Revised EUROGRAPHICS 2008 *Education Paper* submitted as invited paper to Computer Graphics Forum

Interactive Graphics for Computer Adaptive Testing

Irene Cheng and Anup Basu

Department of Computing Science, University of Alberta, Canada
Contact: lin@cs.ualberta.ca, anup@cs.ualberta.ca; Website: crome.cs.ualberta.ca

Abstract

Interactive graphics are commonly used in games and have been shown to be successful in attracting the general audience. Instead of computer games, animations, cartoons, and videos being used only for entertainment, there is now an interest in using interactive 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 3D 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. Methods for estimating difficulty level of a mathematical item type using Item Response Theory (IRT) and a molecule construction item type using Graph Edit Distance are discussed. Evaluation of the graphics item types through extensive testing on some students is described. We also outline the application of using interactive graphics over cell phones. All of the graphics item types used in this paper 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

Instead of giving the same set of questions to test all students, computer-adaptive testing (CAT) focuses on individualized student modeling. CAT [Nce06, Sch06, Wik03] is an effective mechanism not only to reduce student stress, 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).

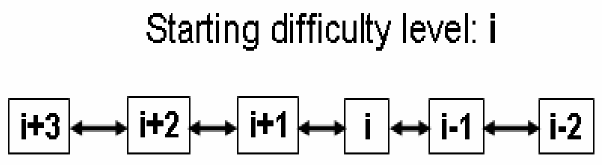


Figure 1: *A strategy for adaptive testing.*

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. In the conventional scoring system, students' skill levels are judged based on the percentage of correct answers for a set of exam questions. In the IRT approach, students' skill levels are evaluated based on the final skill level the performance graph converges to. For example, student *John* can be evaluated highly even with a comparatively lower percentage of correct answers (a decreasing curve) if he starts with an initial question at a high difficulty level; while student *Joe* starting with an easy initial question may obtain a higher percentage of correct answers (an increasing curve) but the convergence level can be lower than *John*.

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:

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

Where c_i is the guessing parameter denoting the probability of guessing correctly on an 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.

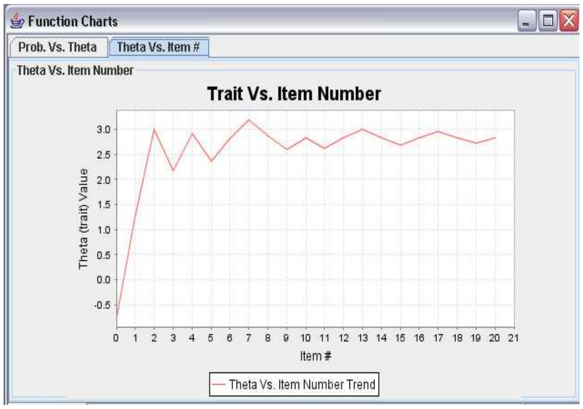


Figure 2: *A snapshot of our interface showing a student's performance based on IRT.*

In addition to questions being selected based on an individual student's ability level, CAT has other advantages over conventional pen-and-paper based testing. Some of these benefits include 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 of 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 authoring 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.

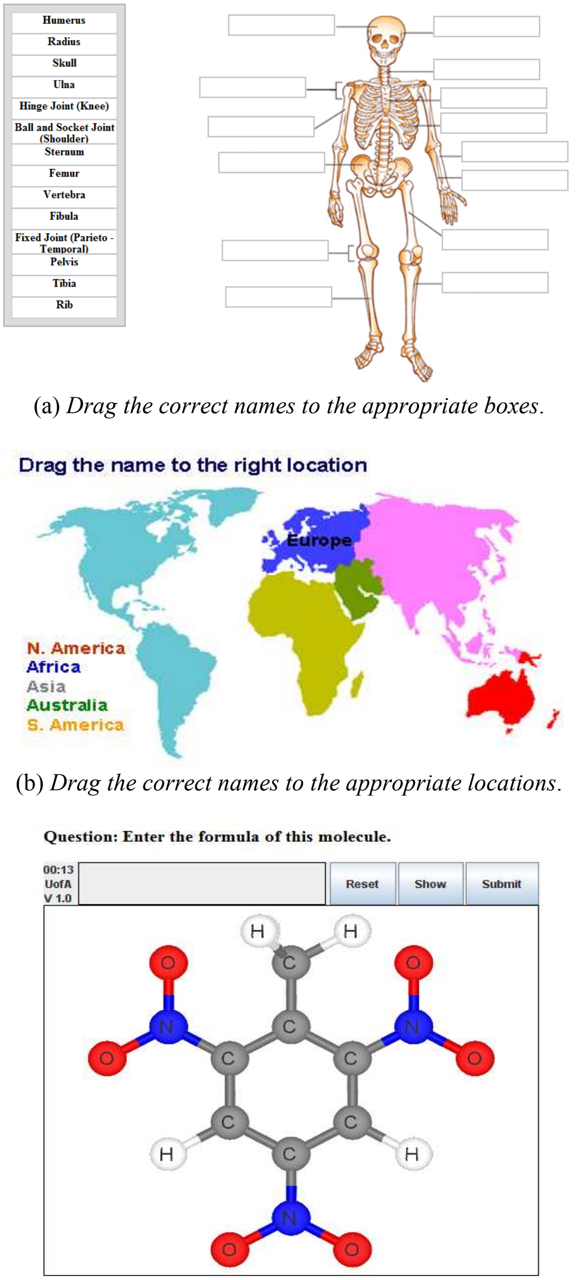
4. The use of educational games to engage students, inspire them to learn and make them feel rewarded.

