Graphics based Computer Adaptive Testing and Beyond

Irene Cheng [1,2] and Anup Basu[2]

[1]Department of Computer and Information Science, University of Pennsylvania, USA
[2]Department of Computing Science, University of Alberta, Canada
Contact: chenglin@seas.upenn.edu, lin@cs.ualberta.ca, anup@cs.ualberta.ca; Website: crome.cs.ualberta.ca

Abstract

Instead of computer games, animations, cartoons, and videos being used only for entertainment by kids, there is now an interest in using graphics for "innovative testing." Rather than traditional pen-and-paper tests, audio, video and graphics are being conceived as alternative means for more effective testing in the future. In this paper we review some examples of graphics item types for testing. As well, we outline how games can be used to interactively test concepts; discuss designing chemistry item types with interactive graphics; suggest approaches for automatically adjusting difficulty level in interactive graphics based questions; and propose strategies for giving partial marks for incorrect answers. We study how to test different cognitive skills, such as music, using multimedia interfaces; and also evaluate the effectiveness of our model. A method for estimating difficulty level of a mathematical item type using Item Response Theory (IRT) is discussed. Evaluation of the graphics item types through extensive testing on some students is also described. All of the graphics implementations shown in this report are developed by members of our research group.

Categories and Subject Descriptors (according to ACM CCS): I.3.2 [Computer Graphics]: Distributed/Network Graphics, K.3.2 [Computers and Education]: Self-assessment

1. Introduction

Computer-adaptive testing (CAT) [Nce06, Sch06, Wik03] is an effective mechanism not only to prevent a student feeling stressed, either because the questions are too difficulty or too easy, but also to assist an educator's understanding of a student's ability and to provide suitable timely advice. CAT involves computerized testing with an adaptive component. Adaptability is the ability to "tailor the difficulty level of each question based on the correctness of the previously answered question" [Wik03]. It is "an innovative, online form of assessment in which items are presented in a sequence that is dependent on the correctness of the examinee's responses to the preceding items" [Cas06]. Figure 1 illustrates the CAT concept by using a linked-list data structure grouping questions of equal difficulty in the same bin. The next test item is adaptively selected, either from the more difficult items in the left bins or from the easier items in the right bins, based on the correctness of the responses given by a student. The selection process is more complex than choosing from the neighboring bins. Which bin to select from is governed by the Item Response Theory (IRT).

IRT is a family of mathematical models that describe how students interact with test items [ER00, LH97]. Regardless of the starting difficulty level given to a student, his or her ability can be assessed with a limited number of items as illustrated by the convergence rate of the curve shown in Figure 2.

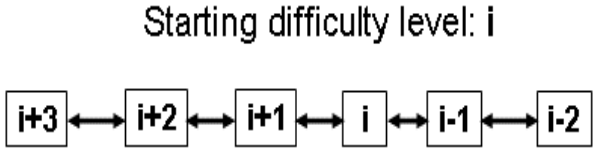

Figure 1: *A strategy for adaptive testing.*

An application can apply one of the three IRT versions: a 1, 2 or 3 Parameter Logistic Model (PLM). The 3-PLM has the following form:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{(1 + e^{-a_i(\theta - b_i)})} \quad (1)$$

Where c_i is the guessing parameter denoting the probability of guessing correctly on this item; b_i is the difficulty parameter; and a_i is the discrimination parameter denoting how well this item can discriminate students of slightly different ability. 2-PLM is obtained by setting c_i = 0, and 1-PLM (Rasch model) is obtained by setting c_i = 0 and a_i = 1.

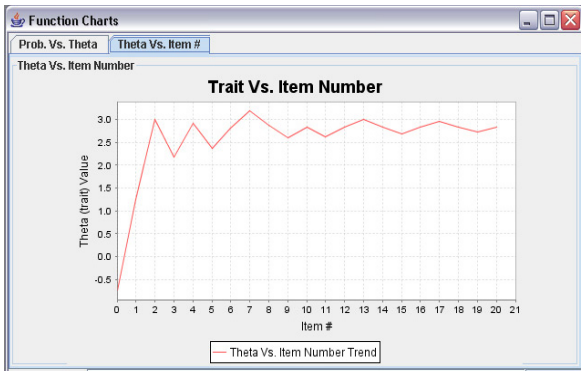

Figure 2: *A snapshot of our interface showing a student's performance based on IRT.*

In addition to questions being selected based on individual student's ability level, CAT has other advantages over conventional pen-and-paper based testing, including: significant cost reduction in administering tests; reduction in test administration time: adaptive methods can estimate a student level much faster, thereby reducing the time needed to administer a test to hours instead of days; immediate test scoring; tests on demand and use of graphics in item types. Detailed review on these topics can be found in [CB06] and thus will not be elaborated here.

