Curtis P. Arledge. Filled-in vs. Outline Icons: The Impact of Icon Style on Usability. A Master's Paper for the M.S. in I.S. degree. April, 2014. 68 pages. Advisor: Robert Capra

This study sought to determine whether single-color, "flat" icons would be more quickly and accurately selected by users when presented in either a filled-in or outline style. An application was developed that allowed participants to take a test measuring their speed and accuracy in selecting prompted icons from an array of distractors. The test was made available on the web and was completed by 1,260 participants. Averaged across the 20 unique icon forms used in the test, the outline style led to slightly longer task times, but only when icons were displayed in white against a black background. For individual icons, the effects of icon style were inconsistent and, except for a few exceptions, quite small. This study concluded that one icon style is not objectively better than the other.

Headings:

icons

icon design

iconography

user interface design

visual design

flat design

iOS 7

Android

Windows Phone

FILLED-IN VS. OUTLINE ICONS: THE IMPACT OF ICON STYLE ON USABILITY

by
Curtis P. Arledge

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2014

Approved by

_____

Robert Capra

INTRODUCTION

In the past few years, a visual aesthetic known as "flat" design has grown in popularity

and exposure in user interfaces on the web and in consumer software. The aesthetic shift

towards simple, minimal forms and fields of uniform color that characterize flat design

has corresponded with the decline of a highly textured, photorealistic design style that

had hitherto dominated the aesthetic landscape of popular user interfaces. This shift is

illustrated most clearly by the overhauled visual design systems introduced in Apple's

iOS7 mobile operating system and Microsoft's Windows Phone and Windows 8

operating systems.

The rise of flat design has produced a host of new considerations for interface designers.

Arguably, the minimal style of flat design offers limited expressive potential for

iconography and other interface elements. One consideration for designers is how best to

indicate an active state for the flat, single-color icons frequently used as buttons and

navigation in flat interfaces. Apple recommends the following design solution in its iOS

Human Interface Guidelines:

> "If you're designing a custom tab bar icon, you should provide two
> versions—one for the unselected appearance and one for the
> selected appearance. The selected appearance is often a filled-in
> version of the unselected appearance, but some designs call for
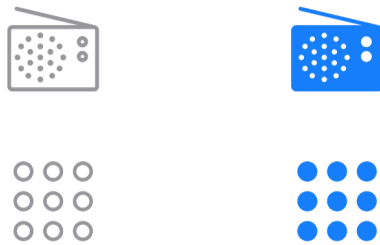> variations on this approach." (Apple, 2013)

Figure 1. From Apple's iOS Human Interface Guidelines, an example of outline-style unselected icons (left column) and filled-in selected icons (right column) (Apple, 2013)

In response to Apple's design guidelines, software designer Aubrey Johnson argued in a post on the blogging site Medium (Johnson, 2013) that icons with the more detail-rich outline style take longer for users to mentally process and should therefore be used sparingly in interfaces. Responses to Johnson's article appeared in the form of blog posts, discussion threads, and polls, as designers weighed in on the subject from various perspectives (Kholmatova, 2013; Solomon, 2013; Wong, 2013; Wroblewski, 2013). The goal of this study was to gather empirical evidence to determine whether flat, single-color icons with a filled-in appearance really are more usable than those with an outline appearance, where usability is defined in terms of the speed and accuracy with which users are able to recognize and select icons. An experiment was designed in the form of a self-directed test, freely available on the Web, to isolate the effects of icon style on speed and accuracy of selection.

It is important to note that creating filled-in and outline versions of a common base icon form involves a degree of artistic discretion, and that the distinction between the two styles can be somewhat loose. However, considered strictly, filled-in and outline versions of a common base icon form are not the same icon. Apple's adoption of a two-version approach to iconography provides ample justification to test the hypothesis that outline

icons come with a usability cost. As Apple proliferates this design pattern to millions of iPhone and iPad users around the world, even very small differences in users' recognition speeds can amount to a significant impact on product efficiency when considered in the aggregate.

There are two other reasons why testing this hypothesis is important. One reason is that there is disagreement among interface designers about the value of a two-version approach to iconography. Among large consumer software vendors, Apple is unique in its use of filled-in and outline versions for each icon. In fact, Google's Android design guidelines discourage the use of outlined forms: "Filled shapes are easier to see than thin strokes" (Google, n.d.). This disagreement may become more visible and contentious as more interface designers come to adopt Apple's design style. Indeed, the two-version approach was included as one of 19 "Top UX [User Experience] Predictions for 2014" in the online UX Magazine (UX Magazine Staff, 2013).

