages on how the graphs were developed, and to the Science Access Project home page.

9.4. Instrument Development

The Survey and Pre-test were identical to those of the Web Pilot. The Main test section however had been considerably altered from those used in the pilot studies. To determine how well subjects were able to identify graphs versus how well they could use graphs for interpretation of physical phenomena, the Main test was divided into two sections, Math and Physics, of 17 questions each.

The Math and Physics sections had virtually identical graphs, and order of graph presentations was the same for the two sections. One question from the Math section contained a graph that was different from the corresponding graph in the Physics section. In the Math section, the graph displayed point discontinuities while the graph in the Physics section was that of a black body spectrum. The rationale for having two sections of similar graphs was so that split-half analysis of the sections could be performed in order to investigate consistency and performance issues relating to identification or analysis type questions.

The first question in each section consisted of a linear graph with a slope of zero. Aside from this graph, there were eight pairings of similar graph types. Thus, each graph type would appear twice in each test section. Graphs were grouped in the following categories: linear, step function, simple positive curvature, simple negative curvature, linear and curved composite, simple curved peak, complicated functions, and multiple peaked. The rationale for having two graphs of each group was to allow for a split-half analysis of each subject test.

As this was a somewhat iterative process, questions were developed based on the graphs, and graphs were chosen based on the types of questions that could be asked of

them. Several, but not all, of the questions from the pilot tests were used for this test. Questions were also chosen based on a diverse range of physical phenomena and their prevalence in the subject matter of introductory math and physics courses. The graphs and their questions were reviewed by Math, Physics, and Science Education faculty for content validity. The graphs and corresponding questions can be found in Appendix C.5.

There were a few modifications of the Web page display between the Web pilot test and the Main Auditory Graph test. As the auditory graphs had no label for their axes, the range of data values was explicitly stated in the questions. Also, several subjects in the Web Pilot test noted that it was difficult to locate the zero point on the auditory graphs. For this reason, a link to a MIDI file that contained the pitch representing zero was included with the auditory graph. The idea was that subjects could compare the "zero" pitch to the pitches of the auditory graphs. The zero sound for all graphs was identical. After taking this test, a couple subjects commented via e-mail that the technique of having the additional zero sound file was not particularly helpful.

Other changes to the test included annotating all images using the "alt" tag field. The graph images were produced with Microsoft Excel 5. Equations in the test were displayed using small graphic images of the equations. These images were created with Microsoft Word'97. All equation images were alt tagged with a linear notation for the mathematics.

The entire test was checked for compatibility with the JAWS screen reader and with Microsoft's Internet Explorer. As noted in the Web Pilot, the auditory graph sound files were displayed in three formats so that users could pick the format that was most compatible with their system. The test was also checked for keyboard access to all links and text entry fields. These last issues were vitally necessary so that the blind subjects could access, take, and understand the test.

The introductory material and Pre-test contained visually presented graphs. Blind subjects had been sent information packets containing these graphs. The graphs were represented as high-resolution tactile graphic images and were produce by the TIGER printer at OSU. Unfortunately, the informational packets were often ignored and the Pre-test questions went unanswered by a majority of the blind volunteers.

9.5. Results

Table 9.2 is a summary of more complete results for the Main Auditory Graph test contained in Appendix C.6. The table is divided by results from the different test groups for the Pre-test and Math and Physics sections of the Main test. Labeling for groups is as follows: S for the group with auditory graphs (Sound), V for visually presented graphs (Visual), B for both auditory and visual graphs (Both), G for graduate student subjects (Grad), and N for non-sighted subjects (Blind).

For the V, B, and S groups, equation 2.2 yields a limit on the error of the averages:

$$L = \left| 1.96 \frac{\sigma}{\sqrt{n}} \right| = 1.96 \frac{0.19}{\sqrt{74}} = 0.04 = 4\%. \qquad (9.1)$$

This result is the 95% confidence limit that the average values for each question are correct to within 4 percentage points. For example, there is a 95% certainty that the Main test question number 33 for the Sound group is between 60% and 68%.

While Table 9.2 provides an accurate listing of the data, it is helpful to view the same data as a bar chart in order to recognize patterns in the data and to easily see where any difficulties may lie. For example, question 34 has an unusually low result for all groups and requires careful analysis of its data. The original data for this question showed an even distribution of answer choices indicative of random guessing. Thus, the data show the effect of a poorly written question.

It should also be noted that the Sound group does not display any increasing trend (either absolute values, or relative to the Visual group) that would indicate better performance as subjects gain experience using auditory graphs. This could be an indication that the introductory material explaining the auditory graphs was sufficient for the purposes of this test.

Table 9.2 Table of Percentage of Correct Answers per Group for Each Problem.

Question	V - Visual	B - Both	S - Sound	G - Grad	N - Blind
Pre-test					
p1	72%	83%	78%	100%	20%
p2	95%	96%	97%	83%	20%
p3	93%	89%	93%	83%	20%
p4	67%	81%	89%	100%	20%
p5	55%	68%	66%	100%	20%
Main Test Math					
m1	84%	64%	57%	100%	100%
m2	86%	79%	46%	83%	100%
m3	82%	80%	77%	100%	100%
m4	78%	79%	76%	100%	100%
m5	80%	83%	69%	100%	100%
m6	61%	42%	24%	67%	80%
m7	83%	81%	70%	100%	100%
m8	76%	78%	68%	100%	80%
m9	66%	68%	45%	100%	60%
m10	62%	75%	55%	83%	80%
m11	58%	58%	36%	100%	80%
m12	38%	38%	14%	67%	40%
m13	68%	65%	49%	83%	80%
m14	28%	17%	22%	83%	40%
m15	66%	63%	59%	83%	80%
m16	49%	46%	46%	100%	80%
m17	30%	26%	28%	100%	80%
Physics Section					
m18	57%	51%	41%	83%	100%
m19	42%	35%	35%	33%	60%
m20	21%	28%	19%	83%	80%
m21	41%	41%	26%	83%	20%
m22	76%	81%	84%	100%	100%
m23	62%	44%	58%	50%	80%
m24	72%	79%	62%	100%	100%
m25	70%	62%	49%	100%	100%
m26	58%	62%	36%	100%	60%
m27	63%	64%	50%	100%	100%
m28	54%	58%	27%	100%	100%
m29	54%	43%	35%	83%	40%
m30	45%	48%	54%	17%	80%
m31	37%	36%	42%	83%	60%
m32	72%	70%	62%	100%	80%
m33	72%	70%	64%	100%	100%
m34	18%	16%	23%	17%	0%

The following charts display the percent correct scores for each testing group vs. the individual test questions. The charts are divided by test section. For Figure 9.1, it should be noted that only one blind subject completed the Pre-test, but that all his answers were correct.

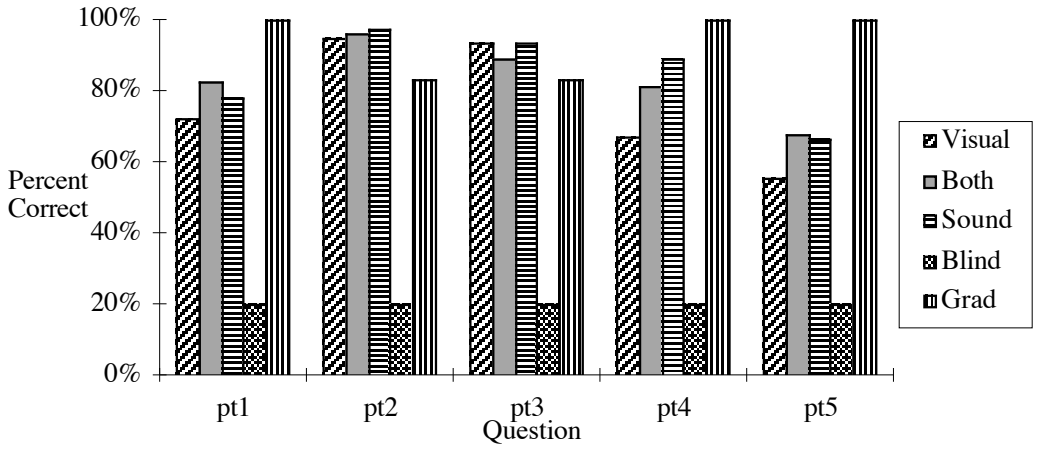

Figure 9.1 Pre-test: Average Percent Correct per Group.

