

Evaluating Glanceable Visuals for Multitasking



*Tara Lynn Matthews
Devin Blais
Aubrey Shick
Jennifer Mankoff
Jodi Forlizzi
Stacie Rohrbach
Roberta Klatzky*

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2006-173

<http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-173.html>

December 13, 2006

Copyright © 2006, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Acknowledgement

Maneesh Agrawala, Bill Prinzmetal, Sara Kiesler, Ken Koedinger, Bonnie John, Laura Dabbish, Carnegie Mellon common meeting participants, and DIS Doctoral Consortium participants. This work was supported by an NSF fellowship and NSF IIS-0205644, IIS-0501895, IIS-0325351.

Evaluating Glanceable Visuals for Multitasking

Tara Matthews¹, Devin Blais², Aubrey Shick², Jennifer Mankoff², Jodi Forlizzi^{2,3},
Stacie Rohrbach^{2,3}, Roberta Klatzky^{2,4}

¹EECS Department
U.C. Berkeley, CA, USA
tmatthew@cs.berkeley.edu

²HCI Inst., ³School of Design, ⁴Psychology Dept.
Carnegie Mellon University, Pittsburgh, PA, USA
{jmankoff, forlizzi, slr, klatzky}@cs.cmu.edu

ABSTRACT

Glanceable visuals enable quick and easy visual information uptake, thus enabling users to monitor secondary tasks while they multitask or divide attention. However, little is known about how to best design visual information for divided attention situations. We present two experiments to address this question, which differ from past work in three ways: (1) We study information uptake speed for *peripheral displays in dual-task situations*; (2) we examine a *wide range of renditions* (graphic objects or text) inspired by existing displays, differing in both visual complexity and the degree to which they convey common meanings; and (3) we investigate how recognizable renditions are together as a set, and how this changes with different set sizes. Our main contributions are best practices for the design and evaluation of glanceable visuals, intended to help designers create better peripheral displays to support multitasking.

Author Keywords: Information visualization, glanceable, peripheral displays, abstraction, multitasking

ACM Classification Keywords: H5.2 *User Interfaces*—Graphical user interfaces. H5.m *Miscellaneous*.

INTRODUCTION

Managing multiple tasks and interruptions is a challenge for information workers, who typically balance 10 basic units of work at once, spend 3 minutes on a task before switching, and are interrupted about once per task [7, 11]. Because glanceable displays require less attention, they better enable users to monitor secondary tasks while multitasking [15].

Little is known about how to best design glanceable visuals for performance-oriented, peripheral displays. By *glanceable*, we mean enabling quick and easy visual information uptake, which is equivalent to Mullet’s *immediacy* principle for design [19]. Our contribution is determining which visual characteristics of glanceable renditions of information best improve multitasking performance. It is commonly believed that simple peripheral display designs are better, but is there such a thing as too simple (e.g., geometric shapes to represent email sender groups [24])? It is also commonly believed that designs should intuitively convey their intended meaning to users without training. Are there situations when it is better for displays not to do so (e.g., items in a beach scene representing various streams of information [20])? Existing peripheral displays challenge common design principles and more knowledge is needed

to inform the effective use of very simplified visuals and indirect mappings that the user memorizes. From here forward, we use the term *renditions* to describe individual graphic objects or text used to convey categorical information to a user, which we test in our experiments.

To inform the design of glanceable visuals, we focus on studying *abstraction*. The term *abstract* has been inconsistently used to describe the way peripheral displays convey information. Renditions have been called abstract for two distinct reasons: (1) less visual complexity (reduced detail), or (2) less symbolism (less capable of conveying common meanings [1]). *Complexity* and *symbolism* are two continuous characteristics that create a space for the design of peripheral displays (Figure 1). For example, if a rendition is high-symbolism (e.g., a picture of coworkers standing for “coworker” as in the top-right of Figure 1), then a simple rendition (e.g., a drawing) is *more abstract* and a complex rendition (e.g., a photo) is *less abstract*. Similarly, low-symbolism renditions (e.g., the colored square, left side of Figure 1) are more abstract than high-symbolism renditions (e.g., the ‘c’, right side of Figure 1).

Another important issue in the design of glanceable displays is how renditions will fit together as a set and still remain individually recognizable, and how this changes with the set size. For example, variations on a simple, low-symbolism visual like a colored square might effectively represent a few email sender groups, but would the same format be effective for many? Past work surveys peripheral display taxonomies and suggests that symbolism and rendition set sizes are important considerations, categorizing displays using similar concepts [21].

Our main contributions are best practices for the design and evaluation of glanceable visuals based on empirical and qualitative results from two experiments: (1) The Unlearned Study compares many renditions users have not seen before; and (2) the Learned Study compares renditions that users learn to identify in sets of 3 or 7. Renditions vary in *visual complexity* (we shorten this term to *complexity*) and *symbolism*, two important tradeoffs in the design of glanceable displays. Among our key findings are that qualitative ratings are especially important, *high-symbolism* renditions only lead to improved performance when unlearned, and *simple* renditions lead to improved performance for learned renditions but people tend to like *complex* renditions more for larger set sizes. These contributions provide design and

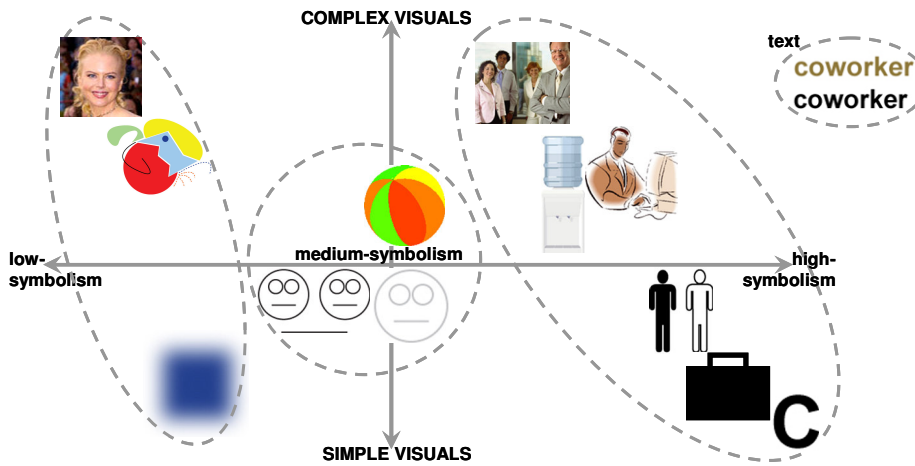


Figure 1: Design tradeoffs (simple↔complex and low↔high symbolism) populated with “coworker” renditions selected for the Unlearned Study. Celebrity photo is from movies.yahoo.com, beach design is from InfoCanvas [20], top-right quadrant images are Microsoft clipart, abstract art was redrawn to protect copyright.

evaluation knowledge that will enable designers to create better peripheral displays to support multitasking.