There are many ways to indicate an icon's active or selected state without using two different versions of the icon: an active icon could be shown in a different color, at a larger size, underlined, enclosed in some other shape whose visual state changes, or moved to some central position. However, each of these approaches has its drawbacks: color differences may not be perceived by the roughly 5% of people with some form of color-blindness (Sharp, Rogers, & Preece, 2007; UX Magazine Staff, 2013); increasing the size of one icon can disturb the organizational grid; underlining or enclosing icons adds nonessential visual noise to the interface; and repositioning icons is not a usable solution, as users will come to expect icons in a consistent position (Blankenberger & Hahn, 1991; Moyes, 1994). Apple's two-version approach avoids the above limitations

while providing a potential accessibility benefit for people with color-blindness. Any

usability costs to using filled-in and outline versions of icons should be considered

alongside this potential benefit.

LITERATURE REVIEW

Since icons first appeared in a digital interface with the introduction of the Xerox Star

computer in 1981 (Sharp et al., 2007), the use of iconography in user interfaces has been

accompanied by research on various aspects of icon design. Early theoretical work sought

to define and classify icons. In a highly cited 1986 paper, David Gittins defined icons as

"pictographic representations of data or processes within a computer system, which have

been used to replace commands and menus as the means by which the computer supports

a dialogue with the end-user (Gittins, 1986)." Various schemes have been suggested to

classify icons by form (Gittins, 1986), function (Rogers, 1989), graphical genre (i.e. user-

facing vs. utility icons) (Sharp et al., 2007), level of abstractness (Garcia, Badre, &

Stasko, 1994; Lin & Kreifeldt, 1992), and the nature of the semantic representation

(Bernsen, 1994; Gittins, 1986; Rogers, 1989; Sharp et al., 2007). One such semantic

classification scheme categorizes icon representations as *similar* (representing a file with

a picture of a file), *analogical* (representing the Cut command with scissors), or *arbitrary*

(representing the Delete command with ×) (Sharp et al., 2007).

A recurring goal in the literature has been evaluating the usability of icons through

"search and select" experiments, in which the test participant is given a textual or verbal

prompt and asked to select the corresponding icon from an array of distractor icons as

quickly and accurately as possible (Blankenberger & Hahn, 1991; Byrne, 1993; Evers,

Kukulska-Hulme, & Jones, 1999; Garcia et al., 1994; Holloway & Bailey, 1996; Kacmar

 & Carey, 1991; Ling & Van Schaik, 2002; Näsänen, Karlsson, & Ojanpää, 2001; Näsänen, Ojanpää, & Kojo, 2001; Näsänen & Ojanpää, 2003; Ojanpää & Näsänen, 2003).

Visual icon search is a complex task involving both semantic factors and visual factors (Byrne, 1993). Some studies have focused on isolating semantic factors (i.e. icons' ability to communicate their intended meaning), and others have focused primarily on visual factors (i.e. icons' ability to be quickly and easily perceived and distinguished from other icons).

Two outcome variables are almost ubiquitous in both types of studies: users' success in mapping icons to their intended meanings (accuracy, error rate, success rate) and the amount of time it takes users to do so (speed, task time). Both variables are needed to make inferences about the semantic or visual factors of icons, but in general, semantic-focused studies tend to emphasize accuracy, and visual-focused studies tend to emphasize speed. Speed of icon recognition is an important factor of icon design because speed is seen as one of the greatest advantages of icons over text in user interfaces. Although text labels in a user interface are unambiguous, pictorial representations have been shown to result in comparatively faster response times (Blankenberger & Hahn, 1991). Other common metrics in icon research include self-reported preferences and variables related to eye-tracking, such as saccade amplitude and fixation duration.

The present study is focused on visual aspects of icon search, but studies focusing primarily on semantic factors offer instructive methodologies and results. For example, concrete icons (i.e. objects) are identified more accurately than abstract icons (i.e. symbols) (Blankenberger & Hahn, 1991; Garcia et al., 1994; Kholmatova, 2013; Passini,

Strazzari, & Borghi, 2008; Schröder & Ziefle, 2008). Therefore, the icon recognition test used for this study used concrete icons with unambiguous names in order to reduce the effects of semantic factors.