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

Innovative graphics item types can be used to test a variety of curriculum subjects. For example, Figure 3 shows drag and drop examples for (a) a biology item, (b) a geography item, and a 3D animation display for (c) an organic compound. These items allow students to drag text or graphics to their appropriate locations on the screen, or visualize complex 3D structure of molecules.

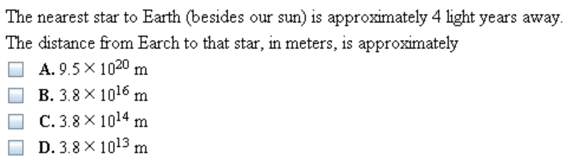


(a) *Drag the correct names to the appropriate boxes*.

(b) *Drag the correct names to the appropriate locations*.

(c)

Figure 3: *Examples of drag and drop question for (a) biology, (b) geography, and (c) a 3D animation display for an organic compound.*



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 conventional multiple choice question.*

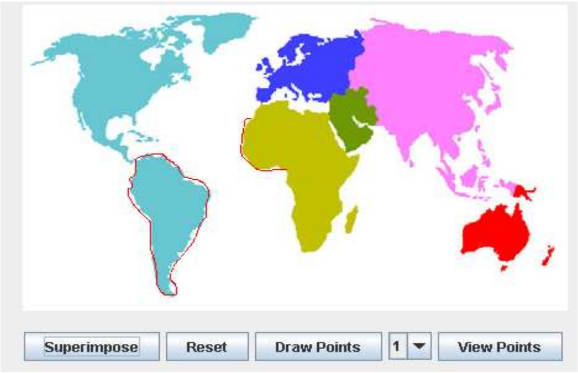


Figure 5: *Our item authoring interface provides user-friendly interaction with the items creator, who can simply define an answer region by outlining the boundary on the screen. The bounding coordinates are then generated automatically by the interface.*

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 has a unique layout and additional parameters are needed to create a question. For example, in the human body item (Figure 3 (a)) the application interface has to know whether or not the student drags and drops the text label to the correct screen location by comparing the coordinates of the answer box with the mouse press/release location.

It is a tedious job for the item creator to use an image editor for locating the (x, y) coordinates, and for defining the answer boxes. To facilitate this process, our authoring interface allows the item creator to draw bounding boxes at appropriate locations on the screen using a mouse. The interface then automatically detects and 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). We follow an object-oriented approach to ensure reusability and portability of each graphics item type. Unique layouts corresponding to different graphics item types are implemented as plug-ins in the authoring framework.

3. Automatic Item Difficulty Level Generation and Scoring

2D pictures can be used in pen-and-paper formats, but 3D interactive 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 constructed molecules can be rotated and manipulated in 3D, allowing students to learn and be tested on structural concepts better than through a 2D interface.

