Multi-Sensory Cognitive Learning as Facilitated in a Multimedia Tutorial for Item Response Theory

Chong Ho YU, Samuel DIGANGI, Angel JANNASCH-PENNELL, Victoria STAY, WenJuo LO, Zeynep KILIC
Applied Learning Technology Institute, Arizona State University
Tempe AZ 85287, USA

ABSTRACT

The objective of this paper is to introduce an application of multi-sensory cognitive learning theory into the development of a multimedia tutorial for Item Response Theory. The cognitive multimedia theory suggests that the visual and auditory material should be presented simultaneously to reinforce the retention of learned materials. A computer-assisted module is carefully designed based upon the preceding theory and also an experiment was conducted to examine the effect of audio types (human audio, computer audio, and no audio) on learner performance measured by an objective test. It was found that while there is no significant performance gap between the human audio and the no audio group, the two groups substantively outperform the computer audio group. A plausible explanation is that un-natural audio requires additional cognitive power to process the information and thus this distraction affects the performance.

Keywords: Multimedia, Hypermedia, Multi-sensory, Cognition, Cognitive Psychology, Item Response Theory, Measurement, Assessment.

1. INTRODUCTION

Since the introduction of the No Child Left Behind Act, assessment has become a pre-dominant theme in the US K-12 system. Schools that fail to demonstrate improvement in their students' test scores may eventually be restructured or even taken over by the state [1]. As a result of the highs stakes involved in assessment, many school districts have taken it upon themselves to develop their own assessments in order to identify and provide extra assistance to low performing students. The test developers within the school district are typically teachers who have no background in measurement theories and already have a full-time job within their classroom during the year. Consequently, the items and tests that are being developed may not be not valid or reliable measures of students' performance.

This study addresses the need to make such statistics accessible to K-12 teachers by providing a multimedia tutorial that helps them interpret their students' and the class' performance while also helping them to identify problems in test authoring so as to write better test items for future assessments. The tutorial is designed to help teachers understand and interpret the psychometric analysis of district tests by teaching item response theory (IRT), one of the most popular measurement theories in the field of educational assessment [2]. Unlike the classical true score theory, in which item difficulty is based upon the pass rate, IRT performs item attribute calibration and student ability estimation simultaneously, and thus it is considered a superior tool to the classical approach.

2. MULTISENSORY LEARNING

In order to make IRT understandable for teachers with no background in measurement theories, effective and user friendly instructional materials are needed. Computer-based materials have been developed by researchers and instructors to help students better understand the complex concepts in statistics and psychology courses [3, 4, 5]. Previous literature has explored the different ways people learn with multimedia applications. Multimedia may be defined as the combination of various types of media including text, images, sounds, voice, and video, integrated into a multisensory presentation that conveys various types of material [6, 7, 8]. A defining characteristic of multi-sensory learning is that it occurs when more than one sense is activated in the learning process.

Various modalities utilized in multimedia applications have been studied in order to identify the most effective combinations in facilitating learning. Mayer & Moreno [9] suggest that a modality effect exists, which posits that individuals learn more when they receive both visual images and narration of text than individuals who receive the same material presented only visually and as on-screen text. The modality effect is based on working memory models which state that visual and auditory materials are processed in different areas of the working memory and both subsystems have a limited processing capacity. Therefore, using only visual or only auditory materials limits the processing capacity that is available, whereas employing both visual and auditory materials provides greater processing capacity and the material can be accessed from two areas of the memory as opposed to only one [7].

Mayer & Moreno [10] offered a cognitive theory of multimedia learning that integrates dual coding theory, cognitive load theory, and constructivist learning, which provides a basis for designing the most effective multimedia instructional materials. Dual coding theory states that visual and auditory materials are processed in different cognitive systems [11]. The cognitive load theory states that there is a limit to the amount of information that can be processed by the visual and auditory systems, and providing too much information with text, pictures, or sounds can overload the systems and inhibit learning [12].

Finally, the constructivist learning theory states that more meaningful learning occurs when individuals take relevant information from the material and integrate it with some of their other knowledge [9]. Based on the cognitive theory of multimedia learning, the most effective modality for instruction involves both audio and visual material presented simultaneously, so that the individual can use complete processing capacity as well as develop a visual and auditory representation of the material which the individual can use to make connections between the material [13, 14].

Likewise, the cognitive multimedia theory suggests that the visual and auditory material should be presented simultaneously to allow the individual to make connections between the types of material rather than presenting the material successively. Lindstrom [15] found that participants could only remember 20% of the total materials when they were presented with visual material only, 40% when they were presented with both visual and auditory material, and about 75% when the visual and auditory material were presented simultaneously. Similarly, Lee and Bowers [16] conducted a study with university students to determine the best combinations of media for learning. Compared to a control group, the pre-tests and post-tests of the treated groups revealed 12% more learning while reading printed text alone; 32% more learning while hearing spoken text and reading printed text; 46% more learning while hearing spoken text, reading text, and looking at graphics; 56% more learning while reading printed text and looking at graphics; 63% more learning while looking at graphics alone; and, 91% more learning while hearing spoken text and looking at graphics.

Research also suggests that adding extraneous sounds or visual stimuli that are not relevant to the material do not add to the ability of the individual to learn [7, 17, 18]. Many instructors believe that adding interesting facts or details to a boring presentation will make students more interested thereby increasing their ability to learn the material. However, the cognitive load theory states that there is a limit to the processing capacity of the visual and auditory systems [12, 14] found that participants receiving only narration and animation performed significantly better than groups receiving narration, animation, and integrated text , or separated text, on both measures of retention and application of the learned material. In addition, Mayer et al. [7] found that irrelevant video clips integrated into multimedia instructional material resulted in less retention of information although the result did not reach significance. By adding background music to a tutorial, Brünken, Plass, & Leutner [19] were able to examine the effect of extraneous and irrelevant audio on student's reaction time while simultaneously completing a task. The results indicated that the addition of the narration in the tutorial with the background music resulted in decreased reaction time during the task. This provides support for the cognitive overload theory and the modality effect, which suggest that the auditory and visual systems have a limited capacity to process information.