Another important aspect of computerized testing is its online digital media component. Audio, video and graphics are being conceived as alternative means for more effective testing in the future [PDP00, ZS02, IB02]. Computer games have been widely used to teach concepts [LH93, Y05]; collaborative augmented reality has been used for math and geometry education [KS03]; an online learning environment [THC94] has been used for the Virtual-U project; virtual reality has been used for medical training and assistance [LPL*05]; education research using web-based assessment has been discussed in [BTB*00]; open exams have been set up for MBA/Business school admission [Syv06]; a virtual environment of water molecules has been used to teach concepts, such as orbits, electron densities, dynamics and so on [TFGT99]; artificial intelligence techniques have been used to recommend research papers to learners [TM04]. However, most of the literature addressed using graphics and multimedia for learning. Other authors, such as [BC07, Cun00, KS96, Tax03], addressed issues relating to teaching computer graphics. The use of graphics in testing has been relatively limited, compared to learning and training. For those systems supporting testing, most of the test items used in current systems still adhere to traditional styles, e.g. True-false, Multiple-choice and Fill-in-the-blank. Our approach on computer adaptive testing differs from other designs discussed in the literature not only because it is enriched with graphics, but also because of the following novelties:

1. The integration with a user-friendly graphics interface for items generation, which facilitates a smooth transition from the traditional pen-and-paper tests to multimedia CAT for item creators and educators.
2. The automatic generation of multiple items and scoring, including partial marks, based on similarity match and predefined parameters.
3. The ability to test not only subject knowledge but also cognitive skills.
4. The use of educational games to get students engaged, to inspire them to learn and make them feel rewarded.
5. The introduction of mobility to CAT.

These aspects characterize innovative item types and our goal is to employ innovative items to inspire students' cognitive powers and make them more engaged in learning. Examples, analysis and evaluations quoted in this paper are extracted from the CROME system implemented by our project team. The remainder of this paper is organized as follows: Sections 2 covers some examples of graphics item types for adaptive testing and explains the item generation process. Section 3 discusses strategies for automatic difficulty level adjustment, scoring and question selection in various cases. In Section 4 we discuss how different types of intelligences can be better measured using graphics and other multimedia based testing. A brief summary of feedback on our graphics item types by students is given in Section 5. Finally, conclusion and future work are discussed in Section 6.

2. Graphics Items and Item Generation

Graphics item types can be used to test a variety of areas. For example, Figure 3 shows a drag and drop (top) biology item, and (bottom) geography item. These items allow students to drag text or graphics to their appropriate locations on the screen. Differing from traditional multiple choice, true/false and fill in the blank type of questions, graphics items are complicated to create. For example, to create a multiple choice item, the item creator only needs to define four choices and identify one as the correct answer (Figure 4). A generic template can be set up and all the questions can be created using the same template by inputting different contents. However, each graphics item is unique and additional parameters are needed to create a question. For example, in the cell item (Figure 3, top) the interface has to know whether the student drags and drops the name to the right location by checking the coordinates of the answer box.

It is a tedious job for the item creator to use an image editor to locate the (x, y) coordinates. To facilitate this process, our generator interface allows the item creator to draw bounding boxes at appropriate locations on the screen by using the mouse. The interface then automatically stores the coordinates. Our interface also allows answer boxes of irregular shapes to be drawn as shown by the contour around "S. America" and the partial contour around "Africa" (Figure 5). Unique layouts corresponding to different graphics item types are implemented as plug-ins in the generator framework.

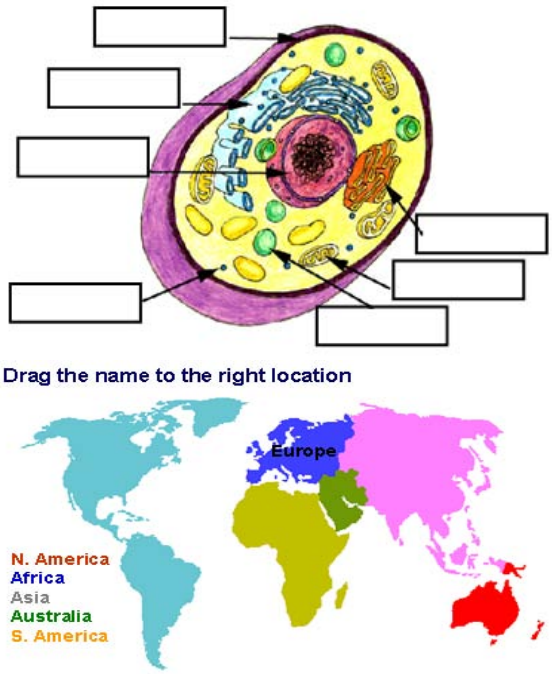

Figure 3: *Examples of drag and drop questions*

The nearest star to Earth (besides our sun) is approximately 4 light years away.
The distance from Earch to that star, in meters, is approximately