The body of studies dealing primarily with visual aspects of icon search offers a number of methodological insights for the design of search and select tests. For instance, test participants have been shown to map the meaning of an icon to its position in the user interface over time (Blankenberger & Hahn, 1991; Moyes, 1994). The test in the present study randomized the positions of icons in each icon-recognition trial to remove the influence of this behavior. The test designs in Huang (2007) and Huang & Chiu (2007) inspired the use of a 20-icon set, which would be positioned in a circular array for each trial in this study.

Finally, unlike more visually complex icon designs, the ones in this study used only one color and so were able to appear in any color against any color background. Different figure/background color combinations have been shown to affect icon recognition speed (Huang & Chiu, 2007; Ling & Van Schaik, 2002). Because color has been shown to effect recognition speed, and because color and style may also work together to influence recognition speed, figure/background color scheme was included as a second independent variable in this study. To keep the appearance of icons as similar as possible across a diverse range of screens, the two "color" variations used were black icons on a white background and white icons on a black background.

In work closely related to this study, interface designer Alla Kholmatova conducted a test of participants' response times in selecting filled-in and outline icons, which she described in a post on the design blog Boxes and Arrows. The results of Kholmatova's

informal study, presented as part of a larger article about optimizing icons for faster

recognition, served as a timely response to Johnson's (2013) warning against the use of

outline icons.

Kholmatova found no significant difference in participants' speed in selecting icons of

either style. Kholmatova's test used two arrays of iOS7-style icons arranged in a 3×6

grid: one with filled-in icons and one arranged identically with outline icons (as shown in

Figure 2). Six participants performed a sequence of search and select tasks in response to

verbal prompts for each of the 18 icons in both arrays. Half of the participants used the

filled-in array first and then the outline array, and the other half used the outline array

first and then the filled-in array. Tasks were presented in the same sequence for both

arrays, and task times were recorded manually using a digital stopwatch (Kholmatova,

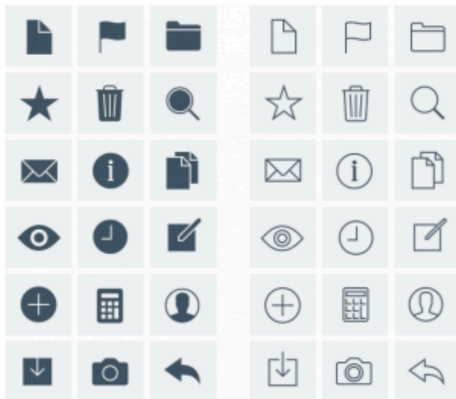2013; Kholmatova, personal communication, April 1, 2014).

Figure 2. The icon arrays used in Kholmatova's experiment (Kholmatova, 2013)

Selection times for each participant's second icon array were more than twice as fast as

those for the first, likely because participants had already encountered all 18 icons in the

same position and sequence during the first set of search tasks. In contrast, the present study attempted to remove the effects of memory and anticipation by positioning icons randomly for each task and by randomizing the order of icon prompts. This study is further distinguished from Kholmatova's by more expansive data collection and analysis and by the introduction of color as a second variable.

METHODOLOGY

Icon Design

The approach used for selecting which icons to use for the study followed Johnson's

initial premise about the distinction between "solid" and "hollow" icons. For general

purposes, the terms "solid," "filled-in," and "fill" are interchangeable, as are "hollow,"

"outline," and "line." For the remainder of this study, the terms "filled-in" and "outline"

will be used, except where quoting.

The flat, single-color icons under study are called "tab bar" or "action bar" icons

(simplified as "bar" icons hereafter) in the parlance of Apple and Android design

guidelines, respectively (Apple, 2013; Google, n.d.). These simple icons are distinguished

from the often more elaborately stylized "launcher" icons that represent applications on a

desktop or mobile home screen. Bar icons are commonly found within mobile

applications, aligned in a row or "bar" of up to five icons or so at the top or bottom of the

screen to act as navigation or shortcuts to common actions. The Android Design

Guidelines describe bar icons as follows:

> "Action bar icons are graphic buttons that represent the most
> important actions people can take within your app. Each one
> should employ a simple metaphor representing a single concept
> that most people can grasp at a glance." (Google, n.d.)

Similarly, Microsoft prescribes the use of iconography in its visual identity guide thusly:

"Windows icons distill and simplify concepts using a minimum of parts and details. They

reinforce the idea of content over chrome by being graphic in nature, flat in perspective, and monochromatic" (Microsoft, 2013).

