Evaluating Performance Features with 3D Item Types for use with Computer-Based Tests in Education

Irene Cheng, Mark Gierl and Anup Basu
Departments of Computing Science and Educational Psychology
University of Alberta, Edmonton, AB T6G 2E8
Email: lin@cs.ualberta.ca, mark.gierl@ualberta.ca, anup@cs.ualberta.ca

Keywords: Computer based testing, 2 Alternative Forced Choice

Abstract

In this work we describe our preliminary research on designing 3D Multimedia Visual Discrimination tests based on an innovative item type called the 2 Alternative Forced Choice (2AFC). We evaluate how human discrimination capability varies with the representation of 3D objects in terms of the structure (mesh), surface colors (texture), size, distance, and complexity of objects. Performance results are summarized. Some implications for educational testing are also discussed.

I. INTRODUCTION

Testing is pervasive throughout North America as almost all states and provinces in the United States and Canada have educational testing programs. Increasing, these tests are being delivered using computers. For example, in the Canadian province of Alberta, mathematics achievement tests will be administered with computers using the internet in Grades 3 to 12 beginning in the Fall of 2006. Starting in 2007, science, social studies, and language arts achievement tests will also be administered with computers. One important advantage of using the computer to administer a test is the inclusion of innovative item formats [14]. An innovative item format makes use of the diverse capabilities in a computer to create new types of test items. For example, a conventional multiple-choice test item requires the student to select one option from a list of 4 or 5 alternatives. An innovative computer-based item, on the other hand, could require the student to highlight text, click on a graphic, move objects, or re-order statements, lists, and graphics on the screen. In fact, the array of innovative item formats that are possible seem limitless when the features of a computer such as graphics, audio, video, and animation are considered. Surprisingly, very little research exists on the use of innovative item formats in education [17]. In particular, researchers know little about how different features in these innovative computer-based test items affect student performance. Hence, the purpose of our study is to evaluate some key features with one promising innovative item format—the 2 Alternative Forced Choice (2AFC) item type—to better understand how this item can be used on educational tests. The alternative forced choice item, also called the multiple true-false item format, is used commonly in many testing situations [5].

Our paper is organized as follows: Section II provides a description of the 2AFC item type and the factors that may affect student performance on this task; Section III contains an overview of the experimental setting and relates scale-space simplification and texture resolution to visual discrimination. Section IV includes the results from the study; finally, in Section V, conclusions and implications for educational practices are presented.

II. THE 2AFC ITEM TYPE: COGNITIVE & AFFECTIVE FACTORS THAT INFLUENCE STUDENT PERFORMANCE

In this paper, we propose an innovative item type for computer-based testing using the 2AFC item. This item type includes two representations of a 3D object side by side, and a student is asked to choose one of the two objects. This type of task could be used in many different content areas. For the purpose of this study, our focus is only on visual discrimination. In psychophysical research this approach is also known as the two alternatives forced choice (2AFC) method and the result is analyzed using a psychometric curve (sigmoid curve) [1, 2-39].

Based on the experimental results, we evaluate the visual discrimination capabilities of individual observers, and generate statistics for the average population. We then relate visual discrimination capability to learning pattern. Geometry and texture are two basic components of a 3D object. Visual discrimination of the geometry and texture may be affected by various environmental and psycho-visual factors [9,10,11,13].

1 This work is supported in part by the Alberta Informatics Circle of Research Excellence (ICORE), Castle Rock Research Corporation, the Canada Research Chair Foundation and NSERC, Canada.
At least ten different known factors in the 2AFC item type could affect student performance. These factors include:

- **Visual threshold** [9,10] — Vanishing point at which a stimulus becomes just visible or invisible.

- **Luminance of the background in contrast with the stimulus** [7,10] — If the background is plain, the visual threshold will change depending on the luminance of the background in contrast to the object. This relationship is described by Weber’s Law, which states that the threshold is proportional to the background luminance.

- **Spatial visual masking** [10] — A large change of luminance may be present across an edge. Such a change reduces the ability of the eye to detect distortions spatially adjacent to the changes. The detection threshold (or the contrast corresponding to the threshold of perception) of a stimulus varies inversely as a function of its distance from the edge. As a target moves across an edge, the threshold on the light side of the edge is increased.