Further, some research has indicated that visual and audio integration does not result in increased learning. In a study by Koroghlanian & Klein [20], an instructional program for biology was given in four forms: one with text and static illustrations; one that had a bulleted outline accompanied by audio narration of the text; one that had text, illustrations, and animated instructional sequences; and one that had a bulleted outline, audio narration, and animated instructional sequences. Results indicated no significant differences between the types of instructional modes on a post-test measure. Similarly, Veronikas and Maushak [21] found no significant differences in learning for the three different modalities (text, audio, or a combination of text and audio) in college students' test scores following a tutorial on software application. However, they did find that students preferred to learn computer application with dual modalities (text and audio). The lack of significance detected in both studies may have been due to inadequate sample sizes and in turn decreased power. Another possible explanation for differing results may have been due to the complexity of the material covered in the multimedia presentation. Tabbers, Martens & van Merriünboer [22] tested the modality effect through a multimedia tutorial of non-technical subject matter, instructional design. Participants receiving visual text reported more mental effort while taking the tutorial than participants receiving the information through audio. However, participants in the visual conditions group scored significantly higher on a test of retention of the material and their ability to apply the material than participants in the audio conditions group. These results suggest that visual text may be more useful with non-technical subject matter.

3. USABILITY IN MULTIMEDIA

The usability of multimedia applications should also be considered when developing instructional materials. Usability is defined as the combination of a number of factors that affect the quality of a user's experience when using a particular program or system. These factors may include ease of learning, effectiveness, efficiency, error frequency, and satisfaction. Usability testing addresses these factors through a variety of methods by looking at how users interact with the prototype. It is usually an iterative process where participants are tested and the prototype is changed based on their feedback or test results [23] (and Usability.gov).

In a study that tracked eye-movement patterns during multi-media presentations, Faraday and Sutcliffe [24] provided guidelines for optimizing learning. These included using speech to reinforce an image; avoiding animation when a label is being mentioned; and, using animation to show results, as well as process. Najjar [25] points out that more interactive media such as user manipulation and periodic quizzes facilitate better learning.

The National Cancer Institute evaluated five different types of multimedia formats for educating people about lung cancer including text paperback booklet, paperback booklet formatted in HTML on the Web, spoken audio alone, spoken audio synchronized with text Web page, and Flash multimedia with animation, spoken audio, and text [26]. There were five testing session, one for each format, with 9 participants per session - 45 participants overall. Participants were shown their assigned program in its entirety; pre-test and post-test multiple-choice quizzes assessed participant learning. Participants were also given design description and short demonstrations of the other four formats. They were asked to rank preference for the five program formats (1-5) along with providing structured and open-ended comments about the usability of each format. Learning improved with the use of all formats, and Flash was preferred by 71.1% of the users regardless of user characteristics.

Loranger and Nielsen [27] conducted a usability study on 46 Flash applications including e-commerce, configurators, news and current events, maps and location finders, e-learning, entertainment, and productivity applications. Overall they found Flash to be a legitimate platform for complex web-based applications. Their results pointed to the ephemeral nature of Flash as an implementation technology used in web-based application. They found that 36% of users did not even make it from the main website to the actual application because the link was difficult to find or too flashy – reminiscent of an advertisement; to combat this they suggest making the link to the application basic text. Among those who did open the application, it was found that users rarely used the application more than once; so maximum impact on the first use is vital. Positive and negative findings were associated with the use of sound and animated objects. Both MacGregor [28] and Loranger & Nielsen [27] suggest using sound and animation judiciously.

The purpose of the following study is to develop a multimedia Flash tutorial on IRT, which is both user-friendly and grounded in cognitive processing theories, in order to maximize the effectiveness of the learner's ability to retain and apply the concepts described in the tutorial.

4. PROGRAM DESCRIPTION

This hypermedia tutorial, which is composed of two modules, is developed with the use of Macromedia Captivate®, Macromedia Flash®, Adobe PhotoShop®, SAS®, SPSS®, Microsoft Excel®, and Microsoft PowerPoint®. Hypertext and multimedia are two major features that are commonly found in many computer-assisted tutorials. However, rich media, such as over-use of animation modules, could lead to cognitive overload [12]. In addition, improper use of hypertext may interfere with instruction. Without prior knowledge pertaining to the subject matter, non-linear jumping across slides may not lead to a full understanding of the material [29]. Hence, contrary to popular practice, the introductory and the Table of Content page of this tutorial emphasizes the following: "Since some concepts are interrelated, readers are encouraged to go through the tutorial in a sequential manner" (Figure 1).

