Proceedings of User Interface Software and Technology'95, Pittsburgh, PA., November, 1995

An Experimental Evaluation of Transparent User Interface Tools and Information Content

Beverly L. Harrison
Dept. of Industrial Engineering
University of Toronto
Toronto, Ontario, Canada
M5S 1A4
beverly@dgp.utoronto.ca

Gordon Kurtenbach
Alias Research Ltd.,
110 Richmond St. East
Toronto, Ontario, Canada
M5C 1P1
gordo@alias.com

Kim J. Vicente
Dept. of Industrial Engineering
University of Toronto
Toronto, Ontario, Canada
M5S 1A4
benfica@ie.utoronto.ca

ABSTRACT
The central research issue addressed by this paper is how we can design computer interfaces that better support human attention and better maintain the fluency of work. To accomplish this we propose to use semi-transparent user interface objects. This paper reports on an experimental evaluation which provides both valuable insights into design parameters and suggests a systematic evaluation methodology. For this study, we used a variably-transparent tool palette superimposed over different background content, combining text, wire-frame or line art images, and solid images. The experiment explores the issue of focused attention and interference, by varying both *visual distinctiveness* and *levels of transparency*.

KEYWORDS: display design, evaluation, transparency, user interface design, interaction technology, toolglass

INTRODUCTION
The central research issue addressed by this paper is how we can design computer interfaces that better support human attention and better maintain the fluency of work. To accomplish this we propose to use semi-transparent user interface objects.

Several key design issues need to be investigated if users are expected to focus on or divide attention between two superimposed semi-transparent images. Can users selectively attend to a chosen "layer" without visual interference from the other?

Are there certain display characteristics or task properties which facilitate or preclude overlapping displays? How do these design choices affect performance? We are conducting a series of controlled laboratory experiments and realistic field studies to answer some of these questions.

This paper reports on one such experimental evaluation which provides both valuable insights into design parameters and suggests a systematic evaluation methodology. For this study, we used a variably-transparent tool palette superimposed over different background content: text, wire-frame images, and solid images. The palette contained text icons, line art icons and solid, rendered object icons. We evaluated both the effect of varying transparency levels (from opaque palettes to highly-transparent palettes), and the interference produced by different types of content information (e.g., Figure 1).

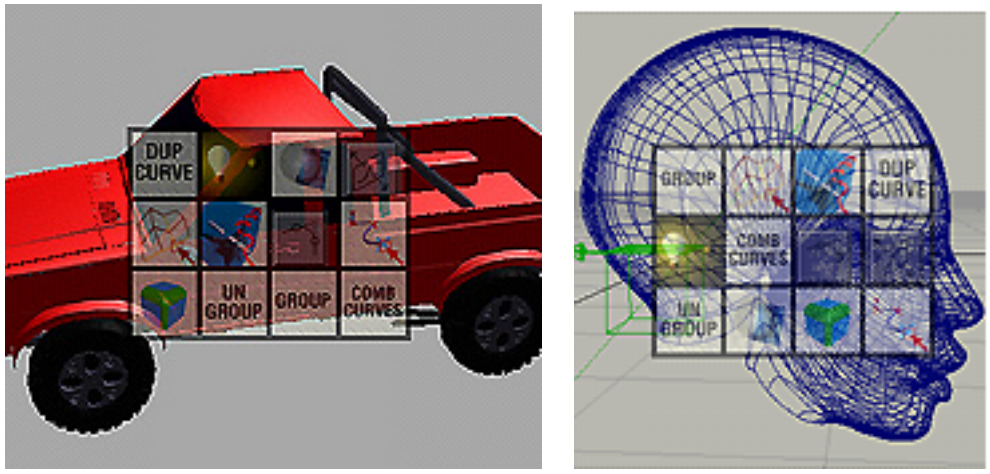


FIGURE 1. Experimental Sample Images

Our approach stems from a technological problem and a psychological problem. The *technological problem* is that

of screen size constraints. Limited screen real estate combined with graphical interface design has resulted in systems with a proliferation of overlapping windows, menus, dialog boxes, and tool palettes. It is not feasible to "tile" computer workspaces to facilitate keeping track of things. There are too many objects. Overlapping opaque objects obscure portions of information we may need to see and therefore may also be undesirable, interrupting our work flow.

The associated *psychological problem* we are addressing is that of focused and divided attention. When there are multiple sources of information (e.g. tool palettes and work areas or multiple windows), users must make choices about what to attend to and when. At times, users need to focus their attention exclusively on a single item without interference from other items. At other times, they may need to time share or divide our attention between two (or more) items of interest. In this case, users rapidly switch attention back and forth between the items (necessitating minimal "switching costs"). Trade-offs among these attentional requirements must be made based on the users' task requirements.

It is our hypothesis that the use of semi-transparent user interfaces can overcome some of these technological and psychological constraints by supporting this attentional trade-off and this will result in an improvement of the fluency of work. In testing this hypothesis, the constraints of the users' existing work domain must be taken into account. Good designs improve task performance by allowing the work to proceed more fluently due to less interference or interruption from the "tools" needed to attain task goals. Our approach is: given an understanding of the task demands, can we manipulate the design characteristics to produce the necessary attentional performance? In particular, does the introduction of less-intrusive, transparent interfaces improve the fluency of work?

INFLUENTIAL EXISTING TRANSPARENT INTERFACES

Transparent user interfaces are not novel per se, though systematic evaluation and experimentation is seldom reported in the literature. A number of researchers and their projects have influenced our thinking with the variety of potential and creative applications where our results might be applied.