- **Temporal visual masking** — Post-masking (backward masking) occurs if the perception of a stimulus is affected by a subsequent strong signal. A visual backward masking model [2] suggests that, when two successive stimuli are presented within 0 to 200ms intervals, the recognition of the first stimulus (the target) can be impaired by the second (the mask). Pre-masking occurs if perception is affected by a strong signal beforehand.

- **Spatial frequency (cycles per degree of visual arc)** — Minimum detectable difference (sensitivity threshold) in luminance between a test spot and a uniform visual field, increased linearly with background luminance at daylight levels. However, virtual scenes and stimuli are not uniform, and contain complex frequency content. Outside a small frequency range, the sensitivity threshold drops off significantly. This phenomenon has led to the study of the perceptibility of contrast gratings (sinusoidal patterns that alternate between two extreme luminance values) and the following terminology:
  
  - **Threshold contrast at a given spatial frequency** is the minimum contrast that can be perceived in a grating.
  - **Contrast sensitivity** is the reciprocal of threshold contrast.
  - **Contrast sensitivity function (CSF)** plots contrast sensitivity against spatial frequency. It describes the range of perceptible contrast gratings [9].

- **Texture masking** — It can be observed that the noise added to a low frequency (or plain) background is much more visible than that added to a high frequency (or flowery) background. A model was developed in computer graphics [6], which allows the choice of texture pattern to hide the effects of faceting, banding, aliasing, noise, and other visual artifacts. This technique is suitable for applications using synthetic texture; however, for applications displaying intrinsic photo-texture of the object, choice of a different pattern is not an option.

- **Short-term memory** — The influence of a strong stimulus will last for a short time, thereby imposing a smoothing effect on the distortion measure. It is worth noting that people are more likely to remember bad quality than good quality.

- **Visual acuity** — The fovea, occupying roughly the central 1° of vision, is the region of highest sensitivity. Visual acuity, measured as the highest perceptible spatial frequency, is lower in the periphery than at the fovea.

- **Visual depth** — A sensation of reality occurs because of visual depth perception. Depth sensitivity — the ratio of viewing distance to depth discrimination threshold — is directly proportional to the viewing distance. Sharp edge, clear texture, shade, and surface gloss strengthen the sensation of depth.

- **Prior knowledge or Expectation** — Prior knowledge imposes on the viewer an expected geometry and texture representation. Deviation from expectation degrades visual fidelity. If a stimulus appears in an orientation different from what the viewer is familiar with, the viewer’s ability to discriminate decreases.

The purpose of evaluating these factors is to better understand how students respond to the 2AFC item type under different experimental conditions. These results will be used, in turn, to monitor, manipulate, and study the psychometric characteristics (e.g., item difficulty level) of this innovative item type when assessing a student’s visual discrimination skills.

### A. Two Alternatives Forced Choice (2AFC)

In the two-alternative-forced-choice (2AFC) procedure, a subject is presented with two stimuli, \( A \) and \( B \) (which represent, for example, \( x, x+\Delta x \)). The stimuli may occur in successive intervals, or they may occur in adjacent locations [1, 2-39]. For the purpose of this paper, we take the latter case, where \( A \) and \( B \) are placed in adjacent locations. The subject’s (judge’s) task is to state whether the target occurred in the left or the right location. In the experiments conducted,
a target is also displayed above the two stimuli so that the subject has a reference of the original object. The subject’s decision is recorded, as either correct or incorrect, for each pair of stimuli. Once the judging is completed, the results are summarized. To find the visual threshold, the percentage of correct judgments is plotted as a function of $\Delta x$ (Fig. 1).

Fig. 1: Percentage of correct judgments plotted as a function of $\Delta x$.

To avoid response bias, sufficient evaluations, e.g., 30, should be collected for each $\Delta x$ value. The line of best fit is obtained by regression, and the threshold can be located at the 75% correct performance [15]. When the two stimuli are clearly distinguishable, the score is 100%. If the difference is not apparent to the judge, he/she is forced to guess, and the possibility of picking the correct stimulus is 50%, after a sufficient number of evaluations have been performed. Discrimination capability of individuals can then be compared with the average value corresponding to the 75% correct judgment.