- A. 9.5×10^{20} m
- B. 3.8×10^{16} m
- C. 3.8×10^{14} m
- D. 3.8×10^{13} m

Figure 4: *A multiple choice question.*

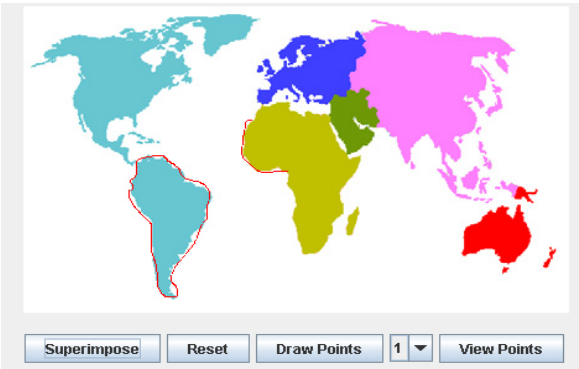

Figure 5: *Our item generator provides user-friendly interaction with the items creator.*

3. Automatic Item Difficulty Level Generation and Scoring

2D pictures can be used in paper-and-pencil formats, but 3D graphics can only be available in digital form. 3D graphics is more intuitive for explaining chemical reactions. In our design, we focus on improving testing at the symbolic and atomic levels. We use 3D objects to test a student's understanding of what atoms and molecules are involved and how they react during various chemical changes. Studies indicate that students tend to experience difficulty with spatially related chemistry problems requiring 3D thinking [TSB91]; thus, graphics item types can assist in understanding some of these processes. The atomic level involves the understanding of molecular structures and the change of structures during a chemical reaction, such as breaking bonds inside a molecule. The objects can be rotated and manipulated in 3D, allowing students to learn and be tested on structural concepts better than through a 2D interface.

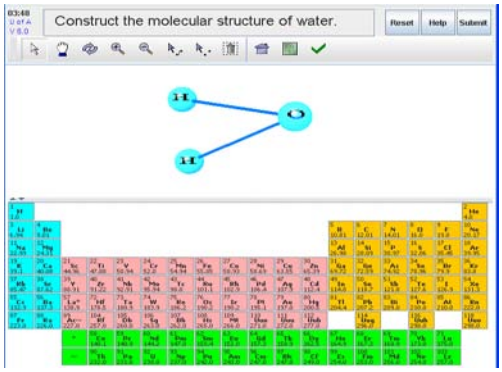

Figure 6: *An example question showing a 3D molecule used to test the atomic and molecular concepts in chemistry.*

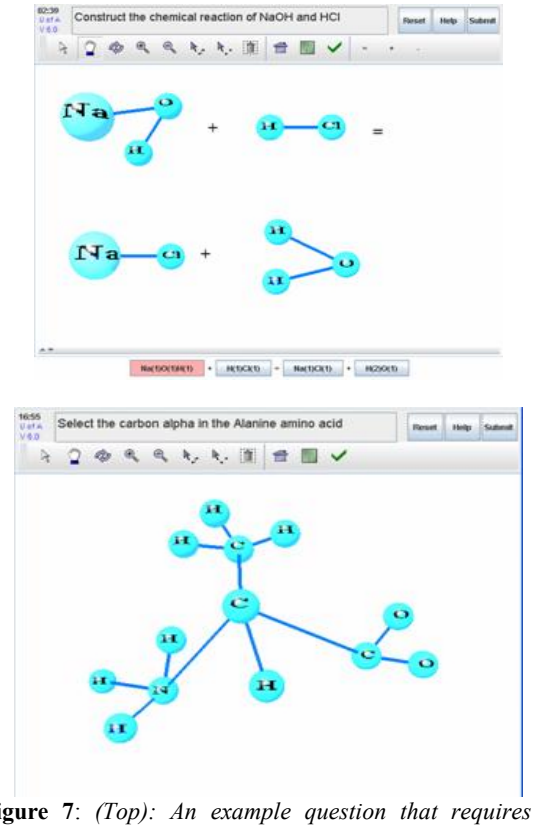

Figure 7: *(Top): An example question that requires a description of the molecular structural changes in a chemical reaction. (Bottom): An example question for testing amino acid structure in biology.*

To demonstrate how our interactive approach can be applied effectively to chemistry questions, we implemented an item type for periodic table related questions. Figure 6 shows an example question asking a student to construct the molecular structure of water (H_2O). Another example question is shown in Figure 7 (top), which asks the student to describe the molecular structural changes in a chemical reaction, *i.e.*, $NaOH + HCL \rightarrow NaCL + H_2O$.

3D molecular structures can also be used in the testing of biology and bioinformatics knowledge; for example, Figure 7 (bottom) shows an example that can be used to test the understanding of amino acid structure.