Fully transparent designs reflect some of the more advanced interfaces, for example, those used in Heads Up Displays (HUDs) in aviation [10,15], in the Clearboard system [5], or in the original Toolglass/MagicLens project [1,2,13]. In HUD design, aircraft instrumentation (a graphical computer interface) is superimposed on the external real world scene, using specially engineered windshields. In the Clearboard system, a large drawing surface is overlayed on a video image of the user's collaborative partner. The superimposed images are presented on a drafting table–like surface. A predecessor to the Clearboard work, TeamWorkstation [6], showed *partially-transparent* views of a collaborative partner's face or of their hands in a computer window superimposed on a task workspace window. The Toolglass project used clear or see-through palettes which could be aligned with underlying objects. Tools were invoked by clicking "through" the desired function, using alignment to specific the target object for the function. Other semi-transparent designs include such things as video overlays (like those used in presenting sports scores while the game is playing), "3-D silk cursors" [16] or more recent, modified "Toolglass–like" tool palettes [7,8]. Transparency has also been applied to tasks such as map reading [11] and to annotation or technical drawing specification [4].

This collection of intriguing applications demonstrates the variety and novelty of transparent user interface design. Most of these applications are striving towards more integration between task space and tool space, between multiple tools, or between multiple views. The transparency allows these multiple "layers" to be simultaneously observed, avoiding problems in divided attention (but possibly creating problems of interference).

FOCUSED ATTENTION AND VISUAL INTERFERENCE

We are concerned with three critical attentional components: the ability to divide attention between two items, the ability to separate the visual characteristics of each source and focus on any single item with minimal interference from other items, and the switching cost (time, mechanism, learning, awareness) of shifting attention from one item to another. This paper deals exclusively with our ability to focus attention under varying visual interference conditions. This focused attention is a critical component in menu or tool palette selection tasks.

To facilitate focused attention (ignoring information from the background layer while focusing on the foreground) we ideally want to make the attributes of the information on foreground objects as different from the background as possible. We might also wish to reduce the visibility of the background objects (e.g., by increasing opacity of the foreground objects). This will minimize interference. By contrast, for divided attention between foreground and background or for focusing attention on the background layer, we need to be able to see through the object in the foreground (i.e., more transparency). Clearly there is a trade-off between these two attentional goals. We need to support this trade-off since most real world jobs require *both* focused and divided attention (Figure 2).

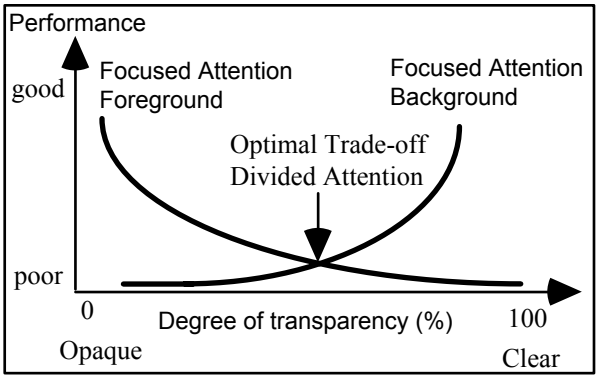


FIGURE 2. A simple model of transparency selection.

As degree of transparency increases, it gets harder to see and hence focus attention on the foreground object. Conversely, higher transparency is required to focus on the background object. The optimal transparency (OT) is a result of a trade-off. The curves and the location of optimal transparency in the figure are hypothetical but may reveal the trend. The non-linear nature of the curves and leveling off points are also proposed and seem to be experimentally supported.

There are many ways of achieving differentiation between layers (with varying success), such as different or distinctive colors, content attributes – analog (images or graphics) versus verbal (text based), font sizes or styles, object movement (motion parallax cues), etc. Many of these features are pre-determined by the user's task, particularly the content of the background or work area. For example, users determine whether they will be doing text editing or drawing, and hence this defines the data content of the work area. However, designers have some control over the presentation of window objects. In particular, the level of transparency can be altered, effecting visibility of the background. Additionally, changes in fonts, colors, sizes, etc. may additionally reduce interference effects. Many of these alternatives suggest experimental scenarios to test.

In the next section we summarize some of our previous experimental findings. We then discuss our progressively more realistic experiments, which test the focused attention aspect of selecting from a foreground tool palette (the topic of this paper).

RESEARCH APPROACH

Within our design space [3], we wish to classify and evaluate a variety of semi-transparent interface objects. Broadly defined these include menus (pull-down, pop-up, and radial or "pie" menus), palettes (tear off tool menus), dialogue boxes (especially interactive scrolling dialogues), windows, and help system screens. These objects appear in many applications and at least temporarily obscure part of our work surface. The degree to which they persist (seconds versus minutes or hours) largely determines how disruptive they may be. In many situations, our primary task or work area becomes the "background" layer while these objects appear in the "foreground". These foreground objects often enable us to carry out activities or change parameters that are ultimately reflected in the now-hidden background layer (e.g., color changes, font changes, view changes). Transparent interfaces allow the user to observe these changes without obscuring the task layer.

We are taking *two* complementary approaches to study our designs: formal experiments and realistic field studies.

To reveal the effects of transparency on focused and divided attention, i.e. how well our model (Figure 2) fits, we are conducting formal experimental studies with well controlled models and simulations (e.g., [3]). However, we realize that controlled experimental paradigms address a restricted set of design dimensions only. Real applications consist of a much richer design space. Therefore, we are developing several prototype systems which are more representative of real world applications. We are evaluating these systems and observing user behavior to gain further insights into the design of transparent user interfaces. These evaluations include progressively more realistic task elements, at the expense of some experimental control. This combined research program allows us to further formulate research issues while remaining confident that our research results have real-world validity. The two approaches are conducted in parallel and as iterative design evolutions.