III. VISUAL DISCRIMINATION WITH SCALE-SPACE SIMPLIFICATION AND TEXTURE RESOLUTION

Luminance is a crucial factor in perceptual experiments. Taking the worst-case scenario, we assume visual sensitivity (ability to discriminate) is the highest when comparing stimuli. In order to enforce this assumption, the experimental environment is set to (1) normal daylight, (2) a background color which has a reasonable contrast with the stimulus’s color, and (3) a small spatial frequency just before the Contrast Sensitivity Function falls off sharply [9], imposing a condition for high brightness adaptation and visual sensitivity.

Classifying stimuli into categories, such as animal and scenery, can divide a perceptual problem into sub-problems; however, presenting similar geometry and texture in consecutive tests can easily confuse judgment. The fact that a strong stimulus lasts for a short time will affect the next discrimination process. Stimuli should be randomly selected from different categories, and displayed in a random order.

A. Scale-Space Simplification and Discrimination on Scales

We use Scale-Space Filtering (SSF) to extract 3D features [3]. Traversal between the different scales is achieved by varying the standard deviation parameter $\sigma$; the higher the value of $\sigma$ the more is the smoothing [8,16]. SSF is based on locating the zero-crossings of a signal at multiple scales. Zero-crossings are used to detect the degree of persistence of a structure (feature) on a 3D surface. Minor structures tend to diminish as $\sigma$ increases, and only major structures survive at higher scales (Fig. 2).

Fig. 2: Increasing scale $S_i$ from top to bottom. $S_0$ is the original signal extracted near the bottom of the Nutcracker model. Note that the local variation (fine detail) in the original signal is gradually removed and the scaled signal becomes smoother.
Table 1: Perceptual values of the nutcracker mesh between adjacent scales.

<table>
<thead>
<tr>
<th>Scale $S_{i-1} - S_i$</th>
<th># of faces</th>
<th>Perceptual value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>1162</td>
<td>0.0410</td>
</tr>
<tr>
<td>1-2</td>
<td>1118</td>
<td>0.0412</td>
</tr>
<tr>
<td>2-3</td>
<td>1074</td>
<td>0.0478</td>
</tr>
<tr>
<td>3-4</td>
<td>1040</td>
<td>0.0468</td>
</tr>
<tr>
<td>4-5</td>
<td>1002</td>
<td>0.0678</td>
</tr>
</tbody>
</table>

Table 2: An example of cumulated perceptual values over varying scales.

<table>
<thead>
<tr>
<th>From scale</th>
<th>To scale</th>
<th>Perceptual value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0.0410</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0.0616</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0.0677</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>0.0759</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>0.1080</td>
</tr>
</tbody>
</table>

When a 3D object moves closer to the viewer in a virtual scene, the mesh needs to be refined only if the resulting mesh improves visual quality. To determine how mesh refinement relates to visual quality requires measuring perceptual impact on the HVS. Adding or deleting a vertex or surface structure from a mesh generates a stimulus to human vision. We used a perceptual value to represent the change in the shortest distance measured from the skeleton of the 3D object, when moving from one scale to another. The perceptual values correspond to the perceptual impacts caused by changes on a 3D surface [4]. Table 1 is an example of perceptual values between adjacent scales. Table 2 shows an example of cumulated perceptual values across scales.

In this set of experiments, the original scale $S_0$ of a nutcracker toy was displayed on top, and two simplified scales (stimuli) were randomly chosen and displayed side-by-side at the bottom. The judge had to decide whether the left or the right stimulus was a finer version with better quality. If a wrong judgment was given, two simplified scales with a larger perceptual value were selected and displayed. If a correct judgment was given, two simplified scales with a smaller perceptual value were selected, and the process was repeated. The total number of correct judgments of each judge was recorded out of the 360 trials together with the time spend on the experiment.