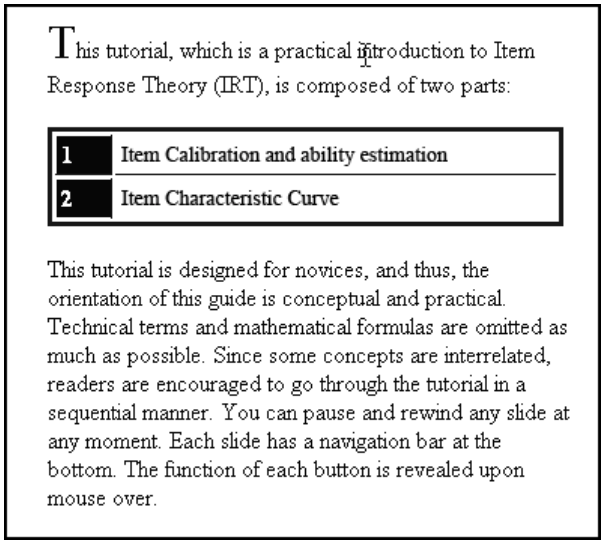
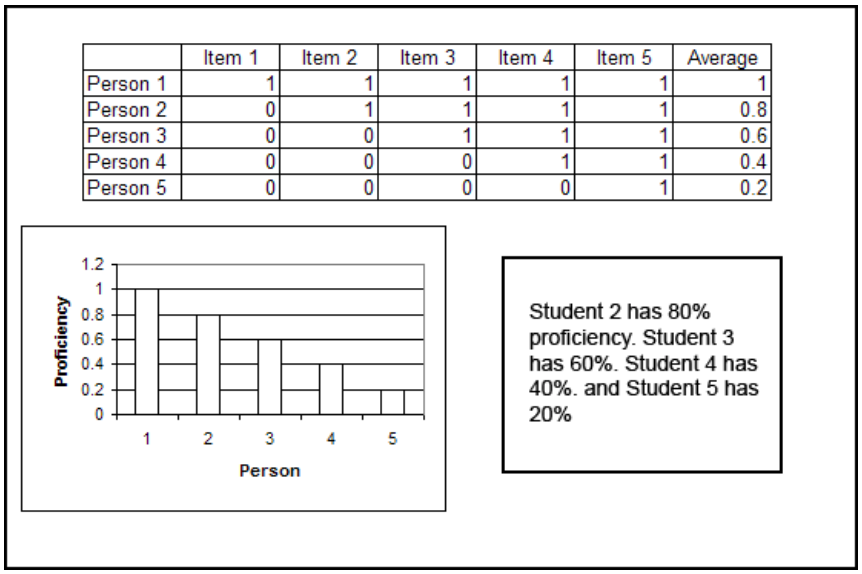


Figure 1. TOC of the tutorial

The content of the tutorial is based on a guide to IRT [30], which is cross-posted on the author's website and *Scientific Software International* ® website. The original text is composed of five chapters but this tutorial, as a pilot project, is reduced to two chapters only. The target audience for this program is undergraduate education students who have learned the basic concepts of statistics. Chapter One is concerned with item calibration and ability estimation whereas Chapter Two pertains to Item Characteristic Curve (ICC).



	Item 1	Item 2	Item 3	Item 4	Item 5	Average
Person 1	1	1	1	1	1	1
Person 2	0	1	1	1	1	0.8
Person 3	0	0	1	1	1	0.6
Person 4	0	0	0	1	1	0.4
Person 5	0	0	0	0	1	0.2

Figure 2. 5x5 item-person matrix

Chapter one starts with a scaled down yet simple example: A data set with five items and five students only. Many instructors use real data sets to illustrate item calibration and ability estimation in a complex simulated environment, and as a result, students may experience cognitive overload. Therefore, the example used in this tutorial was simplified to increase understanding of the material. The example in Figure 2 is ideal as no item parameter can be estimated when all students could answer Item 5 correctly because there is no variation in the distribution.

Nevertheless, many successful scientific "thought experiments" start from "idealization," in which the conditions do not correspond to the real world. In spite of using hypothetical cases, insight may still be gained when idealization makes every variable so simple that the user may "mentally manipulate" them without difficulty [31]. At the end of this session, the tutorial emphasizes that the example is an ideal case that is too good to be true (Figure 3).

	Item 1	Item 2	Item 3	Item 4	Item 5	Average
Person 1	1	1	1	1	1	1
Person 2	0	1	1	1	1	0.8
Person 3	0	0	1	1	1	0.6
Person 4	0	0	0	1	1	0.4
Person 5	0	0	0	0	1	0.2

This example is an ideal case, in which more proficient students answer all items correctly, and less proficient students answer the easier items and fail the hard ones. However, these results rarely occur in reality, and are just "too good to be true." This ideal case is known as the **Guttman pattern**.

Figure 3. Guttman pattern: Ideal case

In Chapter Two, again we adopt the preceding strategy by presenting theoretical modeling but hiding empirical data. In testing regression, it is a common practice for instructors to overlay the data points and the regression line to offer a visual depiction of residuals. However, it would not work well in this situation, because in IRT there is person misfit and item misfit; in each of these categories there is model fit and individual fit; and these may be further analyzed through infit and outfit. The learner will most likely experience cognitive overload if the model and the data are presented together. Hence, our instructional strategy is to illustrate modeling with nice and clean graphics. For example, Figure 4 shows a typical ICC. The tutorial emphasizes that ICC depicts a theoretical modeling where, for instance, in the actual sample there may be no students with -5 skill level. Nonetheless, these extreme cases in a "what-if" scenario could clearly illustrate the point that if the person does not know anything about the subject matter, he or she will have zero probability of answering the item correctly.

The tutorial demonstrates idealization and modeling while also stressing the practical applications of IRT. One of the nice features of IRT is that the parameter values are centered at zero and thus the visual representation of item difficulty is very easy to interpret. For example, Figure 5 is a screenshot about how IRT can be applied to test construction by selecting items with different difficulty levels. As you can see, the bars of the

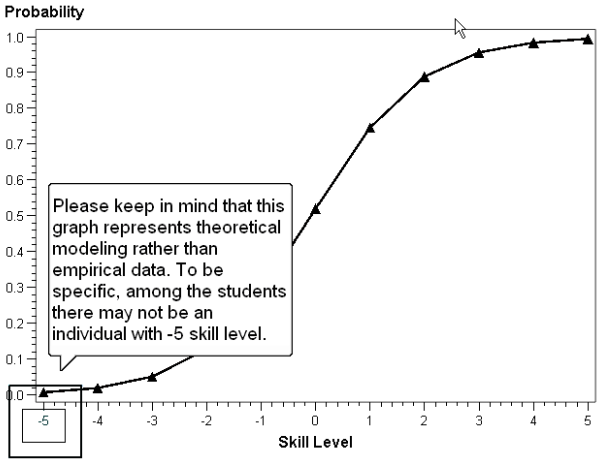


Figure 4. ICC

average items center around zero, hard items are located at the right side, and easy items are placed on the left side. It is notable that this visually compelling illustration is not used by popular IRT programs, such as Bilog, Winsteps, and RUMM. The bar chart in Figure 5 is generated in a SAS macro code written by Yu [32] and is imported into the multimedia tutorial.