This paper reports on an experiment which evaluates tool palette selection, superimposed over various background content information (e.g., Figure 1, 7). The objects on the tool palette were taken from existing icons used in a real product, though these icons were not generally used together on the same palette in the real application. We needed icons of three types: text-based, line art, and solid images. This choice allowed us to evaluate content-based interference problems but meant we combined disparate icons from various tool palettes within the product. The background content was also selected from a set of real working images contained in the product library as released to customers. The experiment did not test product usability or icon purpose, but rather evaluated whether subjects could identify randomly selected icons within the palette, given the backgrounds. The use of icon palettes over library images does reflect realistic usage of UI tools for this product. We do not anticipate that the discrepancies from the real application would confound our results, given the goals of the experiment and the background of the subjects.

PREVIOUS EXPERIMENTAL RESULTS

Our first set of formal experiments used a very simple but robust task to measure interference between two layers called the Stroop Effect [14]. In traditional Stroop tasks, a series of words are presented in randomly chosen colors

(e.g., red, green, blue, yellow). Subjects must name the *ink color* while ignoring the word. Some of the words are neutral (e.g., uncle, shoe, cute, nail); other words are the names of conflicting colors (e.g., yellow, blue, green, red). Consistent, significant performance degradation occurs when conflicting color words are used and subjects attempt to name the color of the ink (e.g., the word "red" appears in green ink; the correct response is green). (For reviews of the over 700 experimental permutations on the Stroop Effect see reviews see [12].) The second part of the Stroop Experiment is a word naming task. We had subjects perform a word reading task, ignoring the color patch. This was to determine how legibility of the word was related to transparency levels (background focused attention task).

We experimented with varying levels of transparency using the Stroop Effect. In our experiment, the word is seen by looking "through" the color patch. At high levels of transparency (e.g., 100% - clear) we anticipate that users will experience high levels of interference from the background word when they try to name the foreground color (difficulty in focused attention on the foreground). In the color naming task, as the color patch becomes more opaque the interference from the background word should decrease (making focused attention easier). The exact opposite should occur for the word naming task (i.e., performance improves with increases in transparency level). This experiment is reported in detail in [3]. The results which are most relevant to this paper are summarized briefly below.

The Stroop test was used to evaluate interference between transparent layers because it provides a sensitive, robust, extreme measure of the extent of interference. As such, it should suggest worst case limitations. In the color naming task, our results suggest that when focusing on the foreground color patch while ignoring the background word, there is a rapid performance degradation between 5% and 20% transparency (Figure 3). For this degradation to occur, the background word must be introducing interference at ever increasing levels. At levels of 5% (and less) minimal or no interference seems to take place, implying that the background word is no longer visible enough to create even minor interference (leveling off point). At 50% transparency, performance is at it worst and does not deteriorate substantially with further increases in transparency.

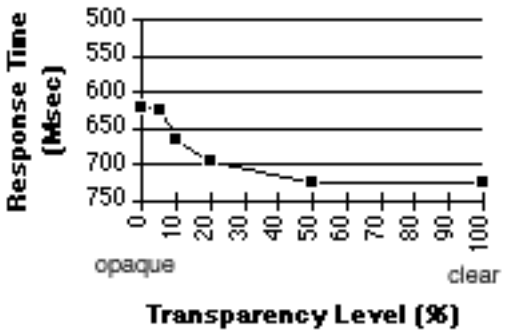


FIGURE 3. Mean Response Time Results from the Stroop Experiment - Color Naming Task

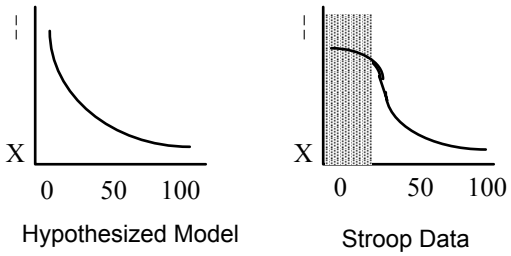


FIGURE 4. Rough visual comparison of actual data to predicted model. Shaded region represents the major difference between the actual data and the model predictions.

A rough visual comparison of the results to the hypothesized model are illustrated in Figure 4. While the curve shape and high-end leveling off point seem to be accurately predicted, we did not predict leveling off points at *both* ends of the curve. In hindsight the low-end leveling off seems reasonable since it indicates the point at which the interfering text is no longer legible and hence performance is not further impacted. It would seem reasonable to modify our hypothesized model to take this into account.

The results for the word naming task did not fit our predicted model well (Figure 5, 6). We believe that this is mainly because the legibility task was extremely simple (Helvetica, 78 point letters tend to be either illegible or very easy to read). This resulted in a much steeper curve indicating rapid performance improvements. We observed a leveling off point at the high end of the curve, indicating that after 50% transparency, performance matched the word only condition (normal reading condition). The predicted low end cut leveling off point did not occur. We believe that this was because users chose the "cannot see any word" option when the effort to read the word became excessive. Additionally, levels lower than 10% were quite error-prone (15-20% errors), and error trials were excluded from the response time analysis. If we were to run a finer analysis using more transparency levels around the lower limit (between 1% to 10%), we might find a more gradual cut-off which better fits our predicted model. It seems reasonable to modify out hypothesized model to include

this high-end leveling off point. We would need further and more precise data to determine whether we should additionally adjust the slope of the curve or the lower end cut off point.