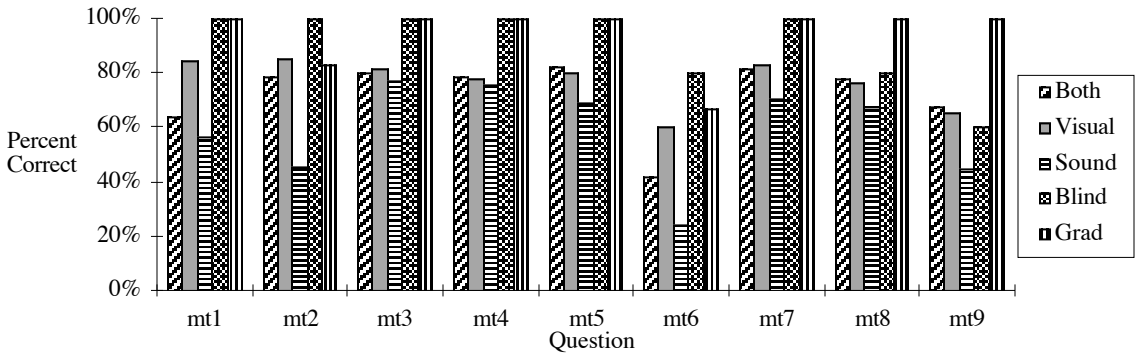

Figure 9.2 Math Section: Average Percent Correct per Group. Questions 1-9.

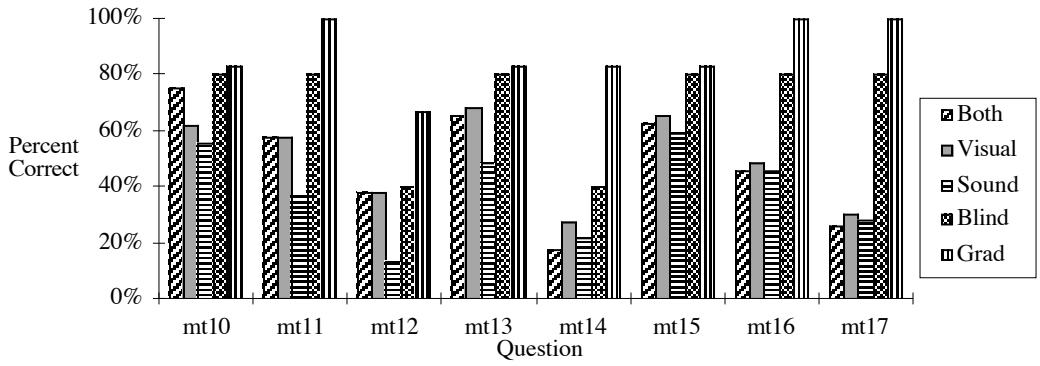

Figure 9.3 Math Section: Average Percent Correct per Group. Questions 10-17.

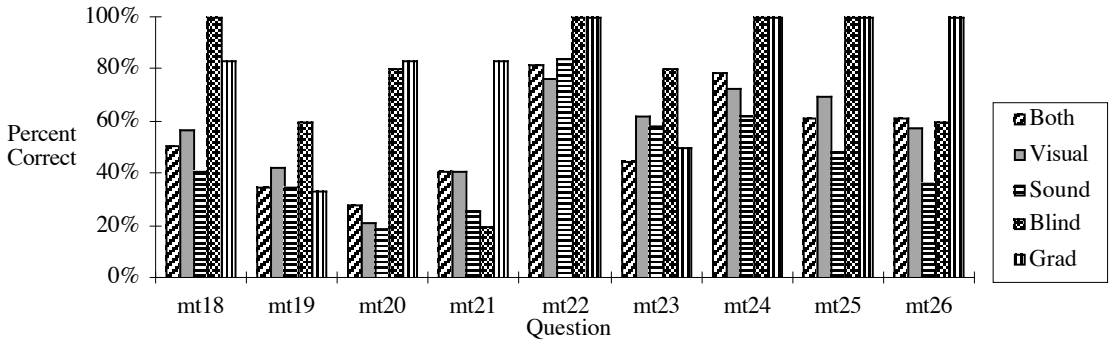

Figure 9.4 Physics Section: Average Percent Correct per Group. Questions 18–26.

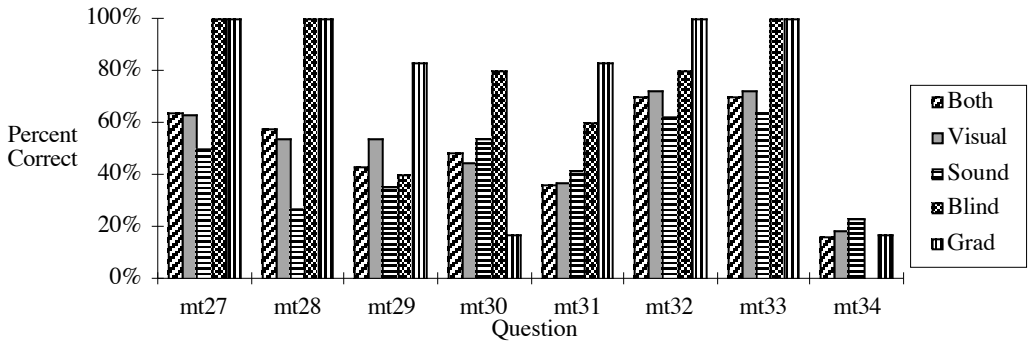

Figure 9.5 Physics Section: Average Percent Correct for Questions 27-34.

The average values for the test sections, standard deviations of the averages, and average time for test completion are given in Table 9.3. The average Pre-test score for the Blind group reflects the result that only one blind subject completed the Pre-test. The large average time for the Blind group was due to several of the subjects starting part of the test, and returning a day or two later to complete the test as their schedule permitted. The average time for the two blind subjects completing the test in one day was 79 minutes.

Table 9.3 Raw Average Percent Correct per Section per Group.

Group:	Both	Sound	Visual	Grad	Blind
Average, Pre-test	83%	85%	77%	93%	20%
Standard deviation ( )	10%	13%	17%	9%	0%
Average, Main	57%	47%	59%	85%	78%
	20%	18%	19%	23%	25%
Average, Math Section	61%	49%	64%	91%	81%
	21%	20%	19%	12%	19%
Average, Physics Section	52%	45%	54%	78%	74%
	18%	17%	18%	30%	31%
Average time to Complete	30 min.	34 min.	24 min.	40 min.	41 hrs.

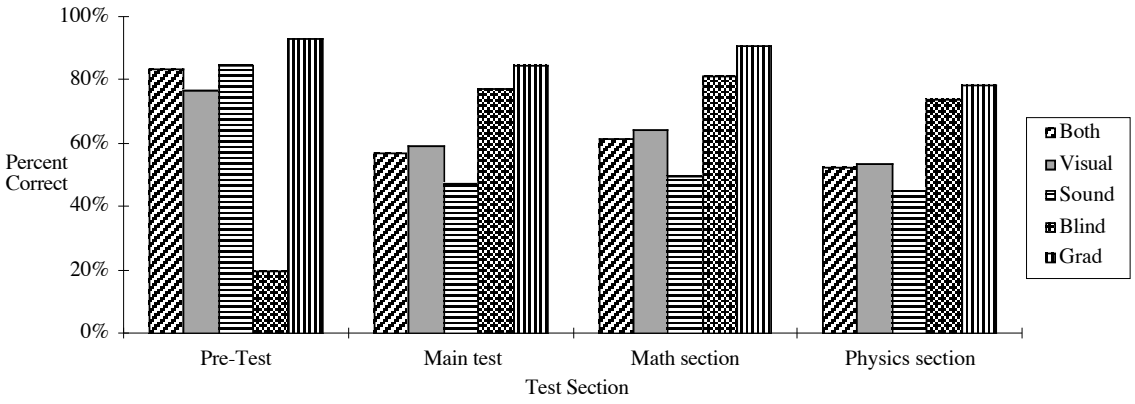

Figure 9.6 Raw Average Percent Correct per Group for Each Section.

The scores in Table 9.3 and Figure 9.6 represent raw averages. That is, they are not corrected for the possibility of subjects randomly guessing answers. To account for this possibility, the scores are modified as noted by Equation 2.9 in section 2.1.9. For the Pre-test, there were an average of seven answer choices, thus the adjusted Pre-test score becomes: $C' = C - W/6$, where C' is the corrected score, C is the percent correct, and W is the percent wrong. For the Main test, there were five answer choices so $C' = C - W/4$.

The adjusted scores are listed in Table 9.4 and shown in Figure 9.7.