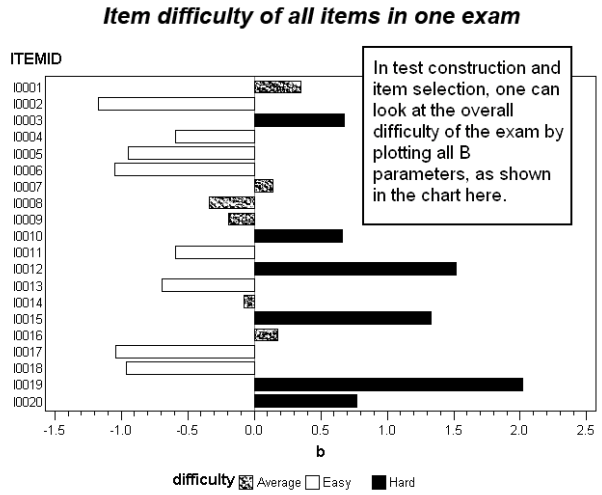


Figure 5. Item difficulty of all items

The computer-based multimedia program is accessible at http://www.creative-wisdom.com/multimedia/IRTTHA.htm. Yu [31] has also prepared a PDF document that presents much of the program content. A version of this document can be viewed at http://www.creative-wisdom.com/computer/sas/IRT.pdf

The multimedia program reflects our pursuit to provide a timely training tool for educational assessment while also enhancing statistical education. Use of the application and dialogue on this topic are encouraged.

5. METHOD

Participants

Participants were 26 students who lacked substantial prior knowledge about Item Response Theory (IRT) recruited from various departments at Arizona State University. Students of different majors were selected for the study because it can broaden generalizability of this study. Participants were randomly assigned in approximately equal numbers to one of three conditions: the human audio (HA) group, the computer audio (CA) group, or the no audio (NA) group. Participation in this study was voluntary and they were informed that they could withdraw from the experiment at any time without penalty. Nonetheless, no students terminated their participation.

The mean age for participants was 24 years (SD = 0.00), ranging from 18 to 38. The gender of participants in the HA group was 28.57% male vs. 71.43% female, 85.71% male vs. 14.29% female for the CA group, and 83% male vs. 17% female for the NA group. The percentage of graduate students was 20% in the HA group, 12% in the CA group, and 8% in the NA group. The mean GPA was 3.65 for the HA group, 3.14 for the CA group, and 3.41 for the NA group.

Measures

The computer-based materials used in this study included an IRT tutorial consisting of three versions (HA, CA and NA) of a multimedia-training program with the same information about IRT and an on-line evaluation regarding the material. The IRT tutorial is comprised of two short animations related to item calibration and item characteristic curves. The on-line evaluation consisted of 20 questions, asking participants about the lesson they had learned in a manner that made them apply it to a novel situation (see Appendix A). These 20 multiple-choice questions were divided into two subsets: the first subset was related to item calibration and the second subset was related to item characteristic curves. The validity of measurement is commonly affected by fatigue; i.e. in an exam that would not affect their GPA, test takers tend to devote more efforts at the beginning, but pay less attention or even rush through the items near the end of the test. As a remedy, the item order is randomized so that no items will always be located at the end.

The display environment includes two major parts: 1) a multimedia panel – for displaying the animated diagrams with or without concurrent narration, and 2) a control panel – allowing the participant to navigate animation with the control functions of pause, forward, or backward.

Procedure

The participants were tested individually and were randomly assigned to one of the three groups. When a participant entered the experimental lab, he or she was seated in a cubicle. Instructions were read aloud by administrators. The administrators then asked the participant to adjust his/her seat, volume of voice in the headphone or the speaker, and display angle of the monitor. When the training program was complete, administrators directed participants to another browser to answer the questions on the test. Participants were given no time constraints to complete both the animation and measure sections, but on the average, the administration lasted 45 minutes.

Data analysis

Descriptive statistics were used to describe participants' profiles. Item difficulty of each item was checked and extremely difficult and extremely easy items were removed from the test. It is assumed that the difficulties of Chapter 1 and 2 exams are equivalent and thus a two-dependent-sample t-test was employed to compare the mean scores. Since there were subject recruitment constraints, the sample size of this study was limited. In order to check whether the comparison in this study is sensitive enough to detect the group difference, a post hoc power analysis was conducted using the η^2, which is the effect sum of squares divided by the total sum of squares [33].

A one-way analysis of variance was conducted to evaluate the relationship between audio types on IRT tutorial and the performance on test. Further, a post-hoc procedure was conducted to evaluate any existing significance of ANOVA results. In addition, diamond plots available in JMP [34] were employed to visually investigate the mean and variance differences, and also the confidence intervals of the means among the three groups.