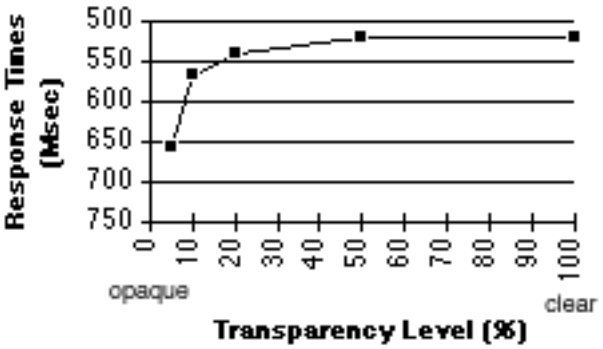


FIGURE 5. Mean Response Time Results from the Stroop Experiment - Word Naming Task

The Stroop Experiment tested a specific and well-known attentional measure for task interference. However, the task components used are visually very simplistic and dissimilar: color and text (though they are *semantically* conflicting). Clearly we are interested in more complex

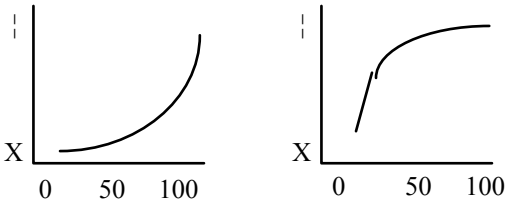


FIGURE 6. Rough visual comparison of actual data to predicted model.

images which better reflect the characteristics of real world tasks. Applying transparency in such domains will almost certainly introduce visually conflicting conditions. Using these more realistic complex image types, we are interested in understanding how our attentional model fits, where the cut-off points in performance are, and the shape of the resulting performance curves. At which points can we select items from the palette with "reasonable" performance, while still maintaining a visual awareness of the background image?

EXPERIMENT - LEGIBILITY AND INFORMATION CONTENT

This experiment attempts to further explore the issue of focused attention and interference, this time varying both *visual distinctiveness* and *levels of transparency*. To accomplish this we had subjects perform a tool palette selection task where the tool palette appeared in the foreground and various images appeared in the background (e.g., Figure 1). The palette transparency levels varied in random order: sometimes the palette was opaque (0%

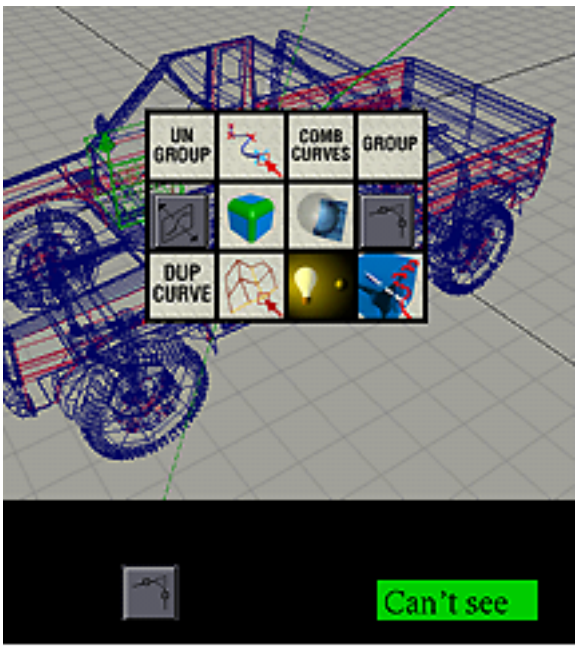


FIGURE 7. Sample Trial Screen showing target icon, stimulus image and "can't see" option

transparent), blocking out the background, other times the background could be easily seen *through* the palette (e.g., 90% transparent). Visual distinctiveness was assessed by the combination of both palette icons and backgrounds. These contained text, line art or wire-frames, or solid images. All combinations of palette icon types X background types X transparency levels were run.

For each trial within the experiment, subjects were shown a target icon image to study. When ready, they pressed a "next trial" button which displayed the palette superimposed over the background at a randomly ordered transparency level (e.g., Figure 7). Icons were randomly distributed on the palette. Subjects had to locate and click on the target icon within the palette. If they could not see any items on the palette (i.e., illegible) they could press a "can't see" button. Response times and errors were logged. The target icon remained on the screen throughout the trial for reference purposes.

Applying Our Predictive Model and Stroop Results

We have briefly outlined a hypothesized model and we later compared this model to results from the Stroop Experiment. Several possible explanations for the differences between predicted and actual data were put forward. The palette selection experiment reflects components of both Stroop tasks. It represents a *foreground* focused attention task and, as such, we anticipate a performance curve which resembles those depicted in Figure 4 (i.e., opaque levels should have "good/fast performance" and transparency increases should degrade performance). However, unlike the color naming task, the palette selection task itself is more similar to a legibility or word naming task. Performance is unaffected

by semantic interference but is expected to be sensitive to changes in visibility.

Our previous Stroop word naming results showed little performance difference from 50% transparent to the "best reading condition" (100% transparent in the case of a background task and presumably 0% in the case of a foreground task). This suggests that the resulting palette selection task might achieve maximum performance at 50% with little significant improvement between 50% and "best reading condition" (0% or opaque). Furthermore, in our Stroop word naming task, levels below 10% were found to be error prone or often illegible. (Levels of 10% transparency when alpha blended roughly means 10% of the word and 90% of the color patch formed the resulting displayed image.) If we again translate this value to our foreground palette selection task, 90% transparency may be a cut-off point (where 90% transparency roughly means that 90% of the background image and 10% of the palette image are used to create the displayed image).

Hypotheses (stated as null hypotheses)

H1: As transparency level increases (i.e., the background is more visible through the icon palette) the response time and errors will be unchanged.

We anticipate *more* interference as transparency increases and therefore reduced performance (slower response time and increased errors).