Table 9.4 Average Percent Correct per Section per Group Corrected for Guessing.

Group:	Both	Sound	Visual	Grad	Blind
Average, Pre-test	81%	83%	73%	92%	9%
Average, Main	46%	34%	49%	72%	72%
Average, Math Section	52%	37%	55%	76%	76%
Average, Physics Section	40%	31%	42%	68%	68%

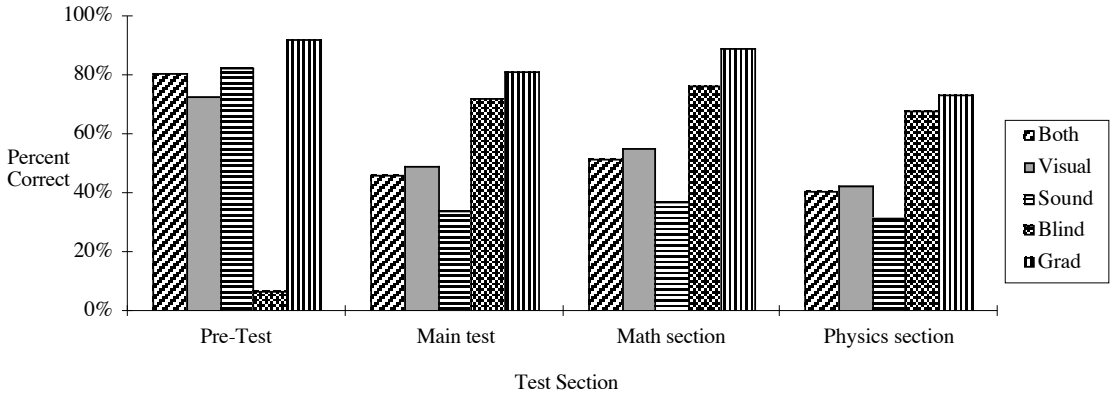

Figure 9.7 Average Percent Correct per Group for Each Section Corrected for Guessing.

9.6. General Analysis of Data

The data were analyzed in several ways. Section 9.6.1 is a description of relevant checks on the Main Auditory Graph test's validity. Section 9.6.2 is a description of issues relating to the reliability of the Main test. In addition, there are sections describing the effect of musical training on performance, the length of time for test completion, and relative performance between the test groups.

9.6.1. Validity Issues

9.6.1.1 Criterion-related Validity

Concurrent validity is a criterion-related validity. This form of validity is established by correlating a new test with a well established test. Unfortunately, there is no similar test for comparative purposes so this form of validity was not well established. Some of the questions in this experiment were modified from questions used in previous research, while the rest were developed to be of a similar format and style. The graph types for questions 1, 2, 3, 7, 8, 11, 12, 18, 19, 20, 24, 25, 28, and 29 were similar in the overall trend as those used in the Flowers study [Flo95], but questioning method was quite different.

The best comparisons of the questions that can be made are Pre-test questions 4 and 5, and Main test questions 18, 19, 20, and 21. For comparative purposes the results from the Visual group are cited as that group is closest to the testing method in Beichner's [Bei94] study. The percent correct values quoted in Beichner's study are approximate as they were displayed in a pie-chart format rather than tabulated values. The wording and graph for Pre-test question 4, which had 80% correct, was similar to Beichner's question 2 with approximately 65% correct. Pre-test question 5 and Main test questions 18, 19, 20, and 21 were variations on the theme of Beichner's questions 3 and 21. Pre-test question 5 had 55% correct while Beichner's questions 3 had 60% correct. The standard deviation for the Visual group's Pre-test was 17%.

Main test question 18 (with 57% correct) used linear graph with zero slope instead of the linearly increasing graph in Beichner's question 3 (with 60% correct). Main test question 19 (42%) was also similar to Beichner question 3 (60%) although it had velocity instead of distance for an axis. Main test question 20 (41%) was similar to Beichner's question 21 (20%) although the wording of the answers was altered. The average standard deviation for the Visual group's Main test was 19%.

The differences in the scores between this experiment and that of Beichner's study may reflect the changes in the wording of the questions, graph type, and axes. However, the results are almost all within one standard deviation and are an acceptable result. Thus, while the test as a whole cannot show concurrent validity, there are indications that at least some of the questions used demonstrate this trait.

9.6.1.2 Face and Content Validity

The objective of the test was stated when volunteers were requested. The objective was also explicitly stated in the welcoming Web page's initial paragraph and again in the Statement of Informed Consent. Thus, subjects were aware of the purpose of the test and face validity established.

Professors in the Math, Physics and Science Education departments at OSU, provided initial input during question development and reviewed the final test questions for appropriateness and content. Several physics graduate students also provided review and commentary of the test. Thus, there is evidence that content validity is established.

9.6.1.3 Construct Validity

Construct validity is a statement about how a measurement performs in accordance with theoretical expectations. Unfortunately, there is no well established theory relating auditory and visual graphs for use as a construct to generate test items. However, a hypothetical relationship was employed. This relationship stated that physics students could use auditory graphs at a performance level equal to that from visual graphs when answering questions that may encounter. Empirical evidence can be gathered by

examining the performance difference between expert s and novices. In the case of the Main Auditory Graph test, graduate students served as the experts, and the first-year subjects served as the novices.

Graduate student subjects were solicited as subjects because of their experience with the physics material and graphs. They received the test with auditory graphs to determine what would be the best expected results for students using auditory graphs. If the graduate students consistently missed specific questions, then careful examination of that question's validity would be necessary.

Two graduate subjects inadvertently received the visual test. These subjects each missed one question (number 8 for one, and number 30 for the other) so these questions are of concern. More importantly are the questions where a majority of graduate students using the auditory graphs gave incorrect answers. This occurred for three questions 19, 30, and 34.

Question 8 involved the graph of $1/X$. It is not clear why one of the graduate students taking the visually presented graph test missed this question, but it could be attributed to misreading the question. No other graduate students missed this question so it remains valid.

Question 19 involved a linearly increasing graph with axes representing velocity vs. time. The most common answer (3 of 6) was "D: The object is moving with a constant velocity" whereas the correct response "The object is moving with a constant, non-zero acceleration" was only answered by two subjects. It is suspected that the subjects were not paying close attention to the statement describing the axes values and representation. Given that all other groups (including the S group) outperformed the graduate students on this question, and that the two graduate subjects did answer this question correctly, the question was retained as valid.

Question 30 involved the identification of an intensity pattern produced by a double-slit source. Five of six graduate students (and one of the visual test grads) identified the pattern as that of a single slit source. While there are similarities between the two patterns under certain circumstances, the other subjects groups correctly identified the pattern at a minimum level of 45% correct. The difference in results may have been due to the graduate students attempting to analyze the problem at a theoretical

level that was more complex than was necessary for the test. Due to the response rate from the other groups, including the Blind group, this question was retained even though the construct validity may be in doubt.

Question 34 involved a determination the initial conditions for the motion of a mass suspended by springs on a cart. Several of the graduate students mentioned that the question was confusing, and responses from all groups followed a random distribution of the possible answers. Thus, question 34 was a poorly designed question, and was dropped from the analyses.

Recalculating the average correct scores for the Main and Physics sections without question 34 adjusts the Main and Physics sub-test averages. Table 9.5 displays the original average scores while Table 9.6 displays the averages after correcting for guessing.

Table 9.5 Recalculation of Raw % Correct Without #34.

	Both	Sound	Visual	Grad	Blind
Main	58%	48%	60%	87%	80%
Math section	61%	49%	64%	91%	81%
Physics section	55%	46%	56%	82%	79%

Table 9.6 Recalculation of % Correct Without #34 Corrected for Guessing.

	Both	Sound	Visual	Grad	Blind
Main	48%	35%	50%	84%	75%
Math section	52%	37%	55%	89%	76%
Physics section	43%	33%	45%	78%	73%

The graduate students result was over two standard deviations greater than that of the Sound group, and over one standard deviation greater than results of the Visual group. Thus, there is some empirical evidence that the test displays construct validity.

9.6.2. Reliability Issues

9.6.2.1 Split-half Analysis

One method of determining the internal consistency, and hence the reliability, of a test is with split-half analysis as noted in section 2.1.6. To reflect the reliability of the test as a whole, the correlation coefficient *r* determined from comparing the two test halves is used with the Spearman-Brown formula: $r = 2r / (1+r)$. When a test is further subdivided, the Spearman-Brown formula becomes: $r = 4r / (1+3r)$. A value of $r = 1$ indicates perfect correlation; the two tests are identical. A value of $r = -1$ indicates that a subject who answered correctly on one sub-test, answered the split question incorrectly and vice versa. A value of $r = 0$ indicates no correlation between the two sub-tests. Ideally, the correlation value *r* should be greater than 0.7 to be considered reliable.