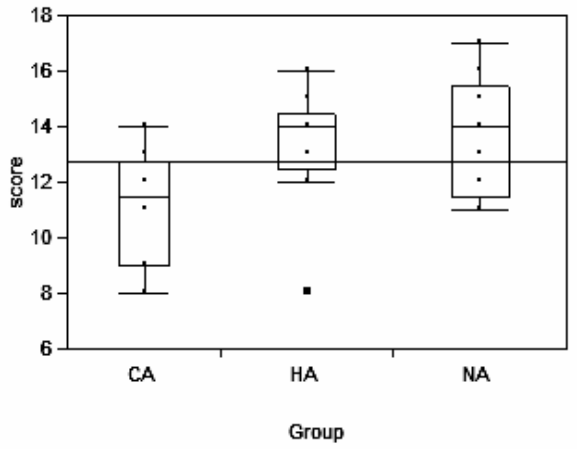
To compensate for the low n and power level, the randomization exact test (RET) available in StaXact [35] was utilized. While bootstrap is resampling with replacement, RET is resampling without replacement. In the classical procedure the F statistics is compared against the F-critical in the F-distribution to determine whether the group difference is significant. However, in RET, instead of consulting a theoretical F-distribution, the researcher asks a "what-if" question: "It may just happen that an over-achiever takes the NA version by chance, and an under-achiever, takes the CA version by chance, too. What if their positions are swapped?" In this process, all or many possible arrangements of the subjects into different groups are enumerated. After exhausting 500 possible combinations, the F-values were used to plot an empirical distribution curve, which was built on the empirical sample data. The original statistics was compared against this empirical distribution to yield the exact p value [36]. Further, since the focal point of the study is multi-sensory and use of audio rather than no audio is important in multimedia research, RET was extended to the comparison between HA and CA. It is important to note that the 26 subjects were re-used numerous times, and hence conventional degrees of freedom could not be applied here.

6. RESULTS

Within Chapter 2, none of the students were able to answer one of the items correctly whereas all of the students answered another item correctly. Since there was no variability in both items, their psychometric attributes are unknown. Therefore, they were not counted toward the total test scores. Before the removal, the one-sample t-test shows that there is a performance gap between the items on chapter 1 and 2; $t(26) = -3.39$, $p = .0023$. After the removal, the performance gap disappeared; $t(26) = -1.01$, $p = .32$. And thus the aggregated test scores without those two items are used for the analysis.

An outlier based on the five-point summary (boxplots) were excluded from the subsequent analysis. This outlier is indicated in Figure 6.

Figure 6. Boxplots showing quantile information by group



The ANOVA was significant, F(2,24) = 6.18, p = .0074. The strength of relationship between the type of audio and the test's performance, as assessed indicated by the effect size (η^2 = .36) was strong, with the audio factor accounting for 36% of the variance of the dependent variable.

Post hoc tests were conducted to evaluate pairwise differences among the means. Because the standard deviations among the three types of audio groups ranged from 1.2 to 2.14 and variances among those groups ranged from 1.44 to 4.58, we chose not to assume that the variances were homogeneous and conducted post hoc comparisons using the Dunnet's C test, a test that does not assume equal variances among the three groups. The results of these tests, as well as the means and standard deviations for the three audio groups, are reported in Table 1.

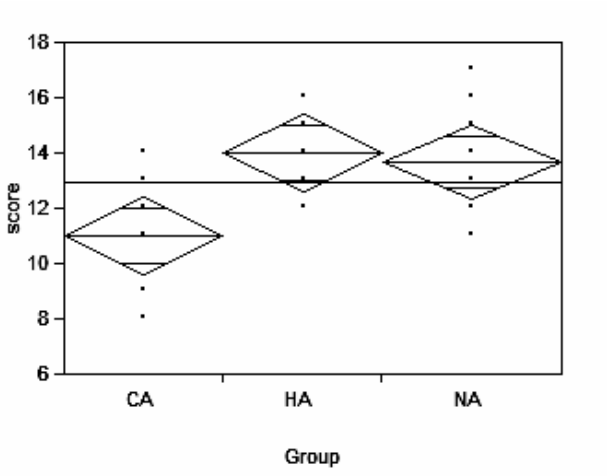
Table 1. Post hoc test results using Dunnett's C

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound	Upper Bound	Lower Bound
CA	HA	-3.00(*)	.866	-5.55	-.45
	NA	-2.67	1.035	-5.67	.34
HA	CA	3.00(*)	.866	.45	5.55
	NA	.33	.824	-2.04	2.71
NA	CA	2.67	1.035	-.34	5.67
	HA	-.33	.824	-2.71	2.04

Based on observed means.
* The mean difference is significant at the .05 level.

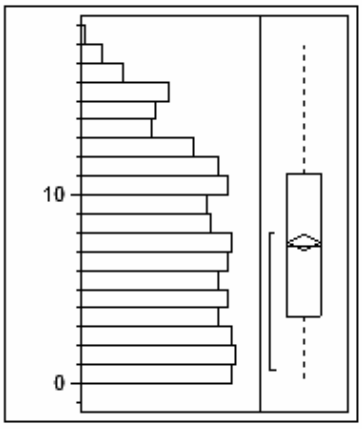
The post hoc analysis is also illustrated by the diamond plots, as shown in Figure 7. In Figure 7, the grand sample mean is represented by a horizontal line across all three groups while the group means are illustrated by a line inside each diamond. And the confidence intervals (CI) for each group are symbolized by the diamond. The visualization of CI is straightforward. The flatter the diamond is, the tighter the CI is. In this analysis, it is obvious that in the performances of HA and NA groups are comparable but there is a significant difference between the former two and the CA group. While the CIs of the former two are basically the same, the lower bound of the former two and the upper bound of the CA barely overlap, which indicates that the performance of the CA group significantly differ from the other two groups.

Figure 7. Diamond plots showing post hoc comparison



The effect size in terms of η^2 is .36. Given that the sample size is 26, the power level for this study is .33, which is not strong enough to detect a true difference. As a remedy, RET with 500 re-samples was used. The result was slightly different from that of conventional ANOVA, with F = 8.36, exact p = .0077, which is significant. Further, HA and CA were compared using permuted t-test, with t = 88, exact p = .0065, which is also significant. Figure 8 shows the distribution of the F-values resulted from 509 resamples.

Figure 8. Distribution of F-values resulted from 500 resamples



The relationships between the IRT test performance and user demographic backgrounds were explored. The average GPA of the CA group is the lowest among the three groups, and there is a strong correlation between GPA and score, r = .613, p =

.0068. At first glance, the poorer performance of the CA group may be attributed to their academic ability.