One main challenge in using 3D graphics is the complexity in assigning scores and estimating the difficulty level of items. In past research [CB06] we had discussed graph-based estimation of difficulty level for Chemistry Item Types. Here we will instead discuss a parameter-based strategy for estimating difficulty levels.

3.1 Parameters based Estimation of Difficulty Level for Math Item Types

The graph based strategy for chemistry discussed in our earlier research may not be applicable for other subjects. Parameter based strategy is a more general approach for assigning initial difficulties to items. We use Math questions as examples to illustrate the concept. Figure 8 shows an item requiring a student to distribute the numbers into four bins so that the sum in each bin is the same. We define parameters to control the generation of multiple questions, as well as the difficulty levels of the questions generated. For example, when solving the question "distribute the numbers so that the sum in each bin is equal" (Figure 8), the difficulty level of a question is defined by the function $f(n_{bkt}, n_{nbr})$, where n_{bkt} is the number of baskets used and n_{nbr} is the number of objects to distribute. The difficulty level increases as n_{bkt} or n_{nbr} increases. Additional difficulty can be introduced by using decimal instead of integer numbers. We verified the feasibility of our approach by conducting evaluation experiments.

3.1.1 Evaluation of the parameter based strategy

Methodology

We extended the concept of IRT and used 2-PLM (see Section 1) coupled with measurement of average time taken to solve problems to fit a linear regression model and examine the correlation between the difficulty levels generated by our strategy with the predefined difficulty levels. The calibration was done by seven students to rate the difficulty of each item based on the percentage of correct responses. 2PLM was used, since it was almost impossible to guess the correct answer for the given question format; the value of parameter *c* was close to zero. Mathematical details will not be discussed here for brevity. However, we will describe the design of the evaluation experiment and discuss results. The user interface of the evaluation program is shown in Figure 9. The questions used for evaluating the automatic difficulty estimation algorithm followed a course of increasing difficulty levels. Participants' familiarities with the questions were not taken into account during the assignment of maximum times, based on the fact that none of the participants had used these test formats before. A participant's answer, time needed and mark for each question was recorded. The mark for an answer was not based on a simple correct or incorrect criterion, and partial mark was awarded. For example, if a participant got the numbers in only one of baskets correct, whereas all together 4 baskets were present, (s)he could still get a mark of 0.25 (the full mark for an answer being 1.0).

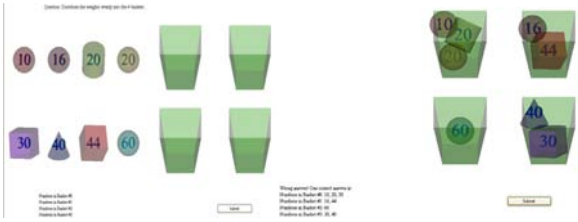

Figure 8: *(Left) an interactive math question, and (right) a student answer.*

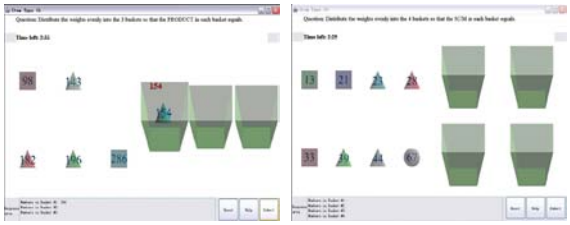

Figure 9: *Interface used in the evaluation experiment.*

Procedure

Seven participants were chosen, who were high-school students in Grade 10 to Grade 12 and understood basic arithmetic including factorization. Two sets of questions were given to the students:

1. Distribute the weights evenly into M baskets so that the SUM of the numbers in each of the basket is the same.

2. Distribute the weights evenly into M baskets so that the PRODUCT of the numbers in each of the basket is the same.

A procedure to solve a SUM question is:

(a) Add up the numbers and divide the sum by the number of baskets.

(b) Move the appropriate numbers into each basket based on the average computed in (a).

A similar procedure can be followed considering prime factors to solve a PRODUCT question.

Results and Analysis

Each participant's ability was considered as his or her total mark scaled in the range between [-3, 3]. Depending on the estimated abilities, each question's difficulty parameter *b* is calculated using IRT. Based on the experimental data (not shown here), the linear regression equation for estimating the difficulty of the SUM questions is:

$$b = -6.44 + 0.47 n_{bkt} + 2.77(n_{nbr}/n_{bkt}) - 0.74\ ID$$

Where *ID* varies between 1 and 6 depending on the calibrated difficulties. The correlation between the calibrated and experimental values was $R^2 = 0.95$.