Our studies focus on the email domain, which can benefit greatly from glanceable displays. People are often distracted by email, which can harm their productivity [7]. At the same time, email is an important work tool that often requires regular monitoring. Knowing whether a new email is important enough to interrupt the current task or can be ignored could significantly improve a user’s ability to maintain task flow and resume tasks at opportune times [15]. Studies have shown that knowing which group a sender belongs to (*e.g.*, coworker, family, *etc.*) is an important factor in deciding when to read a message [8]. Our experiments compare various *sender group* renditions that could replace or enhance existing email notifications or displays (see Figure 2 for an example). Though we map our renditions to email sender information, we study characteristics of these renditions that can be applied to other applications.

A difference between our work and past work is that we aimed to gather practical design knowledge by studying a wide range of renditions, all relevant to a real issue (email). To alleviate concerns of confounds, we chose renditions to evenly cover the tradeoffs, complexity and symbolism (Figure 1). Our results include several correlated metrics that reduce the suspicion of confounds.

We define the scope of our studies in four ways. First, we focused on performance-oriented displays. Second, we studied glanceability rather than *peripheral vision*. Though some renditions afforded peripheral vision with practice, we let users identify renditions in the way they chose – using focal or peripheral vision. Third, we studied information



Figure 2: Mockup of an email display enhanced with a glanceable sender group rendition from our studies.

uptake rather than attention capture. Our studies consistently had stimuli appear abruptly, which can be expected to evoke a response; however, this effect was constant across conditions and should not mask differences in information uptake. Finally, our study was not purely an icon design study. An *icon* is defined as a picture that resembles the thing it represents [18]. We compared various icons and *symbols* (a design with an arbitrary relationship with the thing signified; *e.g.*, text, colors [18]) to recommend the most glanceable visuals for a categorical piece of information.

Overview

We now turn to relating our studies to work on peripheral display studies, cognitive science, human factors, semiotics, and design. Then we describe our formative work exploring glanceable design tradeoffs and selecting designs through interviews and pretesting. We then present our experimental designs and results. We discuss best practices that emerged from our studies and several unexpected findings. We conclude with a discussion of future work.

RELATED WORK

Three characteristics distinguish our research from past work: (1) We study information uptake speed for a *peripheral display in a dual-task situation*; (2) we examine a *wide range of renditions* inspired by existing displays, differing in visual complexity and symbolism (Figure 1); and (3) we investigate how recognizable renditions are together as a set, and how this changes with different set sizes.

The design of glanceable visuals is informed by research in many areas: cognitive science, human factors, semiotics, and design. Since this literature is too large to survey here, we list the most relevant empirical results in Table 1, in which empty cells highlight open questions. Though empirical studies have explored identification and/or search speeds of text [5], color [6], high-symbolism icons [17], and combined text and high-symbolism pictures [2], the effects of using these visuals in peripheral displays is unknown. Visual search studies (upon which many of the empirical results are based) are not representative of actual usage of peripheral displays. Empirical studies of peripheral displays in HCI literature have addressed issues tangential to those studied in this work. For example, a single-task study of the InfoCanvas display showed that pictorial renditions enable recall of more info than text or Web portal displays [20]. A dual-task study showed that finding items on a peripheral displays may not significantly hinder a primary task [22]. However, most dual-task empirical work to date has been on interfaces for safety-critical situations like flying or driv-

ing. Note that these applications are fundamentally different from peripheral task monitoring for information workers. Further, visuals used in past human factors and icon design studies are not characteristic of the visuals in existing peripheral displays, largely because they have focused on icon-function relationships rather than conveying information for secondary task monitoring. This leaves open the exploration of lower-symbolism renditions that may quickly convey information.

Applications that make excellent use of low-symbolism renditions for multitasking have been evaluated: the Scope [24] (email, instant messaging, alerts, calendar) and Info-Lotus [27] (email). Studies of these systems show that low-symbolism renditions can be effective for performance-oriented displays. Our results will inform the design of similar displays using low- and high-symbolism renditions.

From a rich body of empirical work (partially listed in Table 1), icon design guidelines have been created for in-vehicle usage [3] and public information [28]. Both guidelines stress the importance of universal interpretability by using high-symbolism, simple drawings, because they assume viewers will have little or no training with the icons. Glanceable display users will often learn the interface, so these design guidelines are of limited applicability here.

In a past study [15], we explored the effects on desktop multitasking performance of three abstraction techniques used in peripheral displays. Results showed that the peripheral display enabling the best performance showed information relevant to switching tasks, combined with a simple visual cue to signal changed content. In a past, single-task, visual search study [14], we compared learned text, numeric, and simple + low-symbolism renditions of map information. Results showed that text led to the fastest search times. A critical next step is understanding how to convey a wider range of task information with glanceable visuals, an issue we address in this work.

Table 1: Summary of results from related empirical studies.

| | | 1 Task | >1 Task |
|-----------|---|--------|---------|
| unlearned | High-symbolism icons can be processed faster than text (1 task) / similar to text (>1 task). | [5] | [5] |
| | Text can be processed faster than low-symbolism pictures. | [5] | [5] |
| learned | High-symbolism icons can be processed faster than text (>1 task). Text can be processed faster than low-symbolism (1 task). | [14] | [5] |
| | Color leads to faster search & mixed code identification accuracy compared to size, brightness, geometric shapes, & other shapes. | [6] | [6] |
| | Color's beneficial effects lessen with learning. | [6] | [6] |
| both | Concrete (depicts a real object) vs. abstract icons*: concrete lead to faster search times unlearned and have no effect when learned. | [17] | |
| | Complex vs. simple icons*: complex lead to slower search & identification, unlearned & learned. | [17] | |

*All icons were symbolic and black & white.