H2: The content of the background image (text, wire frame, solid) will have no interaction effect with legibility of the icons.

We expect two interaction effect. First, we anticipate that increased complexity or information density on the background will make icon legibility decrease for *all icons types*. Text backgrounds will have the worst performance, followed by wire-frame, then solid images. Second, we also anticipate that visually similar icons types to background types in terms of both colors and content will be most significantly effected in terms of performance degradation (i.e., text icons with text background, line art icons with wire frame backgrounds, and solid image icons with solid image backgrounds).

Lastly, we wish to verify whether our proposed cut off points at 50% and 90% are reflected in the data.

Experimental Design

A fully randomized, within subject, repeated measures design was used. There were three independent variables: type of palette icon, type of background, and transparency level. A total of 576 trials were run for each subject. Trials were presented in random order. Each session lasted about 45 minutes. Dependent variables of selection response time (based on a mouse click) and errors were logged. Two error conditions were possible: the subject pressed the "can't see" button indicating that the item was not legible, or the subject selected the incorrect palette item. In the latter case, the item selected and it's location were logged. Error trials were removed from subsequent analysis of response time data. Error data was analyzed separately.

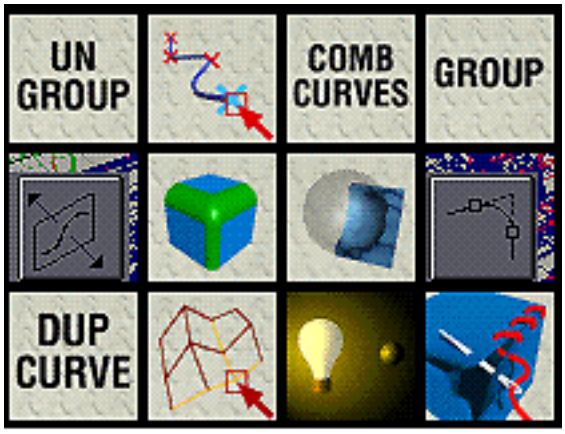


FIGURE 9 Sample Palette (Opaque)

We used three icon types: text, line art, and solid rendered objects (Figure 9). Within each of these three types, we selected four samples from existing product icon palettes. Color icons were used (not shown here). Our resulting tool palette was 3 rows by 4 columns in size. A 12 item palette was felt to be representative of the average menu/palette size used within the actual product. Icons were randomly assigned positions within the palette for each trial. This was done to ensure the experiment was a *legibility* test and not confounded by subjects learning the position of icons. Subjects could not predict the palette location of an icon target based on the item presented, they had to find it each time. The target was presented to the subject throughout the trial as a reminder. This was to prevent memory errors (which we were not testing for).

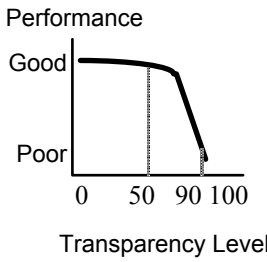


FIGURE 8. Projected curve for the Palette Selection Experiment

We randomly assigned background images of three types: text pages, wire frame images, and solid images. Again four samples of each type were created. Images were 8-bit color rendered images. These backgrounds were aligned such that a major portion of the content was directly under the palette.

Finally we randomly assigned the level of transparency to the palette. These levels were based on our previous

experimental experience [3] and test pilot results with this experiment. Initially levels of 0% (opaque), 10%, 20%, 50%, 75%, 90% and 95% (highly transparent) were used. (100% transparent actually represents the background image only, therefore 95% was tried as an upper limit.). The opaque level represented the baseline condition where the fastest performance was anticipated. Pilot results suggested no substantial performance improvements between 0% (opaque) and 50% (semi-transparent) so intermediate levels within this range were not included in the final experiment. Similarly, images above 90% transparency were found to be completely illegible and were also not included. In summary, levels of 0%, 50%, 75%, and 90% were used.

Experimental System Configuration

The experiments were run on an SGI Indy[a] using a 20 inch color monitor. Subjects sat at a fixed distance of 60cm from the screen (average distance when working normally).

Procedure

Subjects were given 20 practice trials. These trials were randomly selected from the set of 576 possible combinations. Following this, subjects were shown the target icon for each trial and a button to start each trial when they were ready. They could take short rest breaks whenever necessary. Response times and errors were logged. Response selections were made using the mouse. Subjects were debriefed at the end of the experiment. Open ended comments were recorded.

Subjects

A total of 14 students from the University of Toronto were run as subjects. They were pre-screened for color-blindness and for familiarity with the product from which the images and icons were taken. Subjects were paid for their participation and could voluntarily withdraw without penalty at any time.

RESULTS

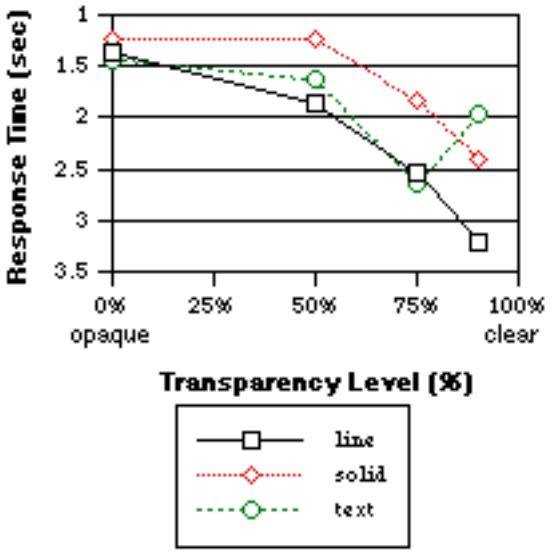
Pilot testing revealed that there did not seem to be any significant performance differences between 0% (opaque) to 50% (semi-transparent). Most of the noticeable differences reflected in performance and legibility seemed to occur between 50% and 90%. Our pilot data seem to partially confirm out initial projected cut-off points (Figure 8). Subsequent detailed analysis was conducted on 14 subjects and is reported below.