The linear regression equation for estimating the difficulty of the PRODUCT questions (*ID* between 7 to 12) is:

$$b = -14.74 + 3.52\ n_{bkt} + 2.77(n_{nbr}/n_{bkt}) \ -1.08\ ID$$

with $R^2 = 0.99$. The high R^2 values (close to 1.0) indicate that the difficulty parameter *b* estimated by our algorithm has very high correlation with the *b* obtained from the calibrated values. Hence, the proposed parameter based strategy for estimating difficulty level is validated. Details on the experimental procedure and analysis of the data collected will be discussed in future work.

4. Testing Cognitive Intelligences in addition to Subject Knowledge

Each person possesses intelligence of one form or another, but intelligence can be discovered only in the correct context. For example, we cannot assess a student's social skill by watching him or her dissecting a frog. Therefore, test item types have to be designed according to the kind of intelligence to be assessed. Based on Gardner, the seven intelligences are skills to resolve problems and to create valuable contribution to society, entailing the potential for finding problems and acquisition of new knowledge [Gar83, Gar83a]. The seven intelligences are:

1. The ability to use words, orally or in writing, effectively (Linguistic).
2. The ability to use and analyze numbers effectively (Logical-Mathematical).
3. The ability to perceive the visual-spatial context and to respond correctly based on the perception (Spatial).
4. The ability to use one's body to express ideas and feelings, including using hands to manipulate or coordinate things (Bodily-Kinesthetic).
5. The ability to perceive, discriminate, compose, express, transform and invent musical forms (Musical).
6. The ability to observe, understand and distinguish moods, intentions, agendas and feelings of other people (Interpersonal).
7. The ability to acquire and be aware of self-knowledge and apply effectively on the basis of that knowledge.

Among these seven intelligences, only two can be expressed in test items based on traditional multiple-choice format (these being Verbal/Linguistic and Logical/Mathematical). We assess student cognitive skills by developing innovative test items, making use of video, audio, graphics, animation, etc. Some examples are discussed below.

4.1 Visual-Spatial Intelligence Item Type (IIT)

Item types for assessing a student's mathematical and logical skill are more commonly used in computer-based testing, and they can be presented using a multiple choice format, provided one only wants to assess the result. In contrast, visual-spatial skills cannot be tested using traditional pen-and-paper format because they need to be tested in a dynamic context. Such context can be simulated using computer-generated navigation. For example, the boxes in Figure 10 continue to move randomly at fast speed on the screen, while the student has to link, by drawing arrows between boxes (from corner to corner) so that the box content is in a particular order. In this example, the question is "to drive through these cities from north to south without revisiting any of the cities." An important consideration is to separate the assessment of visual-spatial skill from knowledge. A student may not know geography but have high visual-spatial skill. Therefore, the question has to be of minimum difficulty, *e.g*., order the numbers in an ascending sequence.

4.2 Linguistic IIT

There are many ways to test linguistic skills. An example is to ask a student to highlight a certain category of words, *e.g*., preposition (Figure 11 (a)), or to highlight a phrase having certain meaning. Vocabulary can be tested using 2D or 3D puzzles (Figure 11 (b) & (c)). Other examples can be drag-and-drop (drag the correct word from a list and drop it in the correct position in a paragraph), listen and dictate (the student types into a text box what (s)he hears from an audio clip) or rearranges words presented in random order into sentences. More detailed discussion on testing multiple intelligences can be found in [CB07].

4.3 Musical IIT

An example of our Musical Item Types (Figure 12 (a)) requires a student to look at video clips showing different dancing patterns. Figure 12 (b) shows a sequence of Korean, Swan Lake, Irish, Jazz and Ribbon dances. A student needs to associate each dance with the correct music, which is played by clicking on the "music" text. Note that no specific meaning is attached to the text to avoid providing any hint. The student has to transform the musical rhythm (s)he perceives to a sequence of artistic body movements. Musical notes' discrimination, or musical instrument and sound mapping can also be used in this item

type to test cognitive skills in music. An alternate format of the video is to use shadow-type dancing figures, like the jazz dancers (the 4th picture in Figure 12 (b)) to avoid disclosing the costume and thus culture as a hint to the music.

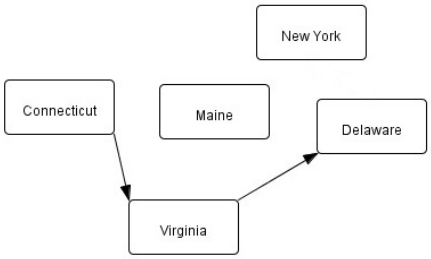

Figure 10: *An example of a Visual-Spatial IIT to test a student's ability to perceive visual-spatial context and respond correctly.*

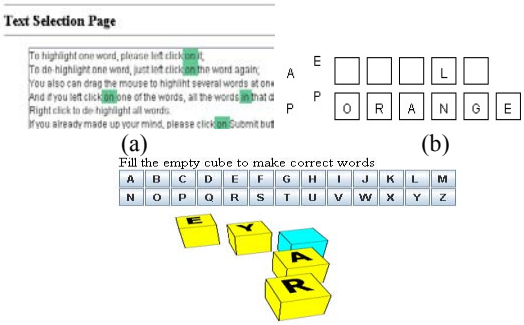

Figure 11: *Use Linguistic IITs to test a student's effectiveness in using words.*

4.3.1 User evaluation of musical IIT

We verified the feasibility of using musical IIT to estimate a student's musical skill. In the context of our evaluation, musical skills mean the general aptitude towards pitch difference, tempo, note duration and rhythm. The evaluation method can be extended to test other IITs.