Several designers suggest approaches to creating glanceable visuals. Kosslyn argues that glanceability is increased when design elements used to convey information are quickly identified and easily related to the content they represent [13]. Mullet's *immediacy* principle for effective renditions [19] (*i.e.*, a rendition's ability to quickly convey intended meaning) is equivalent to glanceability. To accomplish immediacy, Mullet recommends designs that are "reduced to the essence of the underlying sign through a process of simplification and abstraction." To better understand the techniques used by designers to accomplish glanceability in visual interfaces, we interviewed designers.

INTERVIEWS WITH DESIGNERS ON GLANCEABILITY

To inform our studies, we interviewed 9 designers about how to design for glanceability: 3 professional designers, 2 professors of design, and 4 graduate students with design experience. We showed interviewees (directly, not by glancing) 30 email peripheral displays we created that used different visual variables (*e.g.*, shape, color, size, *etc.*) to represent email sender groups (*e.g.*, coworkers, family, *etc.*). We derived the list of visual variables from prior work on design variables [1, 9] and by examining many existing peripheral displays [16]. We asked participants which variables best conveyed email sender group information. Participants did not like the email renditions based on a single variable, such as the orientation of a single black line or the color of a square. They felt that complex visuals could be valuable for conveying information. This motivated us to study a wide range of renditions (a few variables, simple and complex pictures, and text) rather than drawing from a single design variable with a large range of values.

Designers also shared their opinions on the general principles that contribute to glanceable design. The common themes were as follows: (1) visual renditions should logically *match viewer expectations* (*e.g.*, strong match: red=hot; weak match: square=hot), since this will reduce learning effort; (2) renditions should use *abstraction*, simplifying information to its essential qualities to reduce cognitive load for interpretation; (3) visuals are easier to distinguish when they are *distinct*; and (4) *consistency should be maintained* among design elements (*e.g.*, use a similar design language for renditions of email senders).

Most of these themes are common interface design principles, as evidenced by their application in many existing peripheral displays [15, 20, 21, 24, 27]. However, it is not clear how renditions should "use abstraction" to accomplish glanceability, a question we probe in our work.

RENDITION EXPLORATION AND SELECTION

The goal of our design exploration was to create rendition sets that demonstrate major tradeoffs in the design of glanceable displays: *complexity*, *symbolism*, and *set size* (see Figure 1). We study two types of high-symbolism renditions: pictorial renditions that map directly to meaning; and text, which is not pictorial but highly learned. We ex-

plored two other characteristics of renditions: subject (people, objects, situations/ideas, shapes, text) and technique (photographs, complex illustrations, simple illustrations, text). Subject was based on a classic semiotics classification of signs [18]: icon, index, symbol, and metasybol.

Sets of renditions for email *sender groups* were constructed for up to seven sender groups found in [8] (with slight changes): *supervisor, coworker, admin, subordinate, friend, family member, and stranger* (see Table 2 for examples).

Multiple rendition sets were initially used in order to explore the tradeoff continuums. Renditions were next pruned to one set for each combination of simple vs. complex, low vs. high-symbolism, subject, and technique used that made sense (e.g., text and abstract art were not varied by subject). The result was 32 sets (5 renditions per set, 160 total).

To narrow down the rendition sets for our studies, we pre-tested the initial 32 rendition sets with 10 participants (affiliated with a university) using surveys and a pilot of the Unlearned Study (identifying unlearned renditions on a peripheral display while doing a manual tracking task). Based on our results, we chose the 14 sets that led to the highest identification accuracy and user ratings among others positioned similarly in the tradeoff space: 6 simple, 6 complex, and 2 text; 3 low-symbolism (16-29% average accuracy per set), 3 medium (42-77%), 6 high (84-100%), and 2 text. Figure 1 shows the renditions of the sender group *coworker* from these 14 sets.

Despite a relatively small number of participants, results from the pretest clarified the most important design characteristics for the study. First, we found a significant main effect of *symbolism* (low v. med v. high v. text) for all metrics except primary task error time. For *complexity* (simple v. complex v. text) we found main effects for glance times and all qualitative metrics except adoption. Differences among renditions with respect to *subject* and *technique* were largely driven by variations in shape and text that were better categorized by our symbolism factor, so we removed them from consideration in experiments. Finally, a factor analysis of qualitative ratings (users rated renditions on interpretability, memorability, perceptibility, aesthetics, and likelihood to adopt) resulted in three factors driven by *interpretability, aesthetics, and likelihood to adopt*. These were the final ratings used in the remaining experiments.

DUAL-TASK LABORATORY STUDIES

We conducted two studies exploring three tradeoffs in designing glanceable visuals: complexity, symbolism, and the number of renditions in a set. In the *Unlearned Study*, we tested *unlearned* renditions, exploring how symbolism affects user performance and opinions. In the *Learned Study*, we tested different numbers of *learned* renditions, exploring how symbolism, complexity, and the number of renditions to be memorized affects user performance and opinions.

Experimental Setup and Tasks

We ran both studies on a dual-monitor system. Both screens were set to 1280 x 1024 resolution. Users provided input with a keyboard and infrared mouse. The primary task was on the focal monitor (placed directly in front of the user). The email peripheral display was on the second monitor (placed to the right of the focal monitor). The primary task, “the Circle Game,” was a classic continuous manual tracking task [26], where users attempted to keep a blue dot inside a randomly moving red circle. *Error time* measured the total amount of time the blue dot was outside the red circle. The software automatically adjusted the diameter of the red circle to keep the user’s error time per minute below 1.9 seconds. Automatic adjustment compensated for differential user abilities and fatigue during testing, but did not affect error time while a rendition was visible.



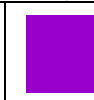
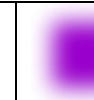









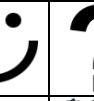







A new rendition abruptly appeared on the peripheral display every 10-15 seconds. Non-text renditions were 1¼ inches on their longest side; text stimuli were rendered using Arial 18 pt. bold font. Users identified a rendition by pressing a keyboard key. The keyboard was covered with a paper mask to expose only keys needed for the study. For set size 3, 3 keys were visible. For set size 7 (Learned Study only), the same 3 keys (each mapped to one sender group for a single-press and a second sender group for a double-press) plus the space bar were used. To control for double-press errors, we removed data points when the user incorrectly identified the rendition. Training enabled users to make these key-to-sender group mappings fluently.