Icons of this style are also used quite often in web design, and many collections of flat, single-color icons can be found online. Some notable examples are the Glyphicons icon set provided with Bootstrap, a popular free framework for styling websites (Otto, Thornton, & Bootstrap Contributors, n.d.); the Noun Project, a large open-source library of single-color icons (The Noun Project, n.d.); and several other collections like Font Awesome (Gandy, n.d.), Foundation Icons (Zurb Inc., n.d.), and FlatIcons.com (Flaticon, n.d.). Google provides its collection of Android bar icons as a free download (Google, n.d.), and numerous designers have made collections of original and derivative icons available for free online. Browsing through these collections of icons makes it clear that small, flat, monochrome icons constitute a discrete, fairly consistent, platform-agnostic phenomenon.

It should be noted that the term "flat" is not sufficient as a descriptor for the style of icon described above. It is possible for an icon to be characterized as flat and still use color, shading, and perceived depth, as shown in Figure 3. In contrast, the icons discussed by Johnson and used in this test can be thought of as *totally* flat: using a single color, hard edges, no shading, and no perceived depth (or very little).
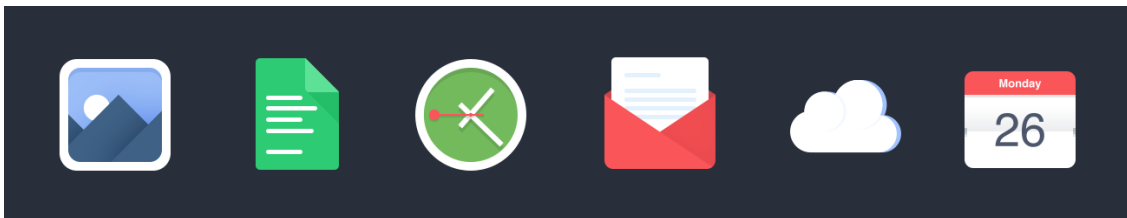

Figure 3. Not all flat icons are totally flat. The icons above are more expressive than the ones used in this study. (Borodin, 2013)

Even with the limited expressive palette of flat, single-color icons, designers have room for considerable variation in each icon's form. For example, the icon set introduced by Apple for iOS7 is unique in its use of two versions of each icon, to which Johnson was responding with his post on Medium. Johnson's analysis, however, is something of an oversimplification; "solid" versus "hollow" does not constitute an absolute, binary choice. For instance, even in the three examples that Johnson used to illustrate this dichotomy (Figure 4), the icon at bottom-right clearly shows aspects of both styles, with a filled-in box and an outlined arrow.

Figure 4. Johnson's argument treats filled-in and outline styles as absolutes, but his own example shows that icons can have characteristics of both styles at once, as in the outlined arrow in the otherwise "solid" icon at bottom-right. (Johnson, 2013)

Indeed, the iOS Human Interface Guidelines are explicit about approaching each icon