Methodology

The evaluation contained three parts: a questionnaire to get a general impression of an observer's musical backgrounds, a Seashore [Sea19] based test used as the ground truth, and the item type being evaluated. The Seashore based test was given to observers at the same time as the questionnaire; using the questionnaire as a distracter task. The questionnaire included several basic questions on the observers' ages, school years, etc., and then asked the observers to describe in as much detail as possible their musical backgrounds including but not limited to any music lessons or classes, the type and quantity of music they listen to, and any other relevant material. The questionnaire also asked observers to rate their own "musicality" on a scale of 1-10. This questionnaire was used when analyzing results of experiments by providing some demographic and background data.

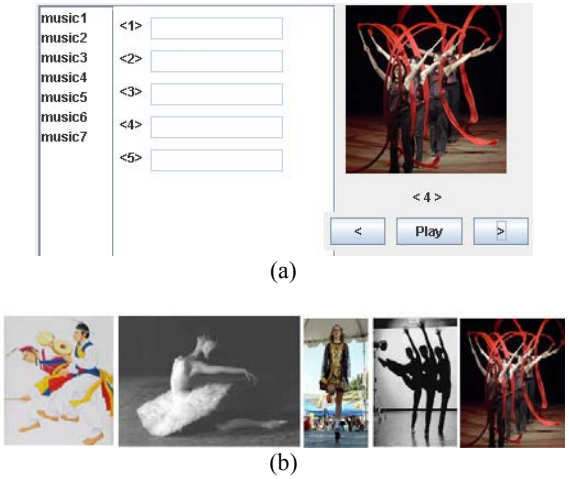

Figure 12: *(a) An example of a Musical IIT to test a student's ability to perceive, express and transform musical forms, and (b) A sequence of video expressing different musical composition.*

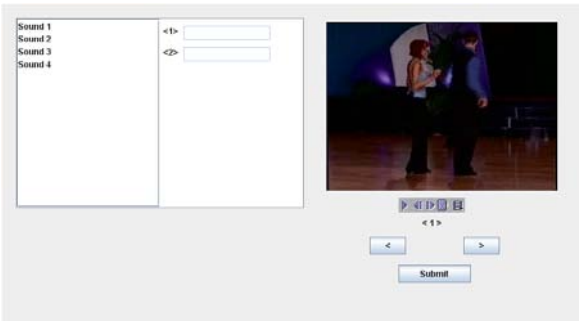

Figure 13: *Item Type used for evaluation of musical skills.*

The ground truth test was made up of paired sound clips, these clips were presented to the observers during the questionnaire with a break of 45 seconds between clips. The combination of the distracter task and the time gap prevented observers from actively rehearsing, and instead forced them to encode the sound into memory and then recall it to compare with the second clip. Observers with greater musical aptitude were capable of encoding the information into memory more accurately. A total of ten pairs covering pitch difference, tempo, note duration and rhythm were presented. The observers were asked to record whether or not the clips were the same, and if they differed, in what way they did so (e.g., shorter or longer, faster or slower, higher or lower, etc.). Once the item type and ground truth assessment test had been taken by observers, the results were analyzed graphically and using rank correlation.

Procedure

Experiments were performed by nine high school students. The item type itself (Figure 13) was presented to observers individually on a computer. An item consisted of a list of four sound clips and two video clips of people dancing (with the sound removed), which the user needed

to match with the sound clips based on the rhythm of the music and the dance. Two of the sound clips matched the videos, while the other two were extracted from unrelated videos. All clips had the same length (15 seconds). A user dragged the corresponding sound clips into the two answer boxes, one for each video, and then pressed the submit button to reach the next question. A total of ten questions, selected from a larger set of questions, were presented to a user, such that the calibrated difficulties of the ten questions formed a uniform distribution of difficulties.