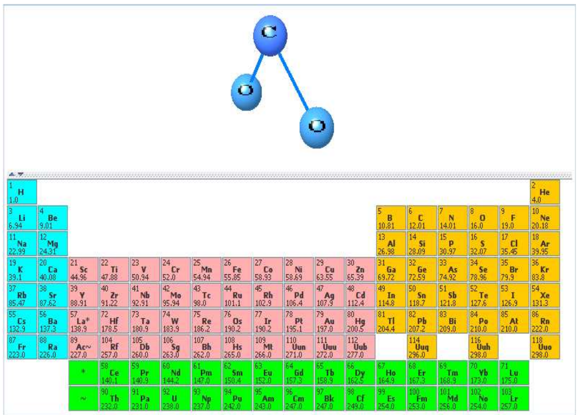
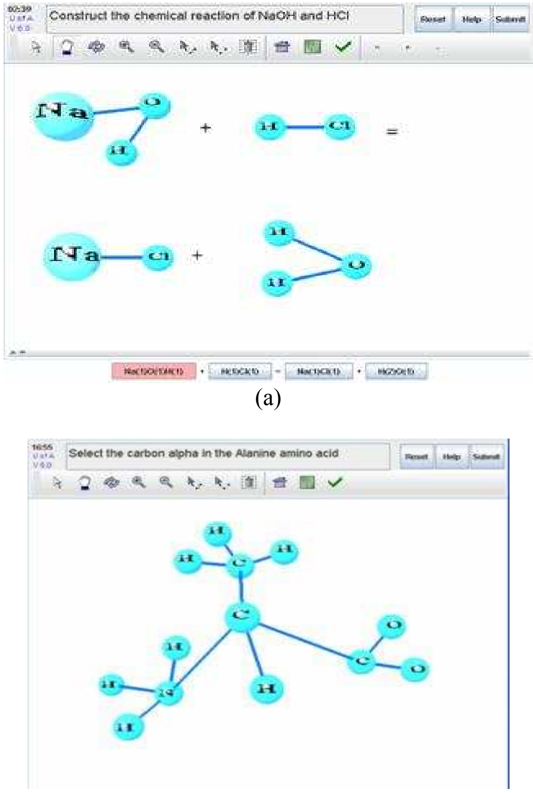


Figure 6: *An example question showing a 3D molecule used to test the atomic and molecular concepts in chemistry.*


02:39
U of A
V 6.0
Construct the chemical reaction of NaOH and HCl

(a)

16:55
U of A
V 6.0
Select the carbon alpha in the Alanine amino acid

(b)

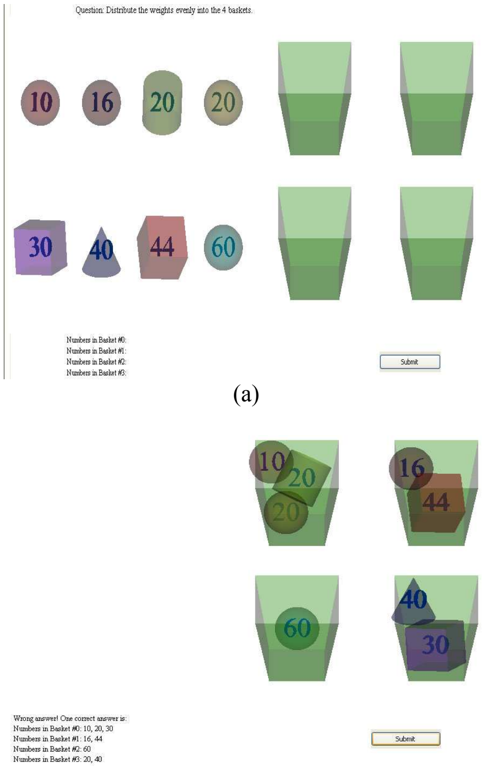
Figure 7: *(a) An example question that requires a description of the molecular structural changes in a chemical reaction. (b) An example question for testing amino acid structures in biology.*

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 carbon dioxide (CO_2). Another example question is shown in Figure 7 (a), 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 (b) shows an example that can be used to test the understanding of amino acid structures.

One main challenge in using 3D graphics to test molecule structure is the complexity in assigning scores and estimating the difficulty level of items. Next, we will discuss parameter-based and graph-based strategies for estimating difficulty levels.

3.1 Parameters based Estimation of Difficulty Level for Math Item Types


Question: Distribute the weights evenly into the 4 baskets.

Numbers in Basket #0:
Numbers in Basket #1:
Numbers in Basket #2:
Numbers in Basket #3:

(a)

Wrong answer! One correct answer is:
Numbers in Basket #0: 10, 20, 30
Numbers in Basket #1: 16, 44
Numbers in Basket #2: 60
Numbers in Basket #3: 20, 40

(b)

Figure 8: *(a) The student needs to distribute the numbers into four bins so that the sum in each bin is equal, and (b) a student's incorrect answer.*

A parameter based strategy is a general approach for assigning initial difficulties to new question items. We use Math questions as examples to illustrate this concept.