Users were shown a point total at the end of the test, encouraging them to get the highest point total possible. Points were accrued for error-free Circle Game performance and for faster identifications of renditions on the peripheral display. To discourage errors, when the user identified a rendition incorrectly, both monitors blacked-out for 2 seconds (pausing the tasks so errors did not accrue).

Participants

We recruited students and staff from two large universities and community members who used the computer and email regularly. Twenty-six users completed the Unlearned Study, (14 female); ages ranged from 19 to 41, with an average of

Table 2: Design sets for the Learned Study. Smaller than actual size used in study.

| | | | | | | | |
|------------------------|---|---|---|---|---|---|--|
| low-symbolic, simple |  |  |  |  |  |  |  |
| high-symbolic, simple |  |  |  |  |  |  |  |
| high-symbolic, complex |  |  |  |  |  |  |  |
| text | supervisor | coworker | admin | subordinate | family | friend | stranger |

25; they had used a computer for an average of 13 years; and all but three checked email several times a day. For the Learned study, 49 different users (31 female) were split into two groups. Ages ranged from 18 to 53, with an average of 24; they had used a computer for an average of 12 years; and all but four checked email several times a day. Users received \$10 (USD) for their time.

Measures

Dependent measures collected during the course of the study included peripheral processing time (time to first look at the display), glance time (time from the first glance to rendition identification), number of glances, primary task error time (duration of Circle Game errors while a rendition was visible), correct rendition identification rates, user qualitative ratings, and overall rendition preferences.

Other than glance data and qualitative data, all measures were automatically collected via logging tools. Glance time was measured by videotaping the user's face and the peripheral display. Video recordings were hand-coded (using software to step through frame-by-frame) to measure the time from the user's first glance at a rendition to the user's key press identifying the rendition. The number of glances at a rendition was also counted by video coding.

UNLEARNED STUDY

Unlearned renditions must be interpretable to be useful. The goal of the Unlearned Study was to help us understand interpretability: (1) what characteristics made renditions quicker and easier to interpret; (2) how accurately renditions were interpreted (indicating their level of symbolism); and (3) to what degree interpreting new renditions distracted from the primary task. High correct identification percentages, low glance times, and primary task error times would indicate easy-to-interpret renditions.

Understanding interpretability is important since not all renditions are learned. Many information sets are too large to easily memorize. Also, displays may need to convey rare or new information. Further, interpretability may help learning since it can provide meaning until users have operationalized use of a display. Finally, it may increase adoption since an easy-to-learn display may reduce frustration.

Hypotheses

We hypothesized that (1) higher-symbolism renditions would be faster and more accurate to interpret than lower-symbolism renditions and (2) text would be fastest and most accurate to interpret (based on [5] in Table 1).

Design

Though symbolism is a continuum, we grouped our renditions into 4 levels for analysis. The study design comprised a 4-level, single factor (symbolism: low v. medium v. high v. text) within-subjects design. The assignment of renditions to the 4 symbolism levels was based on ad-hoc design knowledge and subsequently adjusted according to correct identification rates of renditions. Identification rates of

symbols within each symbolism level did not differ significantly, whereas the identification rates of symbols at different levels of symbolism did differ significantly (by LSD post hoc test). Low-symbolism (correct identification 30-39% of the time) included celebrity photo, abstract art, and colored square (see Figure 1). Medium (52-54%) included one face, two faces, and beach picture. High (74-92%) included the remaining renditions and the text initial. Text (97%) included colored and black words. The presentation order of all renditions was counterbalanced across users.

Method

Users were run individually with an experimenter present. After introductions, the user started with a 5-minute practice trial (using renditions not selected from pretesting). Then users played the Circle Game while classifying 42 renditions they had never seen into three categories: co-worker, family, and stranger. Renditions were drawn from the 14 sets chosen after pretesting. Each rendition was used once. After the dual tasks, users completed a survey, rating each rendition set on a 5-point scale for interpretability, aesthetics, and how likely the user was to use the renditions in their email program (adoption). Last, we debriefed and paid users. Total session time was 30-40 minutes.

UNLEARNED STUDY RESULTS

We used a single-factor (symbolism) ANOVA with four levels to analyze the data presented throughout this section, unless otherwise stated. Where a significant main effect was found, LSD post hoc tests (alpha set to .05) were used to determine which means differed significantly.

For qualitative results, we used data from all 26 users. We used dual-task data from 23 users – data from two users were removed due to English language and manual dexterity issues, and data from one user was missing due to logging errors. For video coding data (glance and peripheral processing times), two additional users were missing due to camera issues. As is standard practice, all time data were transformed into log times to render the distributions normal to deal with skew in the original data (except primary task error time, because the distribution was normal).

Glance Time

We removed outliers ≥ 4 standard deviations from the mean (4 of 857 cases). There was a significant main effect for symbolism, $F(3,60)=4.5$, $p=.006$. Text and high-symbolism led to significantly better times than low- and medium-symbolism. See Figure 3a for a graph.

Peripheral Processing Time

Glance times (mean=34 ms) were a small part of total reaction times (the time from when a rendition appeared to when the user identified it, mean=1931 ms). The additional time can be attributed to what we call *peripheral processing time*. We hypothesize that users began identifying a rendition before looking at it. If they failed to identify the rendition peripherally, they resorted to glancing (a hypothesis

supported by results from the second study in which some renditions were processed entirely without glancing). This is consistent with theories of attention that describe a pre-cognitive phase in which we attempt to recognize primitive features of a rendition before attempting focused recognition [23]. Reaction time is assumed to comprise *attention capture + peripheral processing time + glance time + key press*. The first component, attention capture *via* abrupt onset, is a constant perceptual effect, largely unaffected by rendition properties, that takes on the order of 100 ms [4]. For all further computations and discussion here, peripheral processing time will include attention capture.

From total reaction time data, we removed outliers ≥ 4 standard deviations from the mean (13 of 1005 cases). To measure peripheral processing time, we subtracted glance time from reaction time in cases where both data points were available. We found a significant main effect for symbolism, $F(3,60) = 41.6, p < .001$. Except for a non-significant difference between low- and medium-symbolism sets, renditions that were higher-symbolism led to significantly faster peripheral processing time. See Figure 3b.