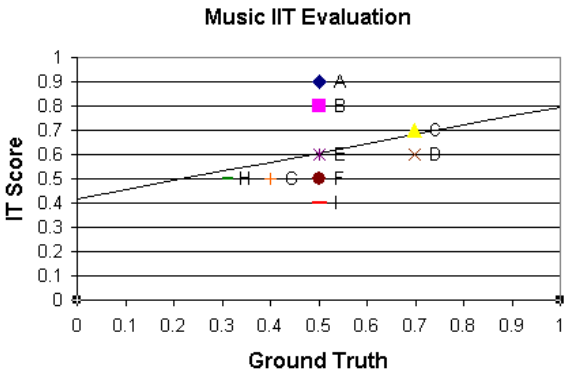

Figure 14: *A graph plotting ground truth vs. item type score.*

Observer	A	B	C	D	E	F	G	H	I
Ground truth	1	2	3	4	5	6	7	8	9
IT Score	3	4	1	2	5	6	8	9	7

Table 1: *Ranked Results.*

Results and analysis

Among the participants, one viewed himself as substantially below average when asked to judge his own musicality, four viewed themselves as average or just slightly above average and the remaining four viewed themselves as having exceptional musical aptitude. These views, did not however appear to have a direct correlation to a subject's performance in the experiment. This could likely be attributed to the Dunning-Kruger effect [DK99], which suggests that those with lower ability in a given area tend to over-estimate their level, due to a lack of skills necessary to properly assess their own ability.

Plotting the results on a graph comparing the ground truth and item type scores, we can see that there is a general trend from the lower left to upper right as would be expected (Figure 14). This trend shows that lower scores on the ground truth result in lower scores on the item type, and higher scores on one imply higher scores on the other as well. The visible appearance of a general trend, however, is not necessarily indicative of an underlying correlation, so we analyze the results further. Using the Kendall Rank Correlation Coefficient (τ) [Ken38] to calculate the correlation between the ground truth and item type result rankings (Table 1) we can see that the results from the two tests have a good level of agreement with $\tau = 0.67$. This shows that the trend visible in the graph does indeed illustrate an underlying correlation between the results of the two assessments. However, for further verification, in our future work we will collect records from musical teachers as ground truth and conduct experiments with more observers.

5. Student feedback on graphics item types

Feedback	Student1 Grade 11	Student2 Grade 11	Student3 Grade 11	Student4 Grade 7
Satisfaction				
Satisfied	31	29	25	24
Neutral	1	4	7	8
Dissatisfied	1	0	1	4
Preference				
Computer-based	30	21	23	26
Pen&paper	3	12	10	10
Time taken				
Less computer	9	14	14	Not
Less pen&paper	18	11	9	recorded
About the same	6	8	10	properly.

Table 2: *Summary of detailed evaluations by some students.*

We have received positive feedback from K-12 students groups visiting our research centre regarding the appeal of graphics item types to students. Extensive user studies with some students were conducted during August 2007. Some of the findings on four students are summarized in Table 2.

Note that the four students in Table 2 had somewhat different backgrounds. Three were in Grade 11 and one was in Grade 7. Among the Grade 11 students, Students 2 and 3 had taken computer-programming courses while Student 1 did not have any programming knowledge. It can be seen that these students in general were both satisfied with the graphics item types and also preferred computer based testing. However, there were some differences in the evaluations: (a) Students 2 and 3 had very similar evaluation and timing results since they were both from the same grade with good programming and user-interface knowledge, these skills may have given them an edge in performing the computer-based tests quite fast; (b) Student 1 though very interested in computer based graphics item types was relatively slower in working with the computer test interfaces and in most cases performed the pen-and-paper tests faster; (c) Student 4 though satisfied and interested in the graphics item types was unable to record precise time data properly. This may be a result of the slight immaturity of a Grade 7 student compared to Grade 11 students. In future evaluations with junior students, it is necessary to find appropriate means of accurately recording the time taken on pen-and-paper tests without involving a costly monitoring process.

6. Conclusion and future work

In this paper we outlined how graphics item types could be used for computer adaptive testing, and how different types of intelligence could be tested through this new approach. There are still various issues that need to be considered in

future research, including: How to precisely measure the effectiveness of graphics in adaptive testing? How to automatically grade graphics based responses to certain questions, *e.g.*, how to evaluate the accuracy of a sketched map? How to use graphics to effectively simulate laboratory tests? How to use graphics and haptics to create interesting testing environments for the visually impaired?

Acknowledgements

The support of NSERC, iCORE and Castle Rock Research in making this work possible is gratefully acknowledged, as is the help of the CROME team members in several implementations. Online demonstrations related to this work can be found under crome.cs.ualberta.ca.

References

[BC07] M. Bailey and S. Cunningham, "A Hands-on Environment for Teaching GPU Programming", *SIGCSE 2007 Conference Proceedings*, Kentucky, 254-258, 2007.

[BTB*00] S. Bonham, A. Titus, R. Beichner and L. Martin, "Education Research using Web-Based Assessment Systems," Journal of Research on Computing in Education, 2000.

[Cas06] CastleRock Research Corp. website: http://www.castlerockresearch.com/caa/WhatisCAA.aspx.