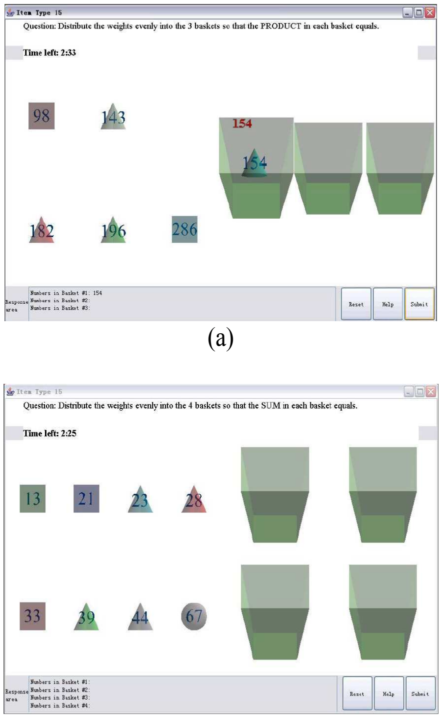
Figure 8 shows an item requiring a student to distribute the numbers into four bins so that the sum in each bin is the same. The parameters defined for this Math item type serve two purposes. They are used to control the generation of multiple questions as well as the difficulty levels of the questions generated. For example, when solving the question "distribute the numbers so that the sum in each bin is equal," 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 be distributed. The state before distribution is shown in Figure 8 (a) and a student's incorrect distribution is shown in Figure 8 (b). 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.

Evaluation of the parameter based strategy

We extended the concept of IRT and used 2-PLM (see Section 1) coupled with measurement of the 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 two baskets correct, whereas all together 4 baskets were present, (s)he could still get a mark of 0.5 (the full mark for a question being 1.0). This is illustrated by a student's answer for the Math question given in Figure 8 (b) where the top right and bottom left baskets contain the correct sum of 60.

Seven participants, who were high-school students in Grade 10 to Grade 12 and understood basic arithmetic including factorization, were chosen. 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 baskets is the same (Figure 9 (a)).
2. Distribute the weights evenly into M baskets so that the PRODUCT of the numbers in each of the baskets is the same (Figure 9 (b)).



Question: Distribute the weights evenly into the 3 baskets so that the PRODUCT in each basket equals.

Time left: 2:33

Response area
Numbers in Basket #1: 154
Numbers in Basket #2:
Numbers in Basket #3:

(a)

Question: Distribute the weights evenly into the 4 baskets so that the SUM in each basket equals.

Time left: 2:25

Response area
Numbers in Basket #1:
Numbers in Basket #2:
Numbers in Basket #3:
Numbers in Basket #4:

(b)

Figure 9: *Interfaces used in the evaluation experiment. The student has to distribute the numbers so that (a) the sum is equal in each bin, and (b) the product is equal in each bin.*

A procedure to solve a SUM question is:

(a) Add up all n_{nbr} numbers and divide the sum by the number of baskets n_{bkt} to compute the average *Av*.

(b) Select a subset of the numbers so that their sum equals *Av*.

(c) Move the subset of numbers into a basket.

(d) Repeat Step (b) and (c) for other baskets.

A similar procedure can be followed considering prime factors to solve a PRODUCT question. This requires breaking up a composite number into its prime components in order to derive the target product in each basket.

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:

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

where ID varies between 1 and 6 depending on the calibrated difficulties. The higher the question ID, the greater is the level of difficulty. 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 can be found in [CSB08].

3.2 Graph based Estimation of Difficulty Level for Chemistry Questions

	H	He	Li	Be	B	C	N	O	-
H		s1	s2	s3	s4	s5	s6	s7	d1
He			s8	s9	s10	s11	s12	s13	d2
Li				s14	s15	s16	s17	s18	d3
Be					s19	s20	s21	s22	d4
B						s23	s24	s25	d5
C							s26	s27	d6
N								s28	d7
O									d8
-									

Table 1: *A score matrix is used for computing the weighted edit distance between two graphs.*

In the multiple choice or true/false format, an answer can only be correct or wrong. There is no partial mark awarded. As a result, students who are completely ignorant about the question are not discriminated from those who are more competent but somehow have made a minor mistake when answering a question. In contrast, when multimedia content, such as 3D items are used in the CAT system, a student's performance can be evaluated more fairly by considering partial scores. For example, for the questions illustrated in Figures 6 or 7, if a student correctly selects two hydrogen atoms and one oxygen atom, but makes a mistake in the connection between hydrogen and oxygen atoms, partial marks should be given.