design on its own merits by highlighting a few specific cases, shown in Figure 5.
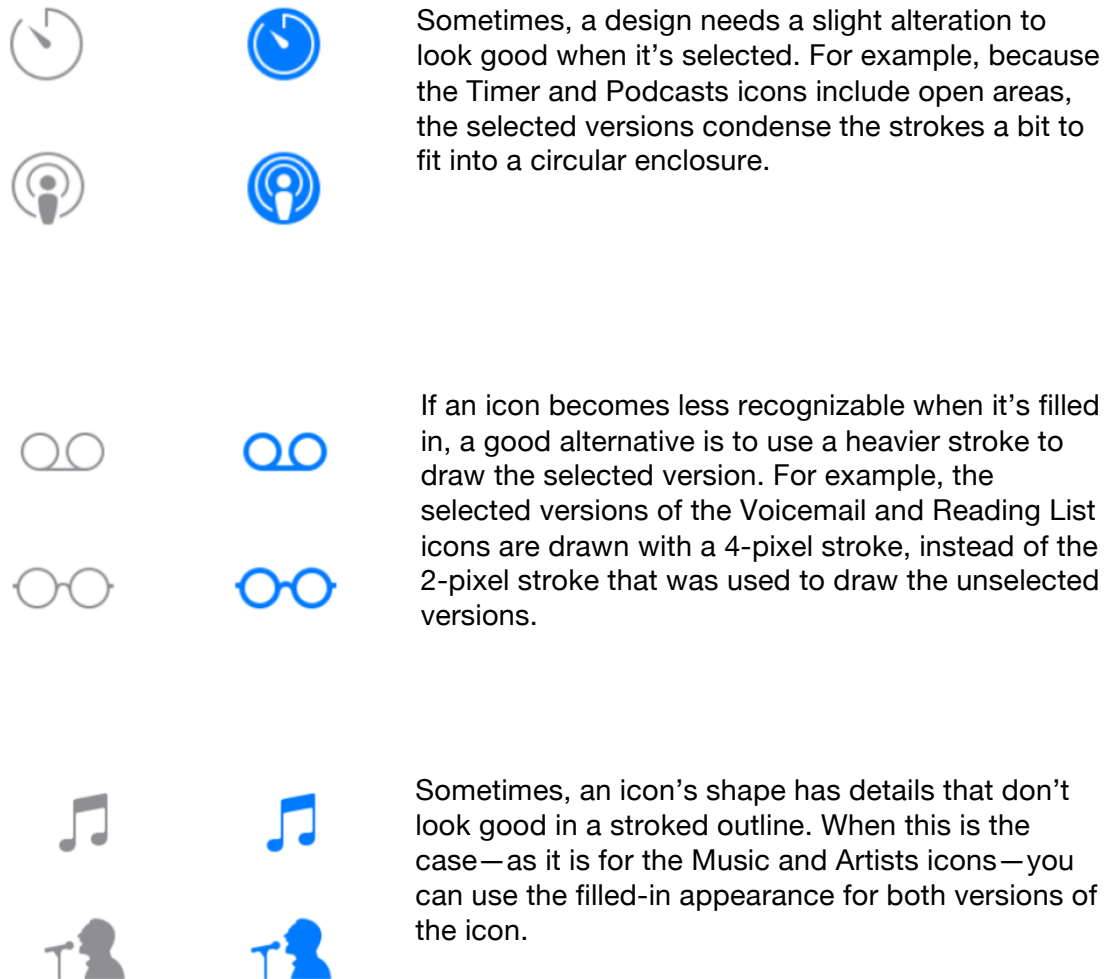
Sometimes, a design needs a slight alteration to look good when it's selected. For example, because the Timer and Podcasts icons include open areas, the selected versions condense the strokes a bit to fit into a circular enclosure.

If an icon becomes less recognizable when it's filled in, a good alternative is to use a heavier stroke to draw the selected version. For example, the selected versions of the Voicemail and Reading List icons are drawn with a 4-pixel stroke, instead of the 2-pixel stroke that was used to draw the unselected versions.

Sometimes, an icon's shape has details that don't look good in a stroked outline. When this is the case—as it is for the Music and Artists icons—you can use the filled-in appearance for both versions of the icon.

Figure 5. Apple provides specific guidelines for creating two versions of an icon. Sometimes an outline icon is not possible or ideal. (Apple, 2013)

There is clearly a degree of artistic discretion implied in Apple's design guidelines, and

in general, icon style is not a binary distinction between purely filled-in or purely

outlined forms. Therefore, in order to measure the relative usability of the two styles in

this study, icon forms were chosen that avoid this stylistic gray area and clearly represent one style or the other.

Altogether, the icon-recognition test used 20 base icon forms borrowed from Apple iOS7 (8 icons), Google Android (8 icons), Microsoft Windows Phone (3 icons), and one directly from the example used in Johnson's post. Forms from multiple icon sets were included to reduce the effect of participants' familiarity with a particular set. Because the Windows Phone and Android icons are only available in a filled-in style, outline versions were created in Adobe Illustrator by inverting the icon color and adding a 2-pixel stroke to the icon shape (centered on the shape path). All icons use a consistent 2-pixel line width in order to give the set a cohesive appearance.

Filled-in and outline versions of Trash Can, Cloud, Flag, Lock, Radio, Shopping Cart, Trophy, and Tools icons that mimic iOS 7 bar icons were downloaded with a free license from Pixeden.com (Pixeden, 2013a, 2013b, 2013c). Filled-in versions of Cog, Person, and Camera icons that mimic Windows Phone icons were downloaded with a free license from Modern UI Icons (Andrews, n.d.). Filled-in versions of Microphone, Magnifying Glass, Phone, Thumbs Up, Scissors (extracted from a larger Cut icon), Star, Key, and Tags icons were downloaded from Android Developer Style Guidelines (Google, n.d.). Filled-in and outline versions of the Speech Bubble icon were recreated from their appearance in Johnson's original post (Johnson, 2013).