Quantitative Statistical Analysis - Response Time

The experimental results indicate highly, statistically significant main effects ($p < .0001$) for icon type, background type, and transparency level. A significant interaction effect ($p < .0001$) was found for: icon type X transparency level, background type X transparency level, icon type X background type and icon type X background type X transparency level. These results are shown in Table 1 below. (All statistics reported in this paper used an alpha level = .05.) A graphical summary of the mean response times for icon type is shown in Graph 1. The mean response times for background type are shown in Graph 2.

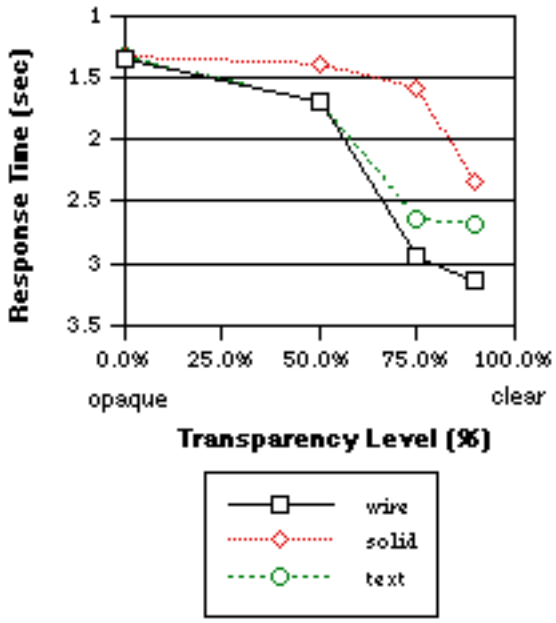
condition icon X bkgrnd X transp	df	F value	p<
icon type	11	8.31	.0001
background type	11	6.81	.0001
transparency level	3	39.04	.0001
icon type X transp	33	6.49	.0001
bkgrnd type X transp	32	9.42	.0001
icon type X bkgrnd type	121	2.81	.0001
icon X bkgrnd X transp	275	4.81	.0001

TABLE 1. Statistical Results for Main Effects and Interactions



GRAPH 1. Mean Response Times for Transparency Levels X Icon Type (across all background types)

To determine if the differences are significant between the individual lines plotted within each of the graphs, a Student-Newman-Keuls (SNK) test was run as a comparison of means. (This determines the clustering of items within icon type, background type, and transparency level. and indicates which items are not statistically different.) For *overall* response time performance per transparency level (across all data points), the groupings were: 90% + 75% (slowest), 50%, and 0% (opaque - fastest). (This measure is not particularly meaningful by itself given the variance by collapsing so many conditions, however it does provide a gross measure of the impact of transparency)



GRAPH 2. Mean Response Times for Transparency Levels X Background Types (across all icons types). There is no statistical differences between points at 100%.

For background type and icon type one would anticipate that 3 groupings would occur which represent the 3 types of items (text, line art/wire frame, and solids). The statistically significant groupings are shown in Figure 10 (collapsed across all transparency levels).

	Background Groupings	Icon Groupings
fastest (good)	Camcorder solid image	All solid icons
	All other solid images	All line art icons
	Wire-frame human head	All text icons
	Wire-frame camcorder	
	All other wire-frame images	
slowest (poor)	All text pages	

FIGURE 10. Statistically Significant Groupings Across Transparency Levels

A detailed analysis was run at each level of transparency (Figure 11). Note that from the graphs we would anticipate that solids were grouped together and fastest, while line art and text performed similarly until highly transparent levels (75% and 90%). The graphs show that text performs better than line art. Figure 11 summarizes which points on the graphs are *not* statistically different. It supports our basic assumptions though wire frame backgrounds and line art performed more poorly than expected.

	Background Groupings	Icon Groupings
fastest (good)	0% (opaque) - no difference	
	50% Transparent	
	All solid images	no difference
	All wire-frame images	
	All text pages	
	75% Transparent	
	All solid images	All solid icons
	All text pages	All line art icons
		All text icons
	All wire-frame images	
slowest (poor)	90% Transparent - no difference	

FIGURE 11. Statistically Significant Groupings Within Transparency Levels

A detailed analysis was also run based on background types to determine how the icon types interacted with the background types. Wire frame backgrounds and text backgrounds were not statistically different and resulted in groupings of icon types: text and line art icons (slowest, no difference), and solid object icons (best). On solid image backgrounds there was no difference between any of the icon types (all performed equivalently well). Wire frame backgrounds (with any icon type) showed no statistical difference in performance between 50% and 0% (opaque). Solid backgrounds (with any icon type) showed no statistical performance difference between 75% and 0% (opaque). Text backgrounds showed significant differences for all transparent levels.

Targeting Error Results

Error trials were removed from the analysis of response time data and were subsequently analyzed separately. In total less than 2% of the trials resulted in *targeting errors* or misses. This suggests that subjects were not guessing when targets were difficult to see. The breakdown of misses is shown in Table 2 below.