B. Texture Resolution and Visual Discrimination

Cubes were used as stimuli in the second set of experiments. Fig. 3 shows an example of our Java3D interface with two stimuli mapped with texture of different qualities. The stimuli rotated slowly so that the judge could compare all views of the cubes. We ensured that the rotating speed was not too fast to lower the discrimination capability of the judge. Five JPEG images (Fig. 4) of $256^2$ pixels ($Res_{256}$) each were scaled down to lower resolution images by reducing $2i$ pixels linearly in the x and y directions, so that the resulting image set was $S = \{Res_{2^i} | 1 \leq i \leq 128\}$. Each lower resolution was generated by scaling down $Res_0$ in order to preserve a high image quality when generating the set $S$. We used pixel mixing for scaling because it tends to replicate what the HVS sees as the image moves closer to or further away from the viewpoint [12]. While the discrete sampling algorithm preserves the same set of colors, it generates less smooth images when scaling down. The linear scaling algorithm is faster but can lead to lower image quality.

![Fig. 3: Each cube face is 160^2 pixels. (Left) Texture of 144^2 pixels interpolated to 160^2 pixels. (Right) Original texture resolution of 160^2 pixels](image)

![Fig. 4: Five images of 256^2 pixels resolutions were used in the first set of experiments.](image)

![Fig. 5: Response curve of an observer using ResL = 90% x ResO.](image)
Six test resolution levels ($n^2 = 256^2, 192^2, 160^2, 128^2, 96^2$ and $64^2$) were chosen. In each loop, the six test levels were shuffled and presented to the observers in random order. At each test level, $Res_0$ and $Res_2$ were displayed side by side, and a randomly selected image was used. The observer had to determine which of the two stimuli had a better quality — $Res_0$ or $Res_2$. The initial value of $2i$ was set at 90% of $Res_0$. The value was then raised or reduced depending on whether the judgment was correct or wrong. There were a total of 60 loops resulting in a total of 360 responses (Fig. 5).

IV. STUDENT EVALUATIONS AND ANALYSIS OF RESULTS

In addition to the scores and time spent recorded in each set of experiments, we also computed the average perceptual threshold (75% correct judgment) or Just-Noticeable-Difference (JND) in the Experiment set A. By comparing with the JND, the visual discrimination capability of an individual student can be studied. Visual discrimination capability is an important factor when considering whether computer-based testing is an efficient studying tool for an individual. If a student’s capability is too far below the JND, computer-based testing may cause distraction or even frustration. In Experiment set B, we found that the HVS responds differently to different texture resolutions. The viewer can benefit from better visual quality if the correct resolution is displayed.

A. Visual Discrimination Experiment in Scale-Space

Fig. 6: An example of different scales of the nutcracker object, $S_0$, $S_3$, and $S_8$ from left to right.

We computed the JND by testing pairs of simplified meshes randomly selected from $S_0$ to $S_{30}$ of the nutcracker object (Fig. 6). The original mesh $S_0$ was displayed as a reference in the upper part of the interface. Two stimuli were displayed side by side in the bottom part. We followed the 2AFC with reference strategy, and a judge was asked to decide which one (left or right) was a finer version closer to the original. The perceptual values in the Look-up-table (LUT) explained in Section III were grouped into 10 sub-ranges. 361 tests were conducted by twenty judges on three monitors of different dimension and resolution, and the percentage of correct judgment in each sub-range was recorded. A JND of 0.096 (Fig. 1) was obtained by locating the 75% correct judgment.

Coarser or more simplified scales are used at a farther distance in order to preserve the perceptual quality. However, the evaluation results show that scale does not relate to distance linearly. When the perceptual value generated by changing scale is very small and below the JND, adjacent scales appear similar without significant impact on visual quality. Based on the evaluation results, we establish a step-like function for the nutcracker top object (Fig. 7). There is a JND below which the discrimination capability is low for most students. Item types associated with graphics, animations and images should avoid generating stimuli below the JND. Also, adjustments should be made for those students with discrimination capability lower than the average.

B. Visual Discrimination Experiment in Texture Resolution

The response curve at resolution level $160^2$ was steeper compared to the others. A similar trend appeared when using $Res_L = 80\% \times Res_0$ (Fig. 5). To verify the consistency of this pattern, we collected the responses from 20 observers (Fig. 8). We used different starting values in the range from 60% to 90% for $Res_L$. When $Res_L$ was too low, some test levels might not be able to achieve equilibrium in 60 loops. If the curve for an observer did not converge in the last 20 loops, the data was excluded from the results. At each resolution level, the average of the last 20 loops was computed. Fig. 8 shows the average taken from each of the 20 observers and the overall average at each resolution level. It was found that at level $160^2$, the visual discrimination ability was weaker for all
observers, compared with that at other resolution levels. When using computer-based testing, it is important to use the correct resolution in images, graphics and animations so that the students can benefit from the quality without wasting the network resources.