[CB06] I. Cheng and A. Basu, "Improving multimedia innovative item types for computer based testing," IEEE International Symposium on Multimedia, 8 pages, San Diego, USA, December 2006.

[CB07] I. Cheng and W.F. Bischof, "Multimedia item type design for assessing human cognitive skills", IEEE International Conference on Multimedia and Expo (ICME), 4 pages, Beijing, China, July 2007.

[Cun00] S. Cunningham, "Re-Inventing the Introductory Computer Graphics Course: Providing Tools for a Wider Audience," *Computers and Graphics*, April 2000.

[DK99] D. Dunning and J. Kruger, "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments", *Journal of Personal and Social Psychology*, Vol. 77, No. 6, pp. 1121-34, December 1999.

[ER00] S.E. Embretson and S.P. Reise, Item Response Theory for Psychologists, Lawrence Erlbaum Associates, 2000.

[Gar83] H. Gardner, Frames of Mind: The Theory of Multiple Intelligences, New York: Basic Books, New York, 1983.

[Gar83a] H. Gardner, "Artistic Intelligences", *Art Education*, Vol. 36, No. 2, Art and the Mind., pp. 47-49, March 1983.

[IB02] K. Ivers and A. Barron, "Multimedia Projects in Education: Designing, Producing and Assessing," 2nd Edition, Libraries Unlimited, 2002,

[Ken38] M. Kendall, "A New Measure of Rank Correlation", *Biometrika*, Vol. 30, pp. 81-89, 1938.

[KS96] L. Kjelldahl and Y. Sundblad, "Experience from 10 years of student projects oriented towards graphic interaction," Computers & Graphics, pp. 463-471, 1996.

[KS03] H. Kaufmann and D. Schmalstieg "Mathematics and geometry education with collaborative augmented reality," Computers and Graphics, vol. 27, 339-345, 2003.

[LH93] R.F. Lyvers and B.R. Horowitz, "A Unique Instructional Tool for Visualizing Equipotentials and its Use in an Introductory Fields Course," IEEE Transactions on Education, Vol. 36, No. 2, 237-240, May 1993.

[LH97] V. Linden, and R.K. Hambleton, *Handbook of Modern Item Response Theory*, London, Springer Verlag, 1997.

[LPL*05] J. Lu, Z. Pan, H. Lin, M. Zhang and J. Shi, "Virtual learning environment for medical education based on VRML and VTK," Computers and Graphics, vol. 29, 283-288, 2005.

[Nce06] National Center for Education Statistics, website: http://nces.ed.gov/nationsreportcard/studies/tbaproject.asp

[PDP00] C.G. Parshall, T. Davey and P.J. Pashley, "Innovative item types for computerized testing," in Computerized Adaptive Testing: Theory and Practice. W. van der Linden & C. Glas (Editors), Kluwer, pp. 129-148, 2005.

[Sch06] Schreyer Institute for Teaching Excellence, PennState, "Using Computers to Administer Tests," website: www.schreyerinstitute.psu.edu/Services/Assessment/Testing/computer.asp

[Sea19] C.E. Seashore, The Psychology of Musical Talent, Silver Burdett Company, Boston, 1919.

[Syv06] Syvum, "Computer-Adaptive Test for GMAT," website: http://www.syvum.com/gmat/cat.html

[Tax03] G. Taxén, "Teaching computer graphics constructively," SIGGRAPH, Educators Program, 1-4, 2003.

[THC94] L. Teles, L. Harasim and T. Calvert, "VIEW: An educational tool for the information highway," Ed-Media Conference, 1994.

[TM04] T. Tang and G. McCalla, "Utilizing artificial learners to help overcome the cold-start problem in a pedagogically-oriented paper recommendation system," AH 2004, pp. 245-254, 2004.

[TFGT99] J.F. Trindade, C. Fiolhais, V. Gil, J.C. Teixeira, "Virtual environment of water molecules for learning and teaching science," Proceedings of Computer Graphics and Visualization Education, 12-15, Coimbra, Portugal, 1999.

[TSB91] H. Tuckey, M. Selvaratnam, and J. Bradley "Identification and rectification of student difficulties concerning three-dimensional structures, rotation, and reflection," *Journal of Chemical Education*, 68(6), pp. 460-464, 1991.

[Wik03] WikEd, "Adaptive Assessment," Education Week on the Web 2003 Vol. 23, Technology's Answer to Testing, website: wik.ed.uiuc.edu/index.php/Adaptive_Assessments

[Y05] H.C. Yang, "A General Framework for Automatically Creating Games for Learning," Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05), 2005.

[ZS02] A.L. Zenisky and S.G. Sireci, "Technological innovations in large-scale testing," Applied Measurement in Education, 15(4), 337-362, 2002.