In order to evaluate the correctness of an answer and award partial marks, we interpret a molecular structure as a graph, where nodes are atoms and edges are bonds. In this way, we can assess the correctness of an answer by comparing the similarity between two graphs. A number of graph similarity matching algorithms can be found in the literature [BS98, WSKR01]. Among these algorithms, graph edit distance [MWH00] is commonly used. In this algorithm, a set of graph edit operations is defined. These edit operations include deletion, insertion, and substitution of a node or an edge. The edit distance of two graphs is defined as the length of the shortest sequence of edit operations required to transform one graph to the other. In our scoring scheme, we extend the edit distance to a weighted version. A scoring matrix is used to store the weights. Since the graphs are non-directed, the matrix is symmetric (implemented as a triangular matrix) in which each entry represents the weight of an edit operation. Table 1 shows the first eight atoms of the periodic table. In the scoring matrix, the entry s_i is the weight of a switching operation, and the entry d_i is the weight for inserting or deleting an atom from the graph. In general, each entry in the scoring matrix records a mistake. We compute the sum of a sequence of edit operations required to arrive at the correct graph. The weighted edit distance of two graphs is defined as the minimum sum of all the possible sequences:

$$D = \min\left\{\sum_{1}^{n} s_i + \sum_{1}^{m} d_j\right\}$$

that can transform one graph to the other. For example, the weighted edit distance of the two graphs in Figure 10 is s_7, because the correct graph (left) can be reached by switching the hydrogen atom and the oxygen atom in the student's graph (right).

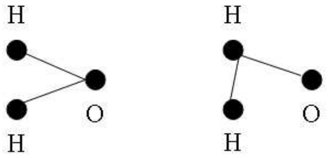


Figure 10: *Comparing a student answer (right) with the correct answer (left). The molecule to be constructed is water (H_2O).*

The general graph edit distance problem has exponential complexity in terms of the number of nodes in the graph [KH05]. However, the graphs used in our 3D model have the following special properties:

- The number of nodes is small. Normally, there are less than twenty atoms in a molecule.
- The degree of each node is small. Because of the limitation in chemical reactions, each atom can connect to no more than four other atoms.
- Very often, the graph follows a tree structure.

These properties reduce the complexity of the edit distance algorithms. In our model, if the number of atoms is less than five, we use an exhaustive search algorithm to calculate the edit distance between the two graphs, otherwise we use the seriation algorithm proposed in [KH05] to calculate the edit distance. For a graph, we define its empty distance D_0 as the edit distance between this graph and an empty graph, which has zero nodes and zero edges. The score of a student's answer is then defined as:

$$S = \max\left(\frac{D_0 - D}{D_0}, 0\right)$$

where D is the edit distance between the student's answer graph and the correct answer graph. If a student's answer is correct, the score is one. If a student does not answer the question, the score is then zero. Sometimes, the value $D_0 - D$ can be negative; for example, when the student's answer is a very complex graph, whereas the answer is a simple graph. In this case the max() function ensures that the student does not get a negative mark. If penalty marks apply, the max() function can be removed.

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 Howard 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 formats (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 2D and 3D 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 11 continue to move randomly at fast speed on the screen and occasionally overlap each other, 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 relating to subject knowledge, *e.g.*, order the numbers in an ascending sequence.