Figure 9(a). Bar chart of academic level

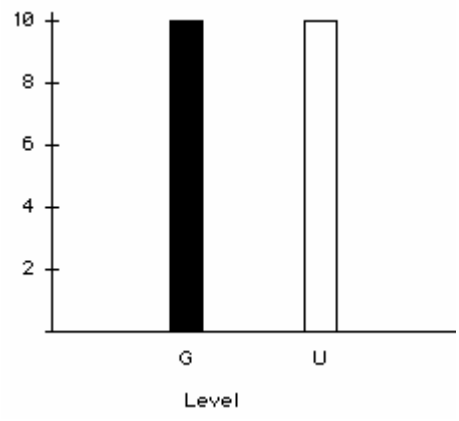


Figure 9(b). Scatterplot of GPA and test score

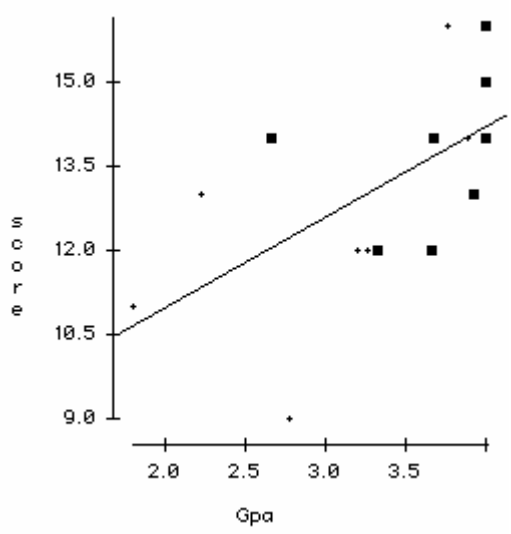
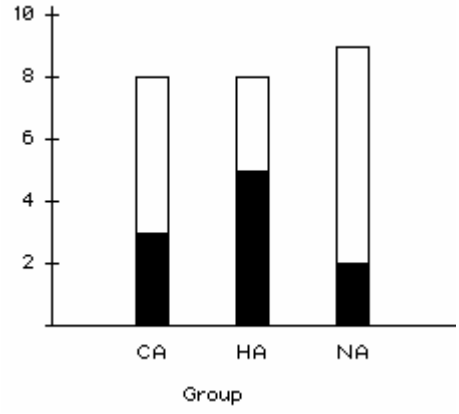


Figure 9(c). Bar chart of grouping by audio type



However, it is important to note that most of the students who did well in the test were graduate students whose GPA was high. In Figure 9a, b, and c, observations from graduate students are highlighted. Among those graduate students whose both GPA and IRT test scores are high, actually more of them were assigned into the CA group than the NA group (see Figure 9c).

7. CONCLUSION

In spite of the low power of the test and lack of generalizability of the results, this exploratory study provides useful insight to pave the path for further investigation. While it was not surprising to see that the mean score of HA group was significantly higher than that of the CA group, it was not within our expectations to find that the NA group outperformed the CA group. Through informal feedback channels, many users complained that listening to the computer voice and reading the text simultaneously is distracting, but this type of resentment was hardly found in the HA group. While the relationships among GPA, academic level, and test performance require further investigation, another plausible explanation is that un-natural audio simulated by computer requires additional cognitive power to process the information, and as a result, this extra cognitive load drags down the performance level.

8. ACKNOWLEDGEMENTS

Special thanks to Ms. Kristina Zeller and Mr. Charles Kaprolet for editing this article, to Ms. Lori Long and Mr. Chang Kim for their assistance in developing the multimedia program, to Ms. Gemma Garcia, Mr. Ruvi Wijesuriya, and Ms. Gladys Caquimbo for their assistance in recruiting subjects for this study, and to Mr. Jin Gong and Ms. Kristina Zeller for helping run the experiment.

9. REFERENCES

[1] M. Goertz. & M. Duffy, “Mapping the Landscape of High-Stakes Testing and Accountability Programs”, **Theory into Practice**, Vol. 42, 2003, pp. 4-11.

[2] S. Embretson & S. Reise,, **Item Response Theory for Psychologists**, Mahwah, N.J.: L. Erlbaum Associates, 2000.

[3] N. Hammond, J. McKendree, W. Reader, & A. Trapp, A., **The PsyCLE Project: Educational Multimedia for Conceptual Understanding**, Proceeding of the ACM Multimedia 95 Conference, San Francisco, CA, 1995.

[4] E. Morris, “The Design and Evaluations of Link: A Computer-Based Learning System for Correlation”, **British Journal of Educational Technology**, Vol. 32, 2001, pp. 39-52.

[5] E. J. Morris, R. Joiner, & E. Scanlon, “The Contribution of Computer-Based Activities to Understanding Statistics”, **Journal of Computer Assisted Learning**, Vol. 18, 2002, pp. 114-124.

[6] M. Neo & K. Neo, “Innovative Teaching: Using Multimedia in a Problem-Based Learning Environment”, **Educational Technology & Society Education**, Vol. 4, No. 4, 2001, pp. 19-31.

[7] R. E. Mayer, J. Heiser, & S. Lonn, “Cognitive Constraints on Multimedia Learning: When Presenting More Materials Results in Less Understanding”, **Journal of Educational Psychology**, Vol. 93, 2001, pp. 187-198.

[8] R. E. Mayer, K. Sobko, & P.D. Mautone, “Social Cues in Multimedia Learning: Role of Speaker’s Voice”, **Journal of Educational Psychology**, Vol. 95, 2003, pp. 419-425.