Primary Task Error Time

We found a significant main effect for the influence of symbolism, $F(3,66) = 3.5, p = .019$. Text distracted from the primary task significantly less than the other three symbolism levels (by 153 ms), which were not significantly different from each other. See Figure 3a for a graph.

Qualitative Ratings and Preferences

Users rated interpretability, aesthetics, and adoption. For all three, there was a significant main effect for symbolism (interpret: $F(3,78) = 144.8, p < .001$; aesthetics: $F(3,78) = 3.9, p = .012$; adoption: $F(3,78) = 23.4, p < .001$). See Figure 3c.

Interpretation and adoption ratings increased with symbolism, while aesthetics was unrelated to the other metrics (all metrics in Figure 3 are correlated). Users rated text significantly higher than all other renditions for interpretability. High-symbolism were rated significantly higher than others for aesthetics. For adoption, text and high-symbolism were rated highest (not differing significantly from each other,

but significantly higher than the others). Low- and medium-symbolism received similar ratings for all metrics

When asked to choose the two rendition sets they liked most, 16 out of 26 users chose the cartoon people (complex + high-symbolism from Table 2), 11 chose the black & white pictures (simple + high-symbolism from Table 2), 8 chose colored text, and 8 chose celebrity photos. When asked which rendition set they liked the least, 13 out of 26 users chose abstract art and 10 chose the people photos. User preferences were somewhat like interpretability ratings, but text and celebrity photos behaved differently. Text was less favored than high-symbolism renditions despite higher ratings, and the celebrity photo was surprisingly favored by 8 people despite very low ratings for low-symbolism renditions. This indicates that user preferences may have been influenced by study task performance somewhat, but that other factors also impacted them.

Summary

Our results confirm our hypotheses: (1) higher-symbolism renditions were faster and more accurate to interpret than lower-symbolism (they also got higher qualitative ratings); and (2) text was the fastest and most accurate to interpret. Reaction times were composed mostly of peripheral processing time, while glance times were only a small part.

LEARNED STUDY

The purpose of the Learned Study was to help us understand what characteristics of visual renditions make them faster and easier to perceive and interpret when *different numbers* of them are *learned*. Does symbolism matter? When fewer renditions are used, do simple renditions enable better performance than complex? When more renditions are used, what characteristics improve performance?

To answer these questions, we compared the four sets shown in Table 2, which produced the best performance in the Unlearned Study. The three high-symbolism renditions (black & white pictures, cartoon people, text) enabled us to compare three levels of complexity (simple v. complex v. text). The two simple renditions (colored squares, black & white pictures) let us compare low- v. high-symbolism.

Half of the users memorized and were tested on three renditions from each set; the other half memorized and were tested on all seven renditions from each set.

Hypotheses

We expected high accuracy for all learned renditions. Differences were predicted for reaction time, however. For small set sizes, we hypothesized that simple renditions would be fastest to identify (based on [17]). For large set sizes, we hypothesized that

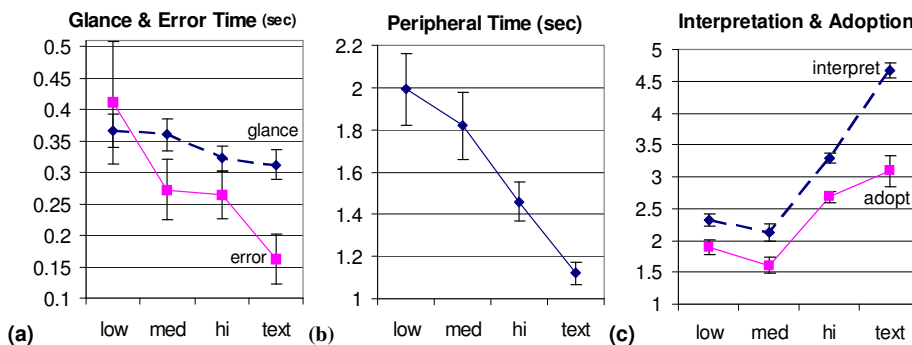


Figure 3: Unlearned Study results by symbolism (low, medium, high, text). All metrics improve as symbolism increases. A factor analysis proves these metrics are all correlated to the same phenomena, which we call *glanceability*. (Error bars depict standard error.)

complex + high-symbolism would be the fastest to identify.

Regarding text, we hypothesized that (1) text would lead to equivalent identification time for small and large set size since the time to read a word is fairly constant; and (2) for large set sizes, text would be faster to identify than low-symbolism renditions (based on [5]).

Design

Subjects in two groups defined by learned set size, saw high-symbolism stimuli varying in complexity and low-complexity stimuli varying in symbolism. Hence the study comprised two sub-designs. The first was a 3 X 2 design with factors of complexity (low v. high v. text – all were high-symbolism) and set size (3 v. 7). The second design was a 2 X 2 design with factors of symbolism (low v. high – both were low-complexity) and set size (3 v. 7). One rendition overlapped in the study designs: black & white pictures were both low-complexity and high-symbolism. For both sub-designs, the first factor was within-subjects and the second between-subjects. We counterbalanced the presentation order of all renditions across users.

Method

Users were run individually with an experimenter present. After introductions, the user completed a standard shape memory test [10]. To control for differences in ability to remember shapes, we split users into our two between-subjects groups so that each group had similar distributions of shape memory test scores. Next, users completed a 5-minute practice trial (using renditions not selected from the Unlearned Study) to get accustomed to the dual tasks.

Users began the study with a training session in which they practiced identifying one set of renditions (either 3 or 7 renditions per set). Training ended once the user correctly identified each rendition five times in a row. Users then played the Circle Game while identifying each rendition in the set three times each. Users repeated the training and dual-task trial four times: once for each set of renditions. Then they completed a survey, rating each rendition set on a 5-point scale for identifiability, aesthetics, and how likely they were to adopt the renditions in their email program. At the end of the session, we debriefed and paid users. Total session time was about 30 minutes for sets of 3 and 60 minutes for sets of 7.