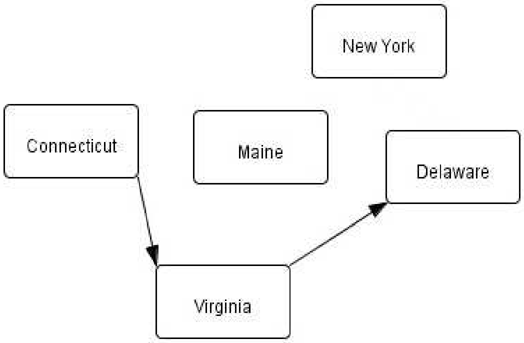
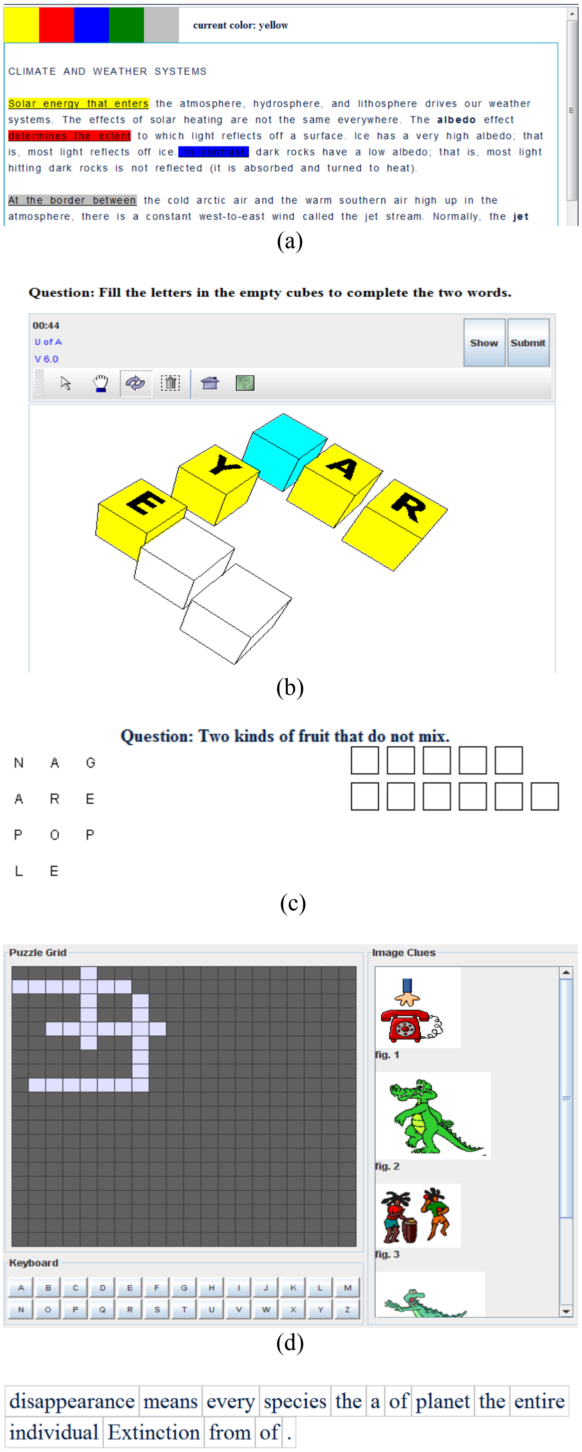


Figure 11: *An example of a Visual-Spatial IIT to test a student's ability to perceive visual-spatial context and respond promptly and correctly.*

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, or to highlight a phrase having certain meaning (Figure 12 (a)). Vocabulary can be tested using 3D or 2D puzzles (Figure 12 (b) and (c)). In addition to spelling, a student can be tested on the meaning or real life object associated with the word. The visual hint is given by a picture, e.g. a dog, or a graphical animation, e.g. running (Figure 12 (d)). Language grammar can be tested by asking the student to form a meaningful sentence by rearranging a set of shuffled words (Figure 12 (e)). Other examples can be drag-and-drop (drag the correct word from a list and drop it into the correct position in a paragraph), or listen and dictate (the student types into a text box what (s)he hears from an audio clip). More detailed discussion on testing multiple intelligences can be found in [CB07].



current color: yellow

CLIMATE AND WEATHER SYSTEMS

Solar energy that enters the atmosphere, hydrosphere, and lithosphere drives our weather systems. The effects of solar heating are not the same everywhere. The **albedo** effect determines the extent to which light reflects off a surface. Ice has a very high albedo; that is, most light reflects off ice. In contrast, dark rocks have a low albedo; that is, most light hitting dark rocks is not reflected (it is absorbed and turned to heat).

At the border between the cold arctic air and the warm southern air high up in the atmosphere, there is a constant west-to-east wind called the jet stream. Normally, the **jet**

(a)

Question: Fill the letters in the empty cubes to complete the two words.

(b)

Question: Two kinds of fruit that do not mix.

(c)

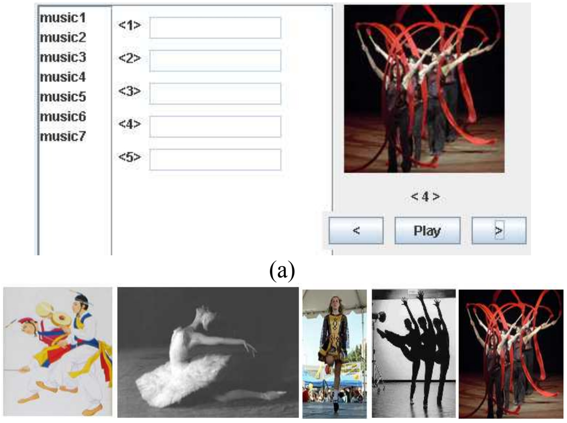
(d)

disappearance	means	every	species	the	a	of	planet	the	entire
individual	Extinction	from	of	.					