When using computer-based testing, it is important to use the correct resolution in images, graphics and animations so that the students can benefit from the quality without wasting the network resources.

Fig. 8: Responses collected from 20 observers show that at test level 160, visual discrimination on the scaled images is relatively lower compared to at other test levels.

C. Analysis of Scores and Time Spent

By analyzing the scores and time spent on this computer-based test using 2AFC, we discovered learning patterns associated with individuals. This conclusion is drawn from the following two case studies. We deliberately relaxed the time constraint so that user fatigue and analytic skill are taken into consideration.

1) Study using different texture resolutions (Experiment in Section III B)

The scores and time spent by six Computing Science Department graduate students at the University of Alberta are given in Fig. 9.

In order to examine the stimuli carefully and make the correct choice, the minimum time required is about 15 minutes. Students spending less than the minimum required time did worse than the others. Also note that after passing the minimum required time, performance does not improve proportionally to the time spent. We found that the student who got the highest score in fact did pretty well in class. The top three scores are associated with a longer break time in between each 60 trials, indicating that fatigue can lower the efficiency.

2) Study using different texture qualities (Experiment in Section III B)

Tasks requiring analysis need more time (effort) in order to obtain a higher score.

This study was aimed at finding out whether the test score is related to his/her analytic skill. Different from Study 1 where the difference between the two stimuli were more obvious, in Study 2 texture qualities very similar to each other were deliberately chosen requiring the judge to spend more time to examine the stimuli in order to get an accurate assessment. Two groups of second-year undergraduate students, composed of four and five members respectively, from the Electronic Engineering and Education Department, Andong National University, South Korea were judges in this study. Group 1 was composed of four students with lower analytic skill and Group 2 was composed of five students with higher analytic skill. The latter group performed better than the former group in solving mathematical and programming problems. The scores are given in Fig. 10. Note that Group 1 had lower scores on average and spent less time in the test. Group 2 did much better in the scores but they spent comparatively longer time to complete the task, even though they had good analytic skills.
V. CONCLUSION AND FUTURE WORK

Increasingly, educational tests will become computer based as schools and testing agencies draw on the administrative potential provided by the internet. Computer-based tests can be delivered to students much more efficiently using the Internet, as test production, printing, and delivery costs are dramatically reduced. Another important advantage of computer-based testing stems from the awesome potential that computers provide for developing innovative item formats. The new items will include features such as graphics, audio, video, and animation that are not available with paper-based tests. To-date, very little research has been conducted with innovative computer-based item formats. The purpose of our study is to describe the factors that can affect student performance when the 2 Alternative Forced Choice (2AFC) items are used to evaluate a student’s discrimination capability with 3D objects.

Computer-based testing with innovative item formats provides a complex, interactive, multi-media testing environment for students. This type of testing draws on new item types that can be easily manipulated by the test developer. In the current study, for example, 10 different factors could be manipulated for the 2AFC item type. These factors must be carefully studied and evaluated in each testing situation as they can affect the testing process altering measurement efficiency.

The results of this study reveal that there exist a JND below which the probability of making correct judgment is low. The JND has two implications in computer-based testing: First, item types should be restrained to use stimuli measured below the JND. Second, students with low discrimination capability relative to the JND should be identified prior to or during testing so their scores are not adversely affected. We also noticed that the HVS responds differently to different texture qualities. If the appropriate resolution is used in images, graphics and animations, students can benefit from better quality making efficient usage of network resources in an online environment. Based on the results of our two case studies, we discovered a pattern between a student’s performance in class and his/her score in the evaluation testing; students perform better in class tend to score higher. Good testing time management is also an important factor for assuring good performance. Working continuously without a break is likely to decrease test performance.

CONTRIBUTIONS
Irene Cheng did most of the research planning and all of the implementations and evaluations related to this project. Mark Gierl assisted with the reviews and associations related to educational evaluation. Anup Basu helped with the planning and structuring of the work.

REFERENCES