LEARNED STUDY RESULTS

In keeping with our design, we performed two ANOVAs. The first was a 3 (complexity: simple v. complex v. text, all high-symbolism) X 2 (set size: 3 v. 7) ANOVA and the second was a 2 (symbolism: low v. high, both simple) X 2 (set size: 3 v. 7). Set size varied between subjects, and complexity and symbolism varied within subjects. Where a significant main effect was found, LSD post hoc tests (alpha set to .05) were used to determine which means differed significantly. For all time data we removed all data points when the user incorrectly identified the rendition.

Users correctly identified renditions most of the time for both set sizes (averages ranged from 93% to 98%), with no significant differences. These results indicate that the training was sufficient for users to memorize renditions.

Glance Times and Number of Glances

Simple renditions enabled users to avoid glancing by using peripheral vision. Out of 2772 total viewings, users used peripheral vision alone in 1074 cases (38.7%). Set size of 3 afforded this better than set size of 7. For set size 3, 90% of colored square views were peripheral alone, 82% for black & white pictures, 38% for cartoon people, and 5% for text. For set size 7, 76% of colored square views were peripheral only, 51% for black & white pictures, 4% for cartoon people, and 0% for text.

Due to unbalanced data resulting from little use of foveal vision for simple renditions, glance time data were uninterpretable. Instead, we relied on total reaction times to describe the time required to identify renditions.

Reaction Times and Primary Task Error Times

Reaction time (RT) is the time from when the rendition appears in the periphery to when the user presses a key to identify it. We removed outliers ≥ 4 standard deviations from the mean (11 of 2944 cases) and transformed all task times into log times (a standard practice for time data, to account for skew). For high-symbolism renditions, there was a significant effect for complexity, $F(2,90)=57.4$, $p<.001$, and an interaction effect between complexity and set size, $F(2,90)=12.5$, $p<.001$. For set size 3, text was slowest, the simple black & white pictures were fastest, with the complex cartoons in the middle, $F(2,46)=57.4$, $p<.001$. For set size 7, text and cartoons were slowest (not significantly different from each other), and the black & white pictures were fastest, $F(2,44)=7.8$, $p=.001$. Effects of complexity were larger for set size 3 than for set size 7. When comparing the low- and high-symbolism, simple renditions (colored squares v. black & white pictures), there were no significant differences. See Figure 4a for a graph.

For primary task error time data, we only observed a significant effect for complexity, $F(2,90)=3.7$, $p=.03$. Simple and complex renditions led to significantly less error time than text, but the difference was very small (26 ms).

Qualitative Ratings and Preferences

Simple v. Complex (High-Symbolism). When black & white pictures (simple) were compared to cartoons (complex), there was a significant interaction between set size and complexity for identification ($F(1,47)=4.8$, $p=.033$) and adoption ratings ($F(1,47)=6.0$, $p=.018$). As Figure 4b shows, simple renditions were rated higher for set size 3 (solid line) (identification: $F(1,24)=4.7$, $p=.041$; adoption: $F(1,24)=8.2$, $p=.008$), while they were not significantly different for set size 7 (dotted line). There was also a significant effect of complexity for aesthetics ratings, $F(2,94)=34.4$, $p<.001$. Complex were rated highest for aesthetics (3.7), then simple (3.2), then text (1.9).

Low- v. High-Symbolism (Simple). There was a main effect of symbolism for identification ($F(1,47)=9.5$, $p=.003$) and adoption ($F(1,47)=5.2$, $p=.028$). High-symbolism renditions (black & white pictures) were rated higher than low-symbolism (colored squares), for both set sizes.

Preferences. When asked to choose a favorite rendition set, most users (20 out of 49) chose the black & white pictures and the cartoons were second (13 out of 49). When asked which rendition set they liked least, 20 out of 49 users chose text and the same number chose the squares. Comparing preferences across groups, the results mirrored these overall results, except users who saw set size 7 favored cartoons and black & white pictures equally (8 out of 24 each).

Summary

Our results confirm that all renditions were identified with high accuracy. We confirm our hypothesis in the small set size case: simple renditions were fastest to identify. We reject our hypothesis for the large set size: complex + high-symbolism renditions were not fastest to identify. For text, we reject both hypotheses: (1) identification with text was slower for the large set size than for the small; and (2) text led to slower reaction times than all other renditions.

DISCUSSION

Here we discuss what we learned from both studies about evaluating for glanceability, the impact of glanceable design on multitasking, the importance of user opinions, and unexpected results. It should be noted that our measured results are a proxy for true multitasking performance. A study of real-world use is required to confirm them.

Overall, our results show that *high-symbolism* only improves performance for unlearned renditions. As shown in Figure 4, when few renditions are learned (solid line), *simple* renditions improve performance and qualitative ratings. This closely mirrors prior single-task empirical results [17]. When many renditions are learned (dotted line in Figure 4), simplicity also improves performance, but not qualitative ratings. From set size 3 to 7, users' ratings of complex rose and of simple fell enough to make the differences between then insignificant. Further, though low- and high-symbolism renditions led to similar performance, low-

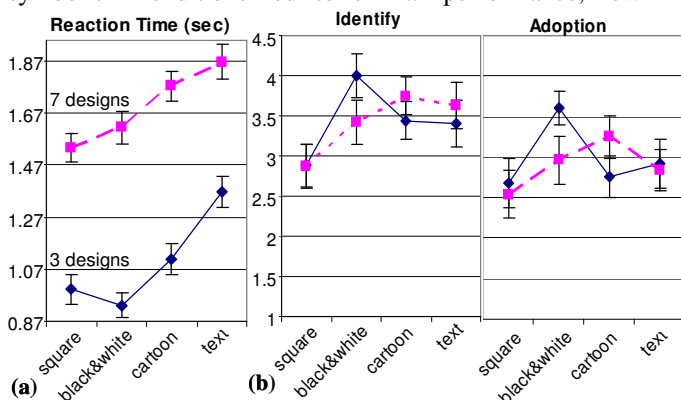


Figure 4: Learned: mean (a) reaction times, (b) user ratings. Solid line for set size 3; dotted line for set size 7. (Error bars show std. error.)

symbolism resulted in significantly lower user ratings. As we discuss below, we recommend that designers take these qualitative ratings seriously since performance differences are small (though similar in magnitude to past studies [17]).