(e)

Figure 12: *Using Linguistic IITs to test a student's effectiveness in using words and the understanding of the words. (a) Highlighting a word or a phrase, (b) forming words in a 3D grid, (c) spelling words using the drag and drop operation, (d) spelling the word corresponding to the give image, and (e) rearranging the shuffled words into a grammatically correct sentence.*

4.3 Musical IIT



(a)

(b)

Figure 13: *(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.*

An example of our Musical Item Types (Figure 13 (a)) requires a student to watch video clips showing different dancing patterns. Figure 13 (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 box on the left of the interface. 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 13 (b)) to avoid disclosing the costume and thus culture as a hint to the music. A student can also be tested on which category of music (s)he can discriminate better, or whether (s)he can discriminate musical pieces oriented from different cultures.

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.

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.

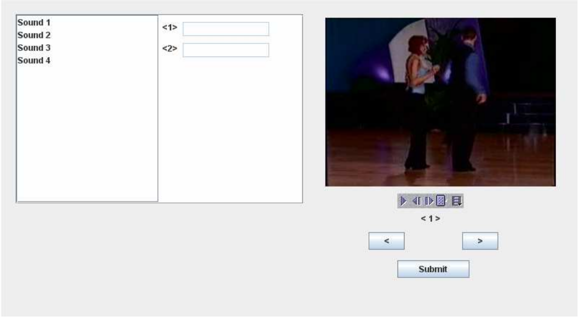


Figure 14: *Item Type used for evaluation of musical skills.*

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.

Experiments were performed by nine high school students. The item type itself (Figure 14) 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.

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.

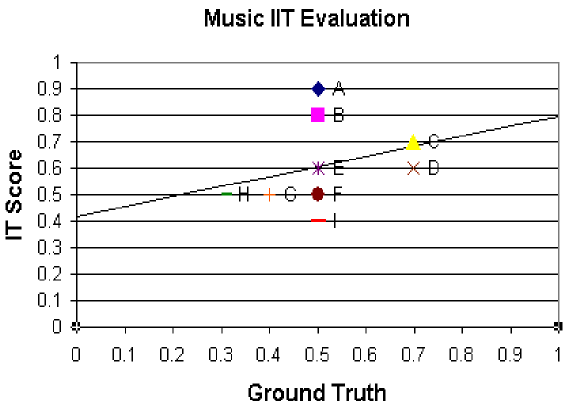


Figure 15: *A graph, plotting ground truth vs. item type score, show a correlation trend.*

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 2: *Ranked Results.*

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 15). 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 2) 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. Note that our evaluation methods make use of computer-based interactive graphics to test an individual's musical skill. Associating graphical visual contents with acoustic effects are more appealing and engaging than the traditional testing method requesting the examinees to discriminate musical beats and rhythms purely based on listening. Since the musical item types can be web-based, it is possible to conduct tests remotely which saves the examiners' traveling and administration time.

5. Student feedback on graphics item types

Feedback on 33 items	Student1 Grade 11	Student2 Grade 11	Student3 Grade 11	Student4 Grade 7
Satisfaction with computer-based items				
Satisfied	94%	88%	76%	74%
Neutral	3%	12%	21%	9%
Dissatisfied	3%	0%	3%	17%
Preference				
Computer-based	91%	64%	70%	76%
Pen-and-paper	9%	36%	30%	24%
Completed items faster when using:				
Computer	27%	43%	43%	Not
Pen-and-paper	55%	33%	27%	recorded
About the same	18%	24%	30%	properly.

Table 3: *User evaluations on 33 interactive graphics items show that in general computer-based items are preferred over pen-and-paper items.*

Overall, 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 3.

Note that the four students in Table 3 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 in general these students 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.