Section 9.4 noted that the Main test had two sections, Math and Physics, containing similar graphs. Each of these sections was designed to be divided into two sub-tests containing graphs of similar nature such as derivative and complexity. For example a graph of $y = x$ was paired with a graph of $y = A - x$, and $y = x^2$ was paired with $y = 1/x$.

Table 9.7 lists the split-half correlation values between and within the Math and Physics sections for each of the groups (B - Both, V - Visual, and S - Sound). Correlation coefficients between the two section tests, and between the sub-test for each section, were calculated for the different groups. The correlation coefficients compared one set of questions, to a second set. The score of each question had been adjusted to account for the possibility of guessing prior to calculating the correlation coefficient. This was accomplished by applying the equation 2.9, $C' = C - W/(A-1)$, to the percent correct score for each question. In this case, *C* is the percentage of students with the correct answer for question *n*, *W* is the percent wrong, and A is the number of choices for that question.

The correlation coefficients were then adjusted by the appropriate Spearman-Brown formula. The adjusted values are displayed in the table.

The Blind (N) and Grad (G) groups are included for completeness, but it should be stressed that due to the small nature of these last two groups, the results have greater errors associated with their results. Since question 34 was dropped due to validity issues, it and its split question (question 17 in the between Math/Physics split, and question 32 in the within Physics split) were removed for calculations of the correlation coefficients. Removing question 30 and its split questions had a negligible effect on the results and the question was therefore retained in the calculations.

Table 9.7 Correlation r Between Sub-Test Groups for the Main Auditory Graph Test

Test	S-B formula	B	V	S	N	G
Main Test A Math questions 1-16 Test B Physics questions 18-33	$2r/(1+r)$	0.48	0.10	0.24	0.49	0.69
Math section Test A: 2,4,6,7,9,12,13,15 Test B: 3,5,8,10,11,14,16,17	$4r/(1+3r)$	0.92	0.95	0.87	0.96	0.29
Physics section Test A: 19,21,23,24,26,29,32 Test B: 20,22,25,27,28,31,33	$4r/(1+3r)$	0.68	0.68	(20.15)	0.66	0.72

The correlation values between the two Math sub-tests are generally very high with the exception of the Grad group. It is suspected that the correlation for the Grad group was so low due to the small number of subjects and low error rate on the questions.

The Physics section produced correlated results that were lower than the Math section for all groups except Grad. The correlation values between the two Physics sub-tests for the various groups are very close to the 0.7 limit for a reliable test. The exception is the Sound group whose adjusted correlation value r reflects the effect of a negative correlation value r. Obviously, this group does not show internal consistency between the Physics sub-tests.

There are very poor correlation values between the Math and Physics sections. Only the Grad group had a result that was close to the acceptable limit. The poor correlations may reflect that the math questions were more of a descriptive choice,

whereas the physics questions involved interpretation and understanding of physics principles. Poor understanding of the physics portrayed in a graph could have played a significant effect on ability to interpret the graph even if the subject could identify the graph.

The split-half analysis for the Math sub-tests demonstrated, for most groups, two equivalent tests. However, it should also be noted that the Grad group did not share the same level of correlation between the two Math sub-tests, in fact this group only had a poor correlation value.

The split-half analysis for the Physics sub-tests demonstrated a less successful attempt at developing two equivalent tests. The Sound group had a greatly reduced correlation result suggesting that the question and graph combination used in this section is least reliable for subjects who are new to auditory graphs and physics. Also, the reduced correlation coefficients may indicate that since the nature of the physics in the questions was different for each question, student understanding and performance ability varied, irrespective of the displayed graph. Subjects were answering questions about physical phenomena, and this was perhaps a more significant effect than the graph type.

9.6.2.2 K-R 20 Analysis

Another method for determining the internal consistency of a test is with the Kuder-Richardson #20 or K-R 20 as described in section 2.1.6. The K-R 20 result varies between 0 and 1 and is interpreted in a similar fashion as *r*. Table 9.8 lists the results of K-R 20 tests for the Main test well as for separate considerations of Math and Physics sections. The groups shown in the table include all of the undergraduate students as a group (All) as well as each group's separate results. All of the scores have been adjusted for guessing as described in section 9.6.2.1.

As can be seen from the table, all of the results show a high degree of internal consistency for the Main test, as well as each of the test sections. The scores are generally well above the 0.70 acceptance level. Thus, the Main test and the test sections can be regarded as internally consistent.

Table 9.8 K-R 20 Results for the Main Auditory Graph Test.

	All (v, b, s)	Both	Visual	Sound	Blind	Grad
Main	0.91	0.89	0.91	0.91	0.85	0.89
Math section	0.86	0.82	0.87	0.88	0.67	0.77
Physics section	0.84	0.84	0.86	0.82	0.80	0.80

9.6.2.3 Correlation of the Pre-test to the Main Test

The correlation values between the Pre-test and the Main test for each group give disappointing results. The correlation values shown in Table 9.9 are well below the 0.7 limit of acceptability. The low correlation scores indicate that the Pre-test is not a reliable indicator of a group's performance on the Main test. Hence, the Pre-test scores are not useful indicators for providing statements about group equivalencies. Unfortunately, these calculations were not performed for the Pilot tests, so the Pre-test was not modified in the Main Auditory Graph Test to provide more useful results.

Table 9.9 Correlation between Pre- and Main Tests.

All (B, S, V)	Both	Sound	Visual
0.29	0.28	0.46	0.33

There are several possibilities for the poor correlation scores: the Pre-test questions were not in the exact same format as the main test, the average question in the Pre-test was not as difficult as the average Main test question, the questions in the Pre-test could be skipped, and there were too few questions in the Pre-test. All of these factors may have contributed to the poor results.

The format of the Pre-test questions was that the questions were all displayed on one web page, and that there were several questions relating to a single graph. The Main

test only had one graph displayed at a time, and one question per graph. The average score on the Pre-test was 77% correct while that of the Main test was only 44% correct. This difference indicates a discrepancy in the difficulty of the questions and suggests that the tests were measuring different constructs.

The questions on the Pre-test were not mandatory as subjects could proceed to the next page before answering all questions. While this was not a common occurrence, seven subjects did not answer all of the Pre-test questions. Subjects were not allowed to proceed to the next question on the Main test until they completed their current question. The five Pre-test questions did not cover the breadth of material that the 33 Main test questions did. Only two of the Pre-test questions were similar to questions on the Main test. These questions did not display a high degree of correlation with the Main test.

Thus, the Pre-test acted as more of a familiarization with the testing interface, and not as a reliable indicator of group consistency. The only statement about the group's comparative ability makes the assumption that since the subjects were randomly assigned to the groups, the groups' average performances should be essentially equal if given the same test conditions. While this is not an ideal situation for the assumption, especially since the Pre-test did show a significant difference between groups, the groups were reasonably large so that any performance differences due to assignment to a group is small. The probability that one of the three test methods had the superior student for half or more of the three-student sets can be determined by looking at the possible combinations and their relative probabilities. For example, since each group had about 75 subjects, the total probability is found by summing the product of the number of combinations, the probability of r cases of the superior student being in the chosen group, and the probability of $75 - r$ cases of the student not being in the group [Sne89, p. 112]:

$$\sum_{r=37}^{75} \binom{75}{r} \left(\frac{1}{3}\right)^{r} \left(\frac{2}{3}\right)^{(75-r)} < 0.30\%. \qquad (9.2)$$

The result of less than a third of a percent gives a good indication that it is unlikely that there was a significantly uneven distribution of students' abilities between the groups.

9.6.3. ANOVA Comparisons of the S, B, and V Groups

Microsoft Excel was used for ANOVA calculations for the Main test and for the two test sections. The ANOVA calculations were single factor with α = 0.05 and compared the differences between the Sound, Visual, and Both groups. The results are given in Table 9.10. The numbers used for the calculation were corrected for guessing in the same manner as in section 9.6.2. Since $F > F_{critical}$ in all cases, these results indicate significant differences between the groups for each of the tests. Thus, it is worthwhile to make detailed comparisons between the groups using other analysis techniques.