Evaluation of Glanceable Visuals

To examine relationships among our metrics, we computed a factor analysis of Unlearned Study data based on correlations between metrics, using rendition set as the unit of observation (the Learned Study included only 4 rendition sets – not enough to compute correlations). The analysis resulted in two factors that explained 81.3% of the variance. The first factor (55.5% of variance) had large loadings from interpretation (.932) and adoption ratings (.892), correct identifications (.816), peripheral times (-.774), error times (-.659), and glance times (-.645). The second factor (25.5% of variance) was driven by aesthetics (.795).

The factor analysis shows *glanceability* is a measurable phenomenon that encompasses all of our metrics except for aesthetics (indicating that aesthetics is not equivalent to likeability). This means that designers can evaluate unlearned glanceable visuals with qualitative metrics (after users use the interface in a divided-attention setting) instead of gathering empirical performance data. Results for both learned and unlearned renditions for set size 3 show that qualitative ratings go up as glance, peripheral, and error times go down. Set size 7 shows the opposite trend: as reaction and error times increase, qualitative ratings also go up. At small set sizes, apparently, qualitative responses reflect ease of processing; at larger set sizes they reflect an assessment of underlying processes that are relatively slow but contribute to a positive impression of performance. Overall, these results suggests that any of our metrics can be used to measure glanceability for smaller set sizes and that qualitative ratings are possibly most important for larger set sizes, since performance differences were not big enough to outweigh user opinions.

Impact on Multitasking Efficiency is Minimal

Primary task error and peripheral processing time data are indicative of a user's ability to maintain task flow while quickly checking a secondary display. Our results show that more glanceable visuals lead to better primary task performance and reaction times. However, improvements are small – 248 ms less error time and 872 ms faster reactions for the best unlearned renditions (text) over the worst (low-symbolism). Information workers have been found to receive 49 emails a day on average [25]. Per day, these improvements add up to only 12.2 seconds less primary task distraction and 42.7 seconds less time reacting to new emails. The time saved is even smaller between different learned renditions. These time improvements are not large enough to make a real impact on primary task performance times in a multitasking situation. (Important to note is the fact that we selected renditions that were most successful in pretesting, so our study was not set up to find large differences in qualitative or performance metrics.)

However, extra effort may be more important than time saved, a theory that is supported by our finding that users preferred a rendition set they felt was easier to use – either because it was faster or because it made the task “easier” or less cognitively demanding. For example, though the performance benefits of text over low-symbolism were small (when unlearned), users indicated they might adopt text (rated 3.1 on a 5-point scale, on average), but they would likely not adopt low-symbolism renditions (1.9). When learned, user preferences dropped for the black & white pictures from set size 3 to 7, while preferences rose for cartoon people. A similar trend appears in the adoption ratings (Figure 4b): black & white pictures were rated higher than the cartoons for set size 3, but the two are not significantly different for set size 7. Users explained that with more renditions to remember, despite slightly better performance with the black & white pictures, it was “not as easy to remember what icon corresponds with which group of people,” while the cartoons were “easy to identify and relate with as they pictorially represent the categories.” These results imply that small differences in performance or perceived difficulty should be taken seriously, since they can affect the effort expended and users’ opinions of a display.

Glanceable Design: User Preference is Important

Though we found significant results that show certain characteristics lead to better multitasking performance, our main recommendation is to listen to what users want. Our results demonstrated that (1) qualitative ratings are correlated with performance for unlearned renditions and are indicative of perceived effort for all conditions, and (2) qualitative differences may be most important since performance differences were not big enough to outweigh user opinions.

When users were unfamiliar with renditions, qualitative and performance results were parallel – high-symbolism and text renditions were best for both. When users learned the renditions, low- and high-symbolism renditions led to similar performance, but low-symbolism resulted in much lower ratings. Overall, users preferred a rendition set they felt was easier to use – either because it was faster or because it was “easiest to remember,” reducing cognitive load.

Surprises

Portions of our results surprised us: the simplest renditions were not favored; complex renditions were popular; peripheral vision alone was used often with learned renditions; and the use of certain photos had a negative impact.

The simplest renditions were not favored. For a small set size of learned renditions, we hypothesized that the simplest rendition with highly distinguishable features would lead to the best performance on all metrics. Colored squares were very distinguishable (as we know from heavy use of peripheral vision alone to identify them), used fewer design variables than the black & white pictures, and the two led to similar performance. However, users preferred the black & white pictures. In fact, though the squares led to slightly

better performance for larger set sizes (another surprise), users still rated them poorly. Users explained: “It is hard to remember what color represents what,” despite both fast and accurate performance. In essence, users favored mnemonic value as a deciding factor over ease of perception. Low-symbolism renditions may have a harder time gaining approval, despite greater visual simplicity.

Complex renditions led to good user opinions for large set sizes. We hypothesized that the simple, high-symbolism black & white pictures would be favored for larger set sizes. We were very surprised that the complex drawings of people were also popular, especially since they were not highly discriminable compared to the simpler renditions. As presented above, complex renditions did *not* lead to better performance than simple, but qualitative ratings and preferences rose for complex and fell for simple from set size 3 to 7. Users explained that complex renditions were easier to remember and identify, and “they are the most pleasing to look at and not as ‘boring’ as the others.” Our results indicate that extra detail may make renditions more attractive and identification easier within large sets.

Heavy use of peripheral vision and color. It is surprising that peripheral vision alone was so prevalently used (33% of the time), and that color supported this (80% of colored square views were with peripheral vision alone, and color was the primary visual difference between renditions in the set). Peripheral vision use was beneficial because the simple renditions that afforded peripheral identification also led to less primary task distraction. It is surprising that colored squares were most effective in the periphery, since color discrimination is poor in peripheral vision, and prior research recommends that color not be used in the periphery [6]. The reason our second-monitor position of the peripheral display was effective is that people have full color use up to 30° from the central fixation area (exact dimensions vary in studies, see [12] for a survey). The peripheral display was about 30° from the primary screen, within color vision range. This indicates that color may be effective for peripheral displays on dual-monitor systems since they are not very far in the periphery, particularly sizeable areas of distinct colors like our renditions. In general, our results show that if use of peripheral vision is important, simple renditions (high or low-symbolism) are best.