There are a few important attributes that can influence a student's satisfaction towards using computer graphics items; not only the knowledge in solving the question items, but also his or her computer interactive skill as well as the appealing factors associated with the items, are essential. The observation gained from Students 2 and 3 completing the items faster when using a computer, and Student 1 being faster on pen-and-paper, suggests that a training period or a gradual migration is necessary before full scale computer graphics based adaptive testing is launched in high schools. The 17% dissatisfaction from Student 4 can be due to the designs and difficulties of the items which target Grades 10-12 and not Grade 7 students. We confirmed from these user evaluations that different graphics presentations and designs are important in order to engage users at different levels. Also, in future evaluations with junior students, it is necessary to find appropriate means for accurately recording the time taken on pen-and-paper tests without involving a costly monitoring process. It is also important to note that although using computer graphics items may not shorten the testing time, it bridges the geographic divide. Furthermore, it engages the students in education as reflected by the satisfaction rates in Table 3.

6. Mobile item types

As mentioned in the introduction, our goal is to extend the use of graphics item types beyond wired computers. The challenges of implementing a similar item type framework on wireless or mobile devices, such as cell phones, include limited display area, bandwidth, network coverage, device processing power and battery duration. Our strategy is to make available mobile item shells, which focus on intelligent brain activity rather than extensive computer interaction. An example of such a mobile item is shown in Figure 16, which requires the student to fill in the blanks so that the sum in each row and column agree. Since CAT requires spontaneous communication between the server and client, mobile items are more effective for learning rather than testing. Our goal is to use mobile items to make education more accessible in terms of time and location, in preparation for testing which is conducted on networked computers.

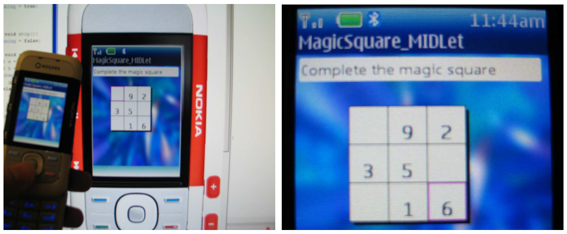


Figure 16: *(Left) is a cell phone displaying a mobile item example, and a computer screen at the background displaying the simulator software. (Right) is a zoom-in view of the cell display.*

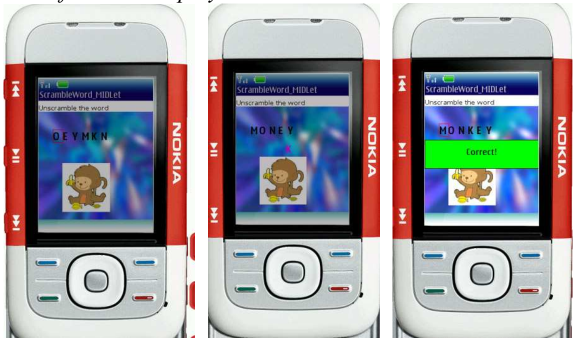


Figure 17: *(Left) another item type "word scramble" on a cell. (Middle and Right) process of selection of alphabets to form the correct word.*

Another mobile item type "word scramble" is shown in Figure 17. In this item type, the goal is to de-scramble the letters using the picture as a hint. The skill level of a user can be judged not only by recognizing the word from the picture, but also in taking the smallest number of steps in re-arranging the letters. Item types like this can be useful for learning new languages and improving vocabulary, as well as stimulate thinking.

7. 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, beyond subject knowledge, 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? By supporting graphics item types in computer-based adaptive testing, our goal is not only to test students' subject knowledge and intelligence; more importantly, we intend to provide engaging and rewarding educational incentives, which are not available in the conventional multiple choice context, so that students are inspired not only to learn online with a computer, but also to explore supplementary offline learning opportunities through their own initiatives.

Acknowledgements

The support of Natural Science and Engineering Research Council (NSERC) Canada and Informatics Circle of Research Excellence (iCORE) Alberta, Canada 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.

[BS98] H. Bunke and K. Shearer, "A graph distance metric based on the maximal common subgraph", *Pattern Recognition Letters*, 19, pp. 255-259, 1998.

[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.

[CSB08] I. Cheng, R. Shen and A. Basu, "An Algorithm for Automatic Difficulty Level Estimation of Multimedia Mathematical Test Items," International Conference on Advanced Learning Technologies (ICALT), 5 pages, Jul 2008.

[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.

[KH05] A. R. Kelly and E. Hancock, "Graph edit distance from spectral seriation", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 27, pp. 365-378, 2005.

[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.

[MWH00] R. Myers, R. Wilson, and E. Hancock, "Bayesian graph edit distance", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22, pp. 628-635, 2000.

[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

[WSKR01] W. Wallis, P. Shoubridge, M. Kraetzl and D. Ray, "Graph distances using graph union", *Pattern Recognition Letters*, 22, pp. 701-704, 2001.

[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.