Table 9.10.ANOVA Results

	F	$F_{critical}$	P-value
Main	10.78	3.04	0.00
Math section	12.84	3.04	0.00
Physics Section	5.23	3.04	0.01

9.6.4. Sheffé Tests for the S, B, and V Groups.

Comparison of the performance between groups can be performed by a series of *t*-tests. However, in order to perform any number of comparative tests, the Sheffé test is necessary in order to limit the probability of finding an erroneous significant result to at most 5%. The Sheffé test, compares the calculated *t* value to the critical value of:

$$\sqrt{(a-1)F_{0.05}} \tag{9.3}$$

where a is the number of comparison groups and $F_{0.05}$ is the 5% level of F dependant on the number of degrees of freedom. In this case, $a = 3$, and there are 2 and 148 degrees of freedom. P is the probability of randomly finding a t value greater than the value calculated from the data. As can be seen in Table 9.11, all of the comparisons with the Sound group lead to significant differences. The comparison of the Visual group to the

Both group does not have a significant difference. The data were compared using values adjusted to account for guessing.

Table 9.11 Sheffé Tests

group pair	test	$\|t\|$	t_{crit}	P	Sheffé	Sheffé P	significant
V-S	main	4.26	1.98	0.00	2.47	0.00	*
	math	4.60	1.98	0.00	2.47	0.00	*
	physics	3.02	1.98	0.00	2.47	0.01	*
B-S	main	3.71	1.98	0.00	2.47	0.00	*
	math	3.99	1.98	0.00	2.47	0.00	*
	physics	2.69	1.98	0.01	2.47	0.03	*
V-B	main	0.80	1.98	0.42	2.47	0.72	
	math	0.99	1.98	0.32	2.47	0.61	
	physics	0.45	1.98	0.66	2.47	0.90	

A t -test comparison between Grad (G) and Blind (N) groups showed that the 7% difference in the results was not significant ($t = 0.97 < t_{critical} = 1.99$). This result should be viewed with caution as the subject sample size was very small for each of these groups (n=16). Equation 2.2 gives an estimate that the average for each question is correct to ±18%, so a significant result may be masked by this uncertainty.

9.6.5. Effect of Music Training

A slight difference was noted in the scores on the Main test when comparing subjects in the sound group who had musical training (47 subjects) to those in the same group, but without any musical training (27 subjects). Musical training was determined from responses to a question on the Survey. Subjects with some music background had an average score of 12.9 of 33 (corrected for guessing), whereas those without music had a score of 10.0 (corrected for guessing). Since there are 27 subjects in the smaller group, equation 2.2 gives a measurement error of less than 8%. F -test results give $F = 3.04$ and $F_{critical} = 3.97$ with $P = 0085$ at the $\alpha = 0.05$ level. Since $F < F_{critical,}$, there is not a significant difference between the groups. A two tailed t-test at the $\alpha = 0.05$ level also

shows that $|t| = 1.74$ and $t_{critical} = 1.99$ with $P = 0085$. When $t < t_{critical}$ the null hypothesis of the two groups being equal is retained. Thus, music training seems to have a small effect, but the difference does not reach a statistically significant level.

9.6.6. Test Completion Times

The Main Auditory Graph test showed a difference between the average times taken by the Sound and Both groups for completing the test. This difference was similar to that seen in the Web Pilot. The time difference between the groups for the whole test was 3.1 minutes, or about 6 seconds per question. This is about the length of time required to listen to the auditory graph once. Thus, the average subject in the Both group either did not listen to the sound graphs, or the average subject in the Sound group played the graphs an additional time. The time difference between the Grad and Sound groups was 6 minutes, or about 10 seconds per question. Thus, the graduate students may have listened to the sound graphs an additional time or given more consideration to the questions.

9.6.7. Blind Subject Performance

The test results for the Blind (N) group were very good. This group, while small, performed at substantially better, on the order of 20%, than any of the undergraduate student groups. Because of the inherent differences in the group composition between the Blind group with the first-year students, there is no method for evaluating group equivalencies. Therefore, ANOVA, Sheffé or *t* -test comparisons between the Blind group and the student groups were not performed.

Although comparisons between the Blind and Grad groups are speculative, and should be viewed only as anecdotal evidence, ANOVA tests give no indication of significant differences. These results are displayed in Table 9.12.

Table 9.12 ANOVA Comparisons Between Blind and Grad Groups.

	F	$F_{critical}$	P	Significant
Main	1.12	5.12	0.32	No
Math	2.34	5.12	0.16	No
Physics	0.25	5.12	0.63	No

There are differences in the average scores between the Blind and Grad groups of 9% for the Main test, 13% for the Math section, and 5% for the Physics section. However, none of the differences between the Blind and Grad groups appear to be statistically significant. This result should be tempered with the reminder that there were only 5 subjects in the Blind group and 6 in the Grad group. However, this comparison, perhaps more than any other test conducted in this study, demonstrates the power of these auditory graphs. The blind subjects were able to access graphical information presented in an auditory format from around the world. They were able to comprehend and answer graph-based questions at a level comparable to physics graduate students at the local test site.

9.7. Conclusion of Main Auditory Graph Test

There was a significant difference between Sound and Visual graph groups. The difference between the average percent correct on the entire test with the scores corrected for guessing was 50 – 35 = 15%. This difference was less than that of 25% observed in the Web Pilot, which was a shorter test and did not have the scores corrected for guessing. The entire test spanned the 17 math and 16 physics questions, with one poorly designed question thrown out due to random answering. These questions had a correct response rate of 50% for the Visual group and 35% for the Sound group. The Sound group thus performed at 70% the level of that shown by the Visual group:

$$\text{Performance Ratio} = 100\% \cdot \frac{(\text{Average Sound Score})}{(\text{Average Visual Score})} = 100\% \cdot \frac{35}{50} = 70\% \qquad (9.4)$$

The effect of a brief, self-guided, introduction and training with several examples seems to have had a substantial increase in the performance of the Sound group between the Web pilot and Main Auditory Graph tests. While these results were from first-year physics students from several institutions, the majority of subjects were from a single course at OSU.

Expert physics students were able to effectively use the auditory graphs to answer questions at an average level of 84% correct for the valid questions. Although a larger number of subjects would be needed to verify this finding, the performance ratio between graduate students using auditory graphs versus those using visual graphs may be as high as 87%.

Blind users demonstrated a 9% difference in average scores on the Main test when compared to physics experts. This result is not a significant difference. However, it should be noted that the 95% confidence limit for a group of 5 subjects allows the average values to have a ± 18% error range which would mask any significant difference between these groups. Nonetheless, it is impressive that blind subjects were able to perform about as well as graduate subjects on this test. Perhaps even more importantly, they were able to answer the questions at a level of 75% correct. While this was not at the 97% level of the two sighted graduate students, it was considerably more than the 50% level of the Visual student group.

The large number of subjects that participated in this test demonstrates the feasibility, practicality, and usefulness of using the World Wide Web as a testing medium. In addition, because the test was available via the Web, blind subjects could participate even from very distant locations. This was particularly important due to the very limited number of blind subjects who have had some training in physics. Furthermore, the results between the Sound and Visual groups demonstrate not only that are auditory graphs practical in tests, but also that they can be used to achieve performances that are within 70% of those obtained when using visual graphs. The performance results for this type of auditory graph are from a very short, self-guided

training session. The new exposure to auditory graphs is an important consideration given the years of experience that subjects have had with visual graphs.

While many parts of this testing process were successful, especially in terms of demonstrating that graph-based physics questions can be answered, to a certain extent, using auditory graphs, there are many areas left to explore. Such questions include: What are the best methods for portraying these graphs? What preferences do people have for sounds used in the auditory graphs? What is the limit of usefulness for these types of graphs? These questions are explored in the next chapter.

9.8. Subject Comments About the Auditory Graphs.

Finally, this chapter will end with several comments made by several test subjects. At the end of the graph test, subjects were invited to e-mail comments to the author. The following quotes are taken from those notes. They are telling as to what subjects found interesting, and which areas still need improvement.

"It's easy to picture the graph being presented with audio tones."

"In general your audible graphs are the greatest thing I've heard about for a long time, and I hope you will continue to work on improving them."

"I think the whole idea is great and I think the drum beats to show curvature and slope are particularly functional and innovative. It is really important to develop the ability to hear negative values."

"I appreciate the value of getting blind users to try this and I am determined to get completely through it. By the way, did you try it blindfolded or you also blind? I want to make sure that you have gone through what I am going through (smile)!"