[9] R. E. Mayer & R. Moreno, "A Split-Attention Effect in Multimedia Learning: Evidence for Dual Processing Systems in Working Memory", **Journal of Educational Psychology**, Vol. 90, 1998, pp. 312-320.
[10] R. E. Mayer & R. Moreno, "Aids to Computer-Based Multimedia Learning", **Learning and Instruction**, Vol. 12, 2002, pp. 107-119.
[11] J. M. Clark & A. Paivio, "Dual Coding Theory and Education", **Educational Psychology Review**, Vol. 3, 1991, pp. 149-210.
[12] J. Sweller & P. Chandler, "Evidence for Cognitive Load Theory", **Cognition and Instruction**, Vol. 8, 1991, pp. 351-362.
[13] S. Tindall-Ford, P. Chandler, & J. Sweller, "When Two Sensory Modes are Better than One", **Journal of Experimental Psychology: Applied**, Vol. 3, 1997, pp. 257-287.
[14] R. Moreno & R.E. Mayer, "Cognitive Principles of Multimedia Learning: The Role of Modality and Contiguity", **Journal of Educational Psychology**, Vol. 91, 1999, pp. 358-368.
[15] R. Lindstrom, **The Business Week Guide to Multimedia Presentations: Create Dynamic Presentations That Inspire**, New York: McGraw-Hill, 1994.
[16] A. Y. Lee & A. N. Bowers, "The Effect of Multimedia Components on Learning", **Proceedings of the Human Factors and Ergonomics Society**, 1997, pp. 340-344.
[17] N. A. Beachman, A.C. Elliott, J.L. Alty, & A. Al-Sharrah, "Media Combinations and Learning Styles: A Dual Coding Approach", **Proceedings of the ED-MEDIA 2002 World Conference on Educational Multimedia, Hypermedia & Telecommunications, USA**, 2002, pp. 2-7.
[18] J. L. Alty, "Dual Coding Theory and Computer Education: Some Media Experiments to Examine the Effects of Different Media on Learning", **Proceedings of the ED-MEDIA 2002 World Conference on Educational Multimedia, Hypermedia & Telecommunications, USA**, 2002, pp. 2-7.
[19] R. Brunken, J.L. Plass, & D. Leutner, "Assessment of Cognitive Load in Multimedia Learning with Dual-Task Methodology: Auditory Load and Modality Effects", Vol 32, No. 1-2, 2004, pp. 115-132.
[20] C. M. Koroghlanian & J.D. Klein, "The Effect of Audio and Animation in Multimedia Instruction", **Journal of Educational Multimedia and Hypermedia**, Vol. 13, 2004, pp. 23-46.
[21] S.W. Veronikas & N. Maushak, "Effective of Audio on Screen Captures in Software Application Instruction", **Journal of Educational Multimedia and Hypermedia**, Vol. 14, No. 2, 2005, pp. 199-205.
[22] H. K. Tabbers, R.L. Martens, & J.J.G. van Merrienboer, "Multimedia Instructions and Cognitive Load Theory: Effects of Modality and Cueing", **British Journal of Educational Psychology**, Vol. 74, 2004, pp. 71-81.
[23] B. Bailey, **The Value of Iterative Design**, Retrieved March 26, 2006, from Web Usability: http://www.webusability.com/article_value_of_iterative_design_7_2005.htm, 2005.
[24] P. Faraday & A. Sutcliffe, "Designing Effective Multimedia Presentations", **Proceedings of CHI '97**, 1997, pp. 272-278.
[25] L. J. Najjar, "Principles of Educational Multimedia User Interface Design", **Human Factors**,Vol. 41, 1998, pp. 311-323.
[26] J.L. Bader & N. Strickman-Stein, "Evaluation of New Multimedia Formats for Cancer Communications", **Journal of Medical Internet Research**, Vol. 5, 2003, p. 16.
[27] H. Loranger & J. Nielsen, **Usability of Flash Applications and Tools: Design Guidelines for Flash Based Functionality on the Web**, Retrieved March 26, 2006, from Nielsen Norman Group: http://www.NNgroup.com/reports/flash, 2002.
[28] C. MacGregor, **Developing User-Friendly Macromedia Flash Content**, Retrieved March 26, 2006, from Macromedia, Inc./MacGregor Media: http://www.macromedia.com/software/flash/productinfo/usability/whitepapers/usability_flazoom.pdf., 2001.
[29] C. H. Yu, **Use and Effectiveness of Navigational Aids in Hypertext**, Unpublished master's thesis, University of Oklahoma, Norman, Oklahoma, 1993.
[30] C. H. Yu, **A Simple Guide to the Item Response Theory**, Retrieved June 5, 2006, from http://www.ssicentral.com/irt/resources.html, 2006.
[31] J. R. Brown, "Why Thought Experiments Transcend Empiricism", In Christopher Hitchcock (Ed.), **Contemporary Debates in Philosophy of Science**, pp. 21-43, MA: Blackwell, 2004.
[32] C. H. Yu, "SAS Programs for Generating Winsteps Control Files and Web-Based Presentations", **Applied Psychological Measurement**, Vol. 30, 2006, pp. 247-248.
[33] G. J. Devilly, **Gpower: International MS-DOS Version (Computer Program)**, Centre for Neuropsychology, Swinburne University, Australia, 1998.
[34] SAS Institute, **JMP Version 6 (Computer Program)**, Cary, NC: The Author, 2005.
[35] Cytel, Inc., **StaXact version 7**. (computer program). Cambridge, MA: The Author, 2005.
[36] C. H. Yu, "Resampling methods: concepts, applications, and justification", **Practical Assessment, Research & Evaluation**, Vol. 8, No. 19, 2003. Retrieved March 6, 2007 from http://PAREonline.net/getvn.asp?v=8&n=19.