Photos of people led to poor performance and ratings. Of all the complex, high-symbolism renditions in the first experiment, we were surprised that *photos of people* (upper-right, Figure 1) led to very low performance and qualitative ratings. We theorize that models in the photos who were not the user’s coworker, family, *etc.*, may have interfered with the user’s ability to associate the photos with sender groups. There was no interference in associating *drawings of people* or *celebrity photos* with these groups, perhaps because these renditions were more removed from actual likenesses of social relations. Also, the photos we used were contained within a square so their shape was not as distinctive as the illustrations’. Similarly, many colors and shapes (people)

are used in the photos so more effort may have been needed to distinguish them. Distinctiveness has been shown to be an important issue for designers [5, 17]. We believe that interference effects are also important to consider when choosing photographs to represent certain information, but more work is needed to confirm this.

CONCLUSION AND FUTURE WORK

Our goal was to inform the design of glanceable visuals, which better enable people to monitor secondary tasks while multitasking. We conducted two studies to understand what rendition characteristics are most glanceable, learned and unlearned, when dividing attention. Our studies explored three major tradeoffs in designing glanceable visuals: complexity, symbolism, and rendition set size. Our main contributions are best practices for the design and evaluation of glanceable displays based on empirical and qualitative results. We found that *high-symbolism* only improved performance for unlearned renditions. When few renditions were learned, *simple* renditions improved performance and qualitative ratings. When many renditions were learned, *simplicity* also improved performance, but not qualitative ratings. From set size 3 to 7, users' ratings of complex renditions rose and of simple fell enough to make differences between them insignificant. Also, though low- and high-symbolism renditions led to similar performance, low-symbolism resulted in much lower user ratings. We argue that since performance differences are small (though similar in magnitude to past studies [17]), user opinions may be more important. These contributions provide design and evaluation knowledge that will enable designers to create better peripheral displays to support multitasking.

In future work, we are interested in applying our design and evaluation findings to peripheral displays for email and task management (e.g., Scalable Fabric [15]), with which we could evaluate the impact of glanceable visuals in the field.

REFERENCES

- Bertin, J. *Semiology of graphics, diagrams, networks, maps*. University of Wisconsin Press, 1983.
- Booher, H.R. Relative comprehensibility of pictorial information and printed words in proceduralized instructions. *Human Factors*, 17: 266-277, 1975.
- Campbell, J.L., Richman, J.B., Carney, C. & Lee, J.D. In-vehicle display icons and other information elements: Volume I: Guidelines, Federal Highway Administration, 2004.
- Card, S., Moran, T. & Newell, A. *The psychology of human-computer interaction*. Lawrence Erlbaum, 1983.
- Carney, C., Campbell, J.L. & Mitchell, E.A. In-vehicle display icons and other information elements: Literature review, Federal Highway Administration, 1998.
- Christ, R.E. Research for evaluating visual display codes: An emphasis on colour coding. In Easterby, R. & Zwaga, H. eds. *Information design: The design and evaluation of signs and printed materials*, John Wiley & Sons, 1984, 209-228.
- Czerwinski, M., Horvitz, E. & Wilhite, S., A diary study of task switching and interruptions. *Proc. of CHI'04*, 175-182.
- Dabbish, L., Kraut, R.E., Fussell, S. & Kiesler, S., Understanding email use: Predicting action on a message. *Proc. of CHI'05*, 691-700.
- Dondis, D.A. A primer of visual literacy. MIT Press, 1973.
- Ekstrom, R., French, J., Harman, H. & Dermen, D. *Manual for kit of factor referenced cognitive tests*, 1976.
- Gonzalez, V.M. & Mark, G., 'Constant, constant, multi-tasking craziness': Managing multiple working spheres. *Proc. of CHI'04*, 113-120.
- Johnson, M.A. Color vision in the peripheral retina. *Am. J. of Optometry and Physiological Optics*, 63(2): 97-103, 1986.
- Kosslyn, S.M. *Principles of graph design*. W.H. Freeman & Co., 1994.
- Lee, J., Forlizzi, J. & Hudson, S.E. Iterative design of MOVE: A situationally appropriate vehicle navigation system. *International J. of Human-Computer Studies*: In press, 2006.
- Matthews, T., Czerwinski, M., Robertson, G. & Tan, D., Clipping Lists and Change Borders: Improving multitasking efficiency with peripheral information design. *Proc. of CHI'06*, 989-998.
- Matthews, T., Forlizzi, J. & Rohrbach, S. Designing glanceable peripheral displays, U.C. Berkeley, 2005.
- McDougall, S., de Bruijn, O. & Curry, M. Exploring the effects of icon characteristics: Concreteness, complexity and distinctiveness. *The J. of Experimental Psychology: Applied*, 6(4): 291-306, 2000.
- Meggs, P. *Type & image, the language of graphic design*. Van Nostrand Reinhold, 1992.
- Mullet, K. & Sano, D. *Designing visual interfaces: Communication oriented techniques*. Sunsoft Press, 1995.
- Plaue, C., Miller, T. & Stasko, J., Is a picture worth a thousand words? An evaluation of information awareness displays. *Proc. of GI'04*, 117-126.
- Pousman, Z. & Stasko, J.T., A taxonomy of ambient information systems: Four patterns of design. *Proc. of AVI'06*, 67-74.
- Somervell, J., McCrickard, D.S., North, C. & Shukla, M., An evaluation of information visualization in attention-limited environments. *Proc. of VISSYM'02*.
- Treisman, A. & Gelade, G. A feature integration theory of attention. *Cognitive Psychology*, 12: 97-136, 1980.
- Van Dantzich, M., Robbins, D., Horvitz, E. & Czerwinski, M., Scope: Providing awareness of multiple notifications at a glance. *Proc. of AVI'02*.
- Whittaker, S. & Sidner, C., Email overload: Exploring personal information management of email. *Proc. of CHI'96*, 276-283.
- Wickens, C.D. The effects of control dynamics on performance. In Kaufman, K.B.A. (ed.), *Handbook of perception and performance*, John Wiley & Sons Ltd., 1986, 39-60.
- Zhang, L., Tu, N. & Vronay, D., Info-lotus: A peripheral visualization for email notification. *CHI'05 extended abstracts*, 1901-1904.
- Zwaga, H. & Easterby, R.S. Developing effective symbols for public information. In Easterby, R. & Zwaga, H. (eds.). *Information design: The design and evaluation of signs and printed material*, J. Wiley & Sons, 1984, 277-297.