10. AUDITORY PREFERENCE PILOT TEST

10.1. Overview

While the Main Auditory Graph test was an effective test using auditory graphs, many unanswered questions arose. First, there were several assumptions inherent in the auditory representation. An example of an arbitrary decision in the Main Auditory Graph test was that data was represented with a piano tone, while a drum tone represented the derivative information. Any of a number of MIDI instruments could have been chosen for these representations. Also, the information for the second derivative used a high drum pitch for negative curvature, and a low pitch for positive curvature. This was a subjective choice by the author as a useful and convenient working model to begin with. There was no indication that these choices were necessarily the best ones to make.

In order to assess the effectiveness and desirability of various auditory graphing techniques, a test was developed that used a combination of pair-wise preference comparisons, graph identification questions, and Likert preference ratings. The preference questions were used to indicate which graphing styles subjects liked best, or thought were most useful. The graph identification questions were used to indicate which graphing style had the highest rate of being answered correctly.

This test was created not only to find better elements for the auditory graph displays, but also to test and evaluate an alternative method of auditory graph production. This alternate technique utilized Microsoft's ActiveX controls to create "live" graphs that have the potential for greater user control, customization, and flexibility than the prerecorded graphs could attain.

The results of preference tests such as this can be used to guide the development of software that uses auditory graphs. The Main Auditory Graph test demonstrated that basic auditory graphs could be used for answering questions. Tests such as the Auditory Preference Pilot can be used to discover what issues should be addressed for the best optimization of auditory graphs.

10.2. Sample

There were 13 subjects who participated in this study. As this was a Web-based test similar to the Web Pilot, one subject attempted the test from a remote location. Due to technical difficulties, the auditory graphs produced using the AudioPlot method were not active, thus the results from this subject were not included. There were twelve subjects that participated locally who used the same computer, but at different times, for the test. The subjects were solicited primarily due to their proximity to the research location at Oregon State University. The subjects included five advanced undergraduate physics students, three science and math education graduate students, three employees of the toxicology department, and an employee of the Science Access Project. The subjects were also chosen because they had not been involved in previous auditory graph research. This choice was an attempt to reduce bias due to familiarity with previous auditory graphing techniques.

10.3. Data Collection

Subjects were invited to an office that contained a desk computer with a Web browser displaying the test's introductory page. Due to the nature of the ActiveX components for creating some of the auditory graphs, Microsoft's Internet Explorer was used as the Web browser. Subjects were told briefly what to expect from the test, that the experiment used a Web browser to display a test consisting of nine questions about auditory graphs. They were also shown the controls for adjusting the volume of sound produced by a pair of speakers next to the computer. The investigator indicated that he would be in a neighboring room in case any technical difficulties arose, and left the subject to take the test.

Data collection was then similar to the method used in the Web Pilot and the Main Auditory Graph tests. A Web browser displayed graphs and information and PERL script programs recorded the answers. The test consisted of an introductory page with the Informed Consent Document and a brief description of the test. Next, subjects were presented with a page to record their names. A scripting program appended the

information to a file and assigned a code number. Subjects were then presented with a series of pages containing one or more auditory or visual graphs, a multiple-choice selection field, and a text entry box for them to comment on their graph choice. Another scripting program appended their code and text answers to a second data file and passed the code and multiple-choice answer to the next page. After completing the last question, the scripting program appended the code number and the string of multiple-choice answers to a third data file.

10.4. Instrument Development

There were nine question pages: four consisted of pair-wise auditory graph comparisons, four involved matching an auditory graph to a visual graph (two questions were matching a visual graph to a choice of auditory graphs, and two were matching auditory graphs to a choice of visual graphs), and one page with five-point Likert ratings of 6 graph types. Each question page had a text field for subjects to provide comments and reasoning for their choices.

The auditory graphs were produced by two methods. The first method played prerecorded MIDI sound files that used a piano instrument to represent the data values. This was the same method as was used in the Web Pilot and Main Auditory Graph tests. For this method, the data were mapped to a chromatic scale. The second method for generating the auditory graphs was with the AudioPlot ActiveX control from Oregon State University's Science Access Project. The AudioPlot (AP) control generated auditory graphs on the subjects' computer from equations specified in the Web page. This method allowed various graphing parameters to be set within the Web page code. The auditory graphs produced by the AP control used linear scale for mapping the data to sound.

Both the MIDI and AudioPlot methods played the auditory graphs when the subject selected a "play" button on the page. The buttons were identical so the subject had no indication of a difference between the methods to produce the graphs. The AudioPlot graphs produced a smooth, continuously varying tone with optional clicks for the derivative information. The MIDI graphs consisted more of a staccato piano note with a courser resolution. The derivative information was represented with a drum like tone.

It should be noted that a potential remote subject did not participate in this study citing security concerns with ActiveX control modules. The choice of using these controls to generate the auditory graphs on the Web was based primarily on the transport of Visual Basic code written for an updated version of the TRIANGLE graphing calculator. This code was able to be quickly modified to produce the AudioPlot control modules that were incorporated into the Web-based testing environment.

10.5. Data Results

Table 10.1 is a summary of the multiple-choice results for each question. The questions and answer choices are abbreviated for reference. The full text for the questions can be found in Appendix D.3.

Table 10.1 Summary of Answer Choice per Question.

Question	Answer Choice as Percentage of Total A	B	C	D	E	F
1. Gaussian curve: A = AP, B = MIDI, C = both, D = neither	33%	58%	8%	0%		
2. Gaussian curve with derivative: A = low +, high -; B = high +, low -, C = both good, D = neither good	33%	17%	33%	17%		
3. $x \sin x$: A = no change at 0, B = instrument change at 0, C = both good, D = neither good	50%	42%	0%	8%		
4. $x \sin x$: A = AP with deriv., B = MIDI with deriv. and pitch change a 0; C = both, D = neither	33%	50%	0%	17%		
5. Match visual graph of $e^{-x} \sin x$ to AP graph D: A = sin x, B = cos x, C = $x \sin x$, D = $e^{-x} \sin x$, E = $e^{-x} \cos x$, F = none	17%	0%	17%	58%	0%	8%
6. Match visual graph of $e^{-x} \cos x$ to MIDI graph E: A = sin x, B = cos x, C = $x \sin x$, D = $e^{-x} \sin x$, E = $e^{-x} \cos x$, F = none	8%	0%	0%	17%	58%	17%
7. Match AP graph of cos x to visual graph A: A = cos x, B = sin x, C = $x \sin x$, D = $e^{-x} \cos x$, E = $e^{-x} \sin x$, F = none	75%	0%	0%	17%	8%	0%
8. Match MIDI graph of sin x to visual graph B: A = cos x, B = sin x, C = $x \sin x$, D = $e^{-x} \cos x$, E = $e^{-x} \sin x$, F = none	0%	92%	0%	0%	0%	8%

9. Likert style 1- 5 preference of $x \sin x$ graph with different sound representations: 1 is bad, 2 is poor, 3 is neutral, 4 is good, and 5 great.	X avg.	std. dev.
A. MIDI	3.75	0.87
B. AP	3.75	1.14
C. MIDI, dx	3.08	0.79
D. AP, dx	3.42	1.08
E. MIDI, 0	3.83	1.34
F. MIDI, dx, 0	3.33	1.44

By equation 2.2, the error associated with the each of the Likert averages can be found. Using a 95% probability limit, the average of the standard deviations ($\sigma_{avg} = 1.12 = 28\%$), and the sample size of twelve subjects,

$$\text{Error} = \left(1.96 \frac{1.11}{\sqrt{12}}\right) = 0.63, \tag{10.1}$$

or about 16% since there was a 4 point range (5-1) in the rating scale.

10.6. Analysis

The results would have provided more consistency if a larger number of subjects had been used. Because of the small sample, the results do not provide convincing evidence of the superiority of any of the graphing methods. The purpose of this pilot test was to discover where any difficulties in the testing process may reside and to evaluate the question statements. Thus, this test should be viewed primarily as anecdotal evidence. However, tentative conclusions about the graphing methods can be made. Comparing the results above to the subjects' written comments about the reasons for their choices was very informative and greatly aided the interpretation of the results.

The first question compared MIDI and AudioPlot (AP) representations of a Gaussian curve. These graphs used only the *y* axis to pitch mapping. The results for question 1 imply that there was a preference (58 to 33%) for the MIDI graph over the AP graph. This is a somewhat surprising result as great effort went to produce a pleasing smooth sound. The commentary is very interesting as unexpected factors played a role in the choice. Subjects choosing the MIDI graph mentioned that it "seemed cleaner," and that the discontinuous sounds produced a more dramatic effect, making it easier to distinguish the maximum point on a graph. In contrast, at least one subject preferred the AP graph because the data were represented with continuous sounds.