10. APPLENIX: TEST ITEMS

There are no question numbers because the item order is randomized.
* Removed items

The B parameter is also known as the _____

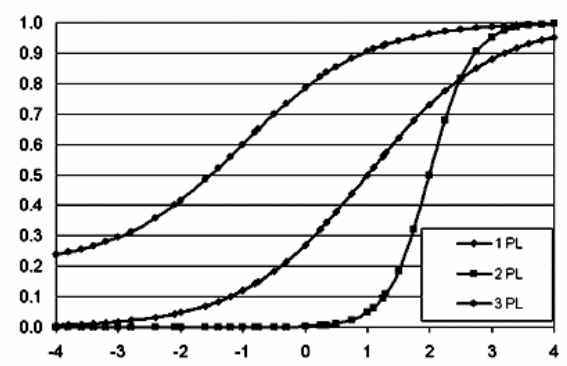
a. A parameter
b. G parameter
c. Discrimination parameter
d. Item difficulty parameter
e. Lord's paradox

Which of the following is not an example of the Guttman pattern?

a. More skilled students could answer all hard and easy items correctly
b. Less skilled students could answer all easy items correctly but failed the hard items.
c. Non-skilled students failed all hard and easy items.
d. Less skilled students could answer some hard item and some easy items correctly.

Which item has a higher guessing rate?

a. Red item
b. Blue item
c. Green item
d. There is not enough information.



Person 1 and Person 2 have the same amount of correct answers (60%). Which one is more likely to be a better student?

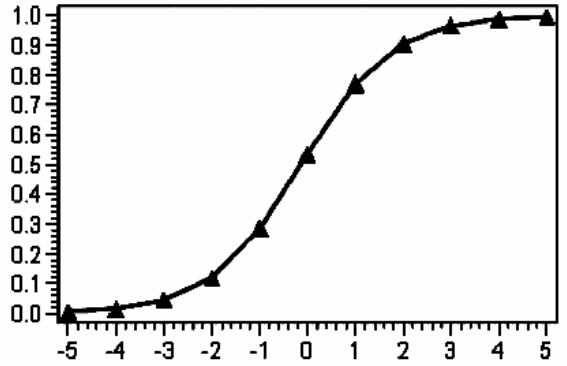
	Hard	**Hard**	**Ave**	**Easy**	**Easy**	**%**
Person 1	0	1	1	0	1	60
Person 2	1	0	1	1	0	60

a. Student 1
b. Student 2
c. They are equally good.
d. There is not enough information.

Which item can do a better job of distinguishing among student abilities?

a. Blue item
b. Green item
c. There is not enough information.

Figure 1



What does the Y-axis (vertical axis) of Figure 1 represent?

a. Student skill level
b. Probability
c. Item characteristic
d. Guttman

What does the X-axis (horizontal axis) of Figure 1 represent?

a. Student skill level
b. Probability
c. Item characteristic
d. Guttman

What does the red curve of Figure 1 represent?

a. Student skill level
b. Probability
c. Item characteristic
d. Guttman

** Figure 1 reflects _____*

a. empirical data
b. theoretical modeling
c. empirical modeling
d. theoretical data

** In Figure 1, if the student skill level is 0, what is the probability of answering the item correctly?*

a. 0
b. 0.05
c. 0.5
d. There is not enough information

When the data and the model do not exactly match each other, what process should be involved to make them eventually converge?

a. calibration
b. point estimation
c. interval estimation
d. iterative pattern

If the tentative student proficiency is .7 and the tentative item difficulty is also .7, what will the probability of passing the item be?

a. .7
b. .6
c. .5
d. .3
e. .25

Which item may be poorly written?

1: Correct
0: Incorrect

	Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**Item 6**
Person 1	1	1	1	1	1	0
Person 2	0	0	1	1	1	0
Person 3	0	0	0	1	1	0
Person 4	0	0	0	0	1	0
Person 5	0	0	0	0	0	1

a. Item 1
b. Item 2
c. Item 3
d. Item 4
e. Item 5
f. Item 6

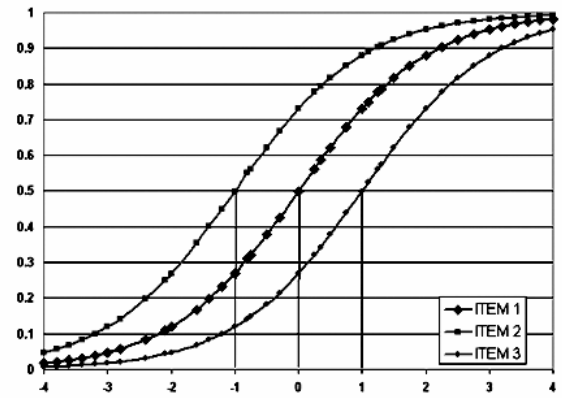
If the student skill level is 0, what can we tell about this student?

a. The student does not know the material at all.
b. The student knows some of the material.
c. The student knows all of the material.
d. There is not enough information

What piece(s) of information do we need to predict the probability of answering an item correctly given a particular ability level?

a. tentative student proficiency
b. tentative item difficulty
c. tentative probability
d. tentative student proficiency and item difficulty

Which item is the most difficult?



a. Item 1
b. Item 2
c. Item 3
d. There is not enough information.

If a more skilled student fails an easy item while a less skilled student could answer it correctly, what is the most likely reason?

a. The ability estimation is incorrect.
b. The item difficulty estimation is incorrect.
c. Both the ability estimation and the item difficulty estimation are incorrect.
d. The wording in the item confuses students.

If one out of ten students could answer item 1 correctly, what is the tentative item difficulty of item 1?

a. 0.1
b. 0.2
c. 0.9
d. There is not enough information.

Why can't we judge a student's ability based on how many items he or she can answer correctly? Because we haven't _____

a. taken the Guttman pattern into account.
b. looked at the tentative student proficiency.
c. taken item difficulty into account for ability estimation.
d. checked the data-model convergence.

Which of the following is not included in a 2-parameter IRT model?

a. A parameter
b. G parameter
c. Discrimination parameter
d. Item difficulty parameter
e. Lord's paradox