For targeting errors, in every case it was due to substituting another icon of the same category for the target icon (e.g., an alternate incorrect text icon was selected instead of the target text icon). Text icon substitutions accounted for 30.1% of total targeting errors respectively, solid object icon substitution 29.7%, and line art icon substitution 39.7%. No targeting errors were due to accidental selection of adjacent items in the palette, suggesting the icon size used was adequate for selection accuracy. More targeting errors occurred at the 75% level versus the 90% level. since users were more inclined to try selections at 75%. At the

90% level users typically marked trials as illegible instead of attempting selection.

transparency level	number of trials	% of total misses
0% - opaque	5	0.06%
50%	20	0.2%
75%	53	0.6%
90% - mostly clear	42	0.5%

Table 2. Errors due to target misses

Legibility "Error" Results

The most frequent source of "error" were trials that the subjects marked as "can't see" (which we believe prevented subjects from guessing randomly). In total, 18.4% of the trials were marked "can't see". The breakdown by transparency level is shown in Table 3. Note that 3/4 of the legibility errors occurred at the 90% level.

transparency level	number of "can't see" trials	% of total legibility errors
0% - opaque	0	0%
50%	20	1.3%
75%	350	23.6%
90% - mostly clear	1113	74.9%

Table 3. Trials marked as illegible

At the 90% level, *all* of the icon types appearing over *text or wire-frame backgrounds* were illegible. This accounted for 98% of the legibility errors at this transparency level (interestingly, only 2% of the illegible trials at 90% transparency were *solid* backgrounds). A rough breakdown by icon type X background type is shown in Table 4.

icon type X bkgrnd type	transp level	% of legibility errors
all X text	90%	46.9%
all X wire-frame	90%	49.5%
solid X all	90%	3%
all X text	75%	37.4%
all X wire-frame	75%	62.5%
line X wire-frame	50%	45%
line X text	50%	65%

Table 4. Trials marked as illegible by type and transparency level

Further investigation showed that line art icon types appear the most problematic across transparency levels (42% of total). (Solid icons were 25% and text icons were 33% of total legibility errors respectively.)

Qualitative Results

In general, subjects felt that they adopted a "categorization" strategy for locating target icons. The categories were based on icon type (text, line art, solid) and then distinctive features within that type (e.g., shape, color, length of text). Subjects first searched for items belonging to the target icon category. They then used the most distinctive features to either locate the target icon or to eliminate the contender icons. All subjects felt the "line art" icons were the most difficult to discriminate since subjects had to search for tiny differences. Subjects commented that solid object icons seemed easier to find, independent of transparency level or background types. They used object color as a major cue. All subjects found the light bulb icon the easiest, primarily because no others had similar colors. Text icons were discriminated based on the shape of the words.

Wire frame backgrounds were perceived as most difficult for any icon type. The more dense wire frames were perceived as slightly easier. Subjects found the darkest solids were easiest (e.g., the camcorder) and commented that the palette seemed to "stand out" best on these images. Most notably, several subjects commented that after a number of trials the opaque palettes seemed "too bright" or "annoying" and that they were "used to the partial transparency".

Although the position of the target icon was randomly selected to avoid learning effects, subjects commented that they eventually learned the entire set of 12 icons (without positional information). This reduced the time required to "study" the target when it was presented. This learning also enabled subjects to eventually determine what they considered to be the emergent features of each icon and how that icon related to the whole set. For example, several subjects commented that since the icon set did not change, it was possible to search for the "light bulb" since they knew there was no chance of getting another light-bulb-like icon.

DISCUSSION

As expected, icon type, background type, and transparency all effected response time performance (Hypothesis 1). Given the response time and error rates combined, 90% (highly transparent) palettes seem unusable. In most cases transparency levels of 50% and 0% (opaque) seem to work about equally well, independent of icon type or background. Our data support both our predicted performance curve (Figure 8) and the proposed cut off points of 50% and 90%.

Contrary to our expectation, wire frame backgrounds seem to perform slightly worse than text (solid backgrounds were the best). The performance difference is more pronounced with increases in transparency. Subjects commented that the wire frame images seem very visually complex and hence interfered the most. This also held for line art icons versus text icons: line art were as bad or worse than text icons as transparency increased. Our hypotheses overestimated the performance degradation resulting from text objects (relative to wire frame or line art). Both solid backgrounds and solid object icons are most resistant to interference and provide the best selection performance.

It seems that the density of some of the wire frame background images skew (and improve) performance more towards solid image performance (e.g., the camcorder wire frame). Additionally within the solid images, we believe that contrasting luminance levels improved performance on the mostly black camcorder background (the palette icons were colored or grays primarily). (We also found a similar effect for color in the earlier Stroop Experiment.) This difference in luminance as it related to visibility was noted by several subjects.

All subjects commented about learning effects and categorization schemes used to facilitate locating items on the palette. Subject performance improved slightly over time and as familiarity with the set of icons improved. In a normal work context, we could assume that the most frequently used icons would likewise become well-known. Subjects would eventually learn which features distinguish the icons that they most frequently use. When this information is combined with consistent palette position, subjects will likely perform well, even if the icons are not clearly visible (i.e., highly transparent).

One aspect of most experimental studies which may exaggerate the error rates is the cost or consequences of errors. In this experiment, although errors were logged, there was no "cost" associated with an incorrect guess. This may have led subjects to select items more often by guessing than one would observe in a real work environment where errors have consequences.

FUTURE RESEARCH DIRECTIONS

The above experiment gave us insights into some of the upper and lower threshold values for transparency as it relates to the visual distinctiveness of the two layers. It suggests some conditions under which transparency works well and works poorly. However, in this experiment we restricted our evaluation to the level of interference as it related to focusing on the foreground information. We additionally need to run an experiment which suggests where the performance cut-off points lie and the shape of the performance curve for selecting objects from the background (i.e., background legibility). Taking the combined results of these two experiments may indicate optimal trade-off points in the design space. This will be our next experiment.