Several subjects commented that their choice was at least partially based on the frequency ranges of the graphs. One subject who chose the MIDI preference noted that "the greater difference between the maximum and minimum tones made the graph easier to visualize." However, another subject chose the AP graph because "I seem to make the

connection better for the higher pitches." Thus, future testing will need to be careful that the different graphing methods display the same range in frequencies.

These choices may also reflect the difference in data mapping methods used by the two auditory plots. As has been noted in previous research by Stevens in Mansur [Man85], pitch has a logarithmic association with height. Thus, the linear mapping method used by the AP graphs had a perceptual effect of flattening the graphs' higher pitches and may have made them seem less distinct.

Question 2 investigated the pitch mapping preference for curvature. A very brief description of what the drum tone represented was given at the top of the page. The first graph used a low drum tone to indicate positive curvature and a high drum tone for negative curvature. The second graph had the reverse mapping. The graphs were again of a Gaussian curve. The results were that 33% (four subjects) preferred the first graph (A), while 17% (two subjects) chose the second (B). However, 33% didn't have a preference (C), and 17% didn't like either (D).

The comments provide the additional feedback that those choosing C or D often did so because they found the graphs confusing, or had a difficult time distinguishing between the graphs. Also, two subjects preferred one graph type gave the opposite graph a higher preference rating in question 9. This indicates the necessity for providing better descriptions and for asking the same question about several different graphs to determine some consistency in the responses.

Question 3 investigated the preference of including a change in the graphs' data sound when the y value was negative. This question was developed in response to comments received during the Main Auditory Graph test. For this representation, the data sound of the graph of $x \sin x$ changed from a piano tone for positive values to a harpsichord tone for negative ones. There was a slight but non-significant preference for the tone change. The reasons for not preferring the change are very informative. Cited complaints were that the tone change created "too many options for the ear to play with" and "broke up the graph a little too much." Those who preferred the tone change found it very helpful. One comment was: "I liked how the pitch changed when the graph went below 0. I think it is important to change the sound when some major distinction (like the

zero line) is involved." A tone change that is more pleasing and less distracting may greatly improve its preference.

Question 4 compared the graphs of $x \sin x$ between the AP and MIDI methods for graphs incorporating derivative information. The AP graph had a score of 33% and represented positive curvature with a high pitch click, and negative curvature with a low pitch click. The MIDI graph had a score of 50% with positive curvature represented by a low pitch drum, and negative curvature by a high pitch drum. The MIDI graph also incorporated a tone change for negative values. This feature was not included in the AP graph as it did not have a similar display option at the time.

Of the subjects choosing the AP graph (A) and providing comments for question 4, there is an indication that improvements were still desirable. Comments included: "A would be better if the drum pitch had those high harmonics for positive values instead of the negative ones," and "A sound is good to me. … Sharp pitch is better to me, but this one also needs some different sound to express the ups and downs." Of the subjects commenting on the MIDI graph (B), they cited that their choice was because "the distinct sounds in B were much more clear than in A." Subjects also chose the MIDI graph because of the "negative change and [because] you can pick up the slope/curvature better." There was also a comment by one subject who chose the neither (D) option because "both seemed rather arbitrary in relation to the graph, at least in the derivative department."

Comparing the results of questions 5 and 6, which were graph identification questions, shows identical results both in the number choosing the correct graph, and in the distribution of incorrect responses. In question 5, subjects were asked to match a visually presented graph of $e^{-x} \sin x$ to one of five AP auditory graphs. These graphs included the derivative indicators. In question 6, subjects were asked to match a visually presented graph of $e^{-x} \cos x$, to one of five MIDI auditory graphs. These graphs included the derivative and negative indicators. Thus, subjects seemed to be able to match a pictured graph to its auditory representation equally well with both methods.

Comments about the AP graphs in question 5 indicated that some subjects found the choices indistinguishable. There were statements of "I started to choose E or D, but really I didn't like any of the choices" from a subject choosing None of the Above (F),

and “frankly, a-d sounded all the same” from a subject choosing the correct answer (D). One subject who chose incorrectly, noted a disparity between the choice and their reasoning: “I just like the sound of C the best[;] however, listening to the pitches, it almost seems like the two maximums reach the same pitch, but on the graph, the second one is lower.”

Several of the subjects who answered question 6 incorrectly provided interesting comments about their choices on the MIDI graphs. One subject who answered incorrectly indicated that “the drums in the background created confusion as to what was going on.” Other comments, such as “A and E sounded nearly the same” from one who chose A, and “E seemed the closest, but the derivative portion seemed wrong” from one who chose F, indicated that several subjects almost chose the correct answer E. One subject who gave an incorrect choice of D noted that “the tempo of the drum was most clear in describing the slope of the line, as was the change in sound describing the negative values of the curve.” These comments may demonstrate the effect of attempting a comparative study on auditory graphs without having a training tutorial such as the one in the Main Auditory Graph test.

For questions 7 and 8, which also involved graph identification, subjects were given an auditory graph and were asked to choose between several visual graphs or a “None of the Above” choice. Question 7 asked subjects to match an AP graph of cos x to one of five visual graphs. This question had a correct response rate of 75%. Question 8 matched a MIDI graph of sin x to one of five visual graphs and had a correct response rate of 92%.

In question 7, one subject who answered incorrectly mentioned a difficulty in identifying the starting of the sound. Question 8 would have had a 100% correct score, but the one subject who chose F instead of the correct answer B mentioned that the graph “seemed to mostly fit B, but I don't think the derivative was correct.” The greater response rate on question 8 than on question 7 may have been a reflection of subjects gaining experience since the two questions were similar. Having a random assignment of which type of graph is encountered first would reduce this type of ambiguity.

The last question asked subjects to rate different auditory representations of the graph of $x \sin x$ on a Likert scale of 1 to 5, where 1 was bad, 2 was poor, 3 was neutral, 4

was good, and 5 was great. The results are given in Table 10.1, but are a bit vague due to the high standard deviations. All rankings should be viewed as essentially equivalent as the averages were all within the smallest standard deviation. ANOVA analysis of the average Likert scores from question 9 shows no significant difference between the methods at the $\alpha = 0.05$ level ($F = 0.82 < F_{critical} = 2.35$, $P = 0.53$). All the average scores were between 3 and 4 indicating that the methods could still be greatly improved.

Table 10.2 Ranking of Preferred Graph Types

Rank	Average Rating	Graph Type
1	3.83	MIDI with 0
2 (tie)	3.75	MIDI plain, AP plain
3	3.41	AP with derivative
4	3.33	MIDI with 0 and derivative
5	3.03	MIDI with derivative

Several subjects provided general comments on what they found helpful or annoying. These comments tended to focus on the drum beat (or clicks) indicating curvature, and the change in tone indicating negative values. A few selected comments demonstrate the greatest strengths and some potential problems with these auditory graphs:

"They all represented the graph well, it just depended on if one was interested in slope and curvature."

"I like hearing positive and negative. I like having pauses between notes instead of one constant sound. I like really hearing the slope. I don't like the soft drums because it's hard to differentiate them from the sound of the computer loading." The subject is referring to the fact that the computer had a somewhat noisy fan.

10.7. Conclusion for Auditory Preference Pilot

The Auditory Preference Pilot demonstrated some useful innovations in the development, production, and comparison of auditory graphing techniques. While the focus of this test was to provide an initial comparison of several of the assumptions used in the Main Auditory Graph test, it also provided a testing medium for a new control module that produced auditory graphs. The AudioPlot controls have the potential to provide auditory graphs with dynamic flexibility and customization for use on the Web.

The results of this pilot test indicate that a variety of graphing techniques is acceptable from a users' standpoint. Also, the results indicate that some auditory graph characteristics tend to be favored by a majority, but by no means all, of the subjects and that subjects' preferences seemed to change over the course of the test.

Comments and preference choices about graphing techniques showed a favoritism toward graphs where the sounds were clear and distinct with a wide tone variation. However, there were also indications that by the end of the test, some of the distinct display techniques became bothersome. In question 9 one subject remarked: "I am starting to find the drum beats to be annoying." From comments such as this, it is evident that there is an inherent need in the design of commercial graphing displays for user configurations of the graphs. Items such as pitch range, the ability to turn on and off derivative sounds, sound transformations at the zero point, and continuous or "broken" sound playback are all important features that should be considered.

There are several reasons why the results of Auditory Preference Pilot test are unable to be used to make a determination of which auditory graphing techniques are ultimately preferred. These reasons include the small sample size, the desire to test auditory graph options that had not been implemented in the AuditoPlot controls, and the limited number of test questions.

The use of the AudioPlot controls for graph generation has many powerful advantages. Once the control is loaded on a remote computer, many complex auditory graphs can be produced with little more than embedded commands in a Web page. The use of these controls eliminates the need for pre-produced graphs, and creates a dynamic display where users can provide a more thorough investigation of the graph than from

passive listening. The use of Visual Basic to create the ActiveX controls can result in short development times when adding features. Disadvantages of the ActiveX control system are potential security risks for users, the potential for missing support files (.dll files) on user computers, and the limitation to a single platform and virtual limitation to a single Web browser. The use of the JAVA language to create the auditory graphs is a possible candidate to remove the limitations of ActiveX controls.

Future studies will need carefully constructed questions as well as many graphing variables in order to provide definitive answers. A longer set of questions, with repetition of graph types to provide multiple comparisons is highly desirable as subjects tend to change their views about which styles are favored as they gain experience with the graphs. Ultimately, the goal is to have the graphing method controlled by the end user when he or she is selecting the styles that are most evocative for the particular graph that is being listened to.

11. CONCLUSION

11.1. Summary of Conclusions of Test Results

11.1.1. Triangle Pilot

The Triangle pilot test was useful for gaining experience in question development. It also provided a method to gauge the potential of an auditory graph display to be used in a testing environment. From the results of this test, the initial auditory graphing technique was modified in two significant ways. First, in order to accentuate the curvature, a derivative tick mark was added. The tick mark was represented by a drum beat where the tempo of the drum represented the magnitude of the graph's slope. Second, the data were mapped to a chromatic scale rather than the previous linear scaling. The tonal quality of the sound was also modified, but this was a result of using MIDI to implement the sound files rather than from research findings. The testing method was modified after analyzing the results from the Triangle Pilot. Testing was changed from a guided interview method to a Web based test so that subjects would be less influenced by time or environmental conditions. Although the Triangle Pilot demonstrated that there was a difference of 34% between the Sound and Visual groups, the group sizes were far too small for meaningful results.

11.1.2. Web Pilot

The Web Pilot test was important for gauging student participation in a self-guided test. Participation was not a problem when the test was offered for token credit. The Web Pilot demonstrated that the PERL scripting method used to display the test and record subjects' results worked well. This pilot test also demonstrated the inadequacies of the initial set of questions when used in a full comparative test. Thus, while the testing

environment did not need to be modified, the questions used in the test were extended and reworked to provide a more complete comparison with a higher level of internal consistency. The difference between the Sound and Visual groups for this test was 25%.

11.1.3. Main Auditory Graph Test

The Main Auditory Graph test effectively compared the performances of several groups of subjects when answering graph based questions. Unfortunately, because of poor correlation of the Pre-test with the Main test, the Pre-test was not a reliable indicator of subjects' performance on the test and was therefore dropped from the analysis. Because subjects were randomly assigned to the different testing groups, and because of the large number of subjects in each group, the subject groups were assumed to be equivalent for analysis purposes. The Main Auditory Graph test's Main test was shown to have strong indications of validity and reliability from Split-half analysis, K-R 20 tests, and comparison of the scores of novices to experts. However, Split-half analysis did show poor correlation between the Math section and Physics section tests indicating that these sub-tests were not equivalent tests. If a subject performed well in the Math section, there was no guarantee that he or she would perform well in the Physics section.

ANOVA and Sheffé tests indicated that there were significant differences between the Visual and Sound groups, and between the Both and Sound groups for the Main test as well as each of the sections. After correcting the scores for the possibility of guessing, the difference between the average percent correct scores for the Visual and Sound groups was 15%. Subjects in the Sound group performed at 70% of the level of the Visual group subjects. The difference between the Sound and Both groups 13%. While these are significant differences, the results demonstrate that subjects are able to use auditory graphs to answer many math and physics questions at a fairly high level given very little self-guided training.

The Main Auditory Graph test also demonstrated that blind subjects around the world could not only access the test, but could effectively complete and answer the questions. In addition they were able to do so at a level that exceeded the student subjects, and was not significantly different from local graduate students taking the test

with auditory graphs. Although small in size, the Blind group had an average percent correct score of 75%, which was considerably greater than that of the Visual group, and is not statistically different from the Grad group's score.

11.1.4. Auditory Preference Pilot

The Auditory Preference Pilot test was an initial attempt to determine how well subjects liked the auditory graphing techniques that had been developed, and which elements of the auditory graphs they thought were most useful. The test was as a Web-based tool displaying auditory graphs in several formats, the questions, and related visual graphs. Scripting programs served to generate the questions and record subjects' answers.

The Auditory Preference Pilot test results indicate that subjects' opinions of which items in an auditory graph are important change as they gain familiarity with this new graphing method. Thus, the results indicate the need for flexible graphing displays that have the ability to play the data with and without certain indicators, such as the derivative markers. It was also shown that many of the subjects found the technique of changing the tone quality to indicate when a data value is positive or negative (the zero indicator) was more helpful than the derivative indicators.

11.2. Further Studies Suggested By Test Results

The Auditory Preference Pilot test explored only a few of many areas of interest for future research on auditory graphs. One alternate avenue of research involves multivariate graphs. All graphs in these studies have used single-valued, single data sets. Construction of auditory representations for multiple data sets, for comparisons between data sets, as well as for the display of multi-valued functions needs development

Another important area for further study is an analysis of the effect of training times on performance. The relative differences between the Sound and Visual groups' scores on the Web Pilot and the Main Auditory tests indicate that the amount of training plays a role in auditory graph comprehension. It is unknown how much training is required, or how training times affect the relative performance. Furthermore, there may

be an effect that certain graphing techniques or indicators are only valid or useful under unique circumstances.

Since the initial pilot studies and the Main Auditory Graph test concentrated on simple graphs, and the Auditory Preference Pilot on only marginally more complex graphs, it is unknown at what point the graphing techniques used in these studies are no longer useful. As graphed data become more complex, there may be a preference for audification (directly representing the data values as a wave pattern, and then using that pattern to drive a sound source) or other data sonification methods.

The auditory graphs in these studies allowed only limited control of the sound parameters. Playback rate, the ability to listen to the sound forward or backward, and point by point control of data sonification, could effect graph comprehension. These points demonstrate many areas open for future research.

11.3. Practical Application of Results From This Study

Several of the auditory graphing features used in these experiments have had direct application in current software development. The Triangle Calculator is a scientific graphing calculator for Windows'95. This program is an updated version of the DOS Triangle program used in the Triangle Pilot study and is designed to display functions and data sets not only with visual graphs, but also with auditory graphs as well. The Calculator implements the use of the derivative tick-mark display as was found necessary in the initial pilot tests. It also incorporates a method for the user to enable or disable the tick-marks. The Auditory Preference Pilot test respondents indicated that this feature, while useful, became annoying after repeated listening.

In addition, the Triangle Calculator incorporates a method for altering the sound quality when representing negative *y* axis values. This characteristic was met with general approval from the subjects involved in the Auditory Preference Test. Thus, the effect of this research has led to significant changes in the display methodologies employed in real world applications.

11.4. Final Comments

The series of experiments described in this work has been an effort to demonstrate not only why there is a need for auditory graphs, especially in scientific areas such as physics, but how these graphs can be implemented and used. The use of auditory graphs benefits not only visually disabled people who have the right, and with these techniques, the ability for quick access of data displays, but also allows anyone to effectively use the displays with very little training. With the equivalent of a short description and a few examples, subjects demonstrated the ability to perform at a level that was at least 70% of what they would have achieved with visual graphs. With more training or experience with graphs, this can easily be increased to 85% or more. It was also demonstrated that auditory graphs are not limited to displays in research laboratories with fixed environments, but can be effectively utilized throughout the country and world. The Main Auditory Graph test demonstrated that subjects do not have to be sighted to accomplish this feat.

Auditory graphs hold great promise as a display technique. The Auditory Preference Pilot test demonstrated some of the many areas that future research can be focused on to provide for even more effective displays. Finally, here are two last quotes from subjects. The first is from the Main Auditory Graph test:

"I think the whole idea is great and I think the drum beats to show curvature and slope are particularly functional and innovative. It is really important to develop the ability to hear negative values."

The last comment is from the Auditory Preference Test demonstrating the accomplishment of the previous subjects' request:

"Again, I really like the negative value changing tone. It really helped to see the graph with my eyes closed."