The experiment reported here varied the location of the target item on the palette randomly with each trial to avoid learning effects. This provided a more accurate assessment of palette legibility. However, in real applications the location of items on a palette is generally constant and predictable. Results from a prototype system suggest that as familiarity with the interactive window layout improved, users preferred corresponding increases in transparency. They preferred to see "less" of the interactive dialog boxes and more of the underlying image. The dialog box items were needed only as outlines to target selections - the actual legibility of the text was substantially less important. It may be possible (or desirable) to handle the borders of windows and buttons and data entry areas in a different way than the actual names and labels. This suggests new and intriguing possibilities for dynamically evolving interfaces based on increased expertise. We wish to experimentally test this using more longitudinal studies of skilled users.

Finally, for simplicity and more experimental control we used static images to test our current attentional model. In more realistic applications, users would be moving the palettes and windows, particularly if the UI tools resemble Xerox PARC ToolGlass/MagicLenses. We know from preliminary prototyping (and the literature in visual perception) that motion parallax greatly helps users discriminate which features belong to which objects. We believe that the addition of motion will also benefit our transparent UI tools. This remains to be experimentally evaluated.

CONCLUSIONS

We have described a method of empirically testing transparent UI tools within the context of a attentional framework. We attempted to illustrate an approach which combined realism (by using actual content information) with experimental control to systematically evaluate user performance. Our results suggest design parameters for tool palettes which not only relate to transparency levels but also to icon design. While we recognize the limitations of working with static fixed images, we have proposed a series of progressively more realistic, experiments which more closely reflect existing or new user interface tools.

We believe that interface designers can take advantage of both the intrinsic properties of the task and of an understanding of human visual attention to design new display techniques and systems. We believe that results thus far show promising advantages for creating new user interfaces and interaction techniques. We are exploiting possibilities of new technology in a way that is sensitive to both psychological and task constraints.

ACKNOWLEDGMENTS

Primary support for this research is gratefully acknowledged from Alias Research Inc. We also wish to particularly thank Bill Buxton for his substantial insights and contributions to this research program. Support for our laboratory is gratefully acknowledged from the Natural Sciences and Engineering Research Council (NSERC), Apple Computer. the Information Technology Research Centre (ITRC), and Xerox PARC. We would also like to thank Dr. Hiroshi Ishii, Dr. Colin MacLeod, Dr. Chris Wickens, Shumin Zhai, and members of the Graphics Lab and Cognitive Engineering Lab for their comments and contributions.

REFERENCES

1. Bier, E. A., Stone, M. C., Pier, K., Buxton, W., and DeRose, T. D. (1993). Toolglass and magic lenses: The see-through interface. *Proceedings of SIGGRAPH'93*. Anaheim, CA. 73-80.

2. Bier, E. A., Stone, M. C., Fishkin, K., Buxton, W., and Baudel, T. (1994). A Taxonomy of See-Through Tools *Proceedings of CHI'94*, Boston, Mass. 358-364.

3. Harrison, B. L., Ishii, H., Vicente, K. J., Buxton, B. (1994). Transparent Layered User Interfaces: An Evaluation of a Display Design Space to Enhance Focused and Divided Attention: . *Proceedings of CHI'95,* Denver, Colorado. . 317-324.

4. Kramer, A. (1994). Translucent Patches: Dissolving Windows. *Proceedings of UIST'94*. 121-130.

5. Ishii, H. and Kobayashi, M. (1992). Clearboard: A seamless medium for shared drawing and conversation with eye contact. *Proceedings of CHI'92,* Monterey, CA, 525-532.

6. Ishii, H. (1990). TeamWorkstation: Towards a Seamless Shared Workspace. *Proceedings of CSCW'90*, Los Angeles, CA. 13-26.

7. Kabbash, P., Buxton, W. A. S., and Sellen, A. (1994). Two-handed input in a compound task. *Proceedings of CHI'94,* Boston, MA., 417-423.

8. Kabbash, P. and Buxton, W. A. S. (1995). The Prince Technique: Fitts' Law and Selection Using Area Cursors. *Proceedings of CHI'95*, Denver, Colorado. 273-279.

9. Kobayashi, M. and Ishii, H. (1994). DispLayers: Multi-Layer Display Technique to Enhance Selective Looking of Overlaid Images. Poster from *CHI'94 Conference*, Boston, MA.

10. Larish, I and Wickens, C. D. (1991). Divided Attention with Superimposed and Separated Imagery: Implications for Head-Up Displays. *University of Illinois Institute of Aviation Technical Report* (ARL-91-4/NASA HUD-91-1).

11. Leiberman, H. (1994). Powers of Ten Thousand: A Translucent Zooming Technique. Demonstration at *UIST'94*.

12. MacLeod, C. M. (1991). Half a Century of Research on the Stroop Effect: An Integrative Review. *Psychological Bulletin,* Vol. 109, No. 2, 163-203.

13. Stone, M. C., Fishkin, K., and Bier, E. A. (1994). The Movable Filter as a User Interface Tool. *Proceedings of CHI'94*, Boston, Mass. 306-312.

14. Stroop, J. R. (1935). Factors affecting speed in serial verbal reactions. *Journal of Experimental Psychology,* 18, 643-662.

15. Wickens, C. D., Martin-Emerson, R., and Larish, I. (1993). Attentional tunneling and the Head-up Display. *Proceedings of the 7th Annual Symposium on Aviation Psychology,* Ohio State University, Ohio, 865-870.

16. Zhai, S., Buxton, W., and Milgram, P. (1994). The "silk cursor": Investigating transparency for 3D target acquisition. *Proceedings of CHI'94,* Boston, MA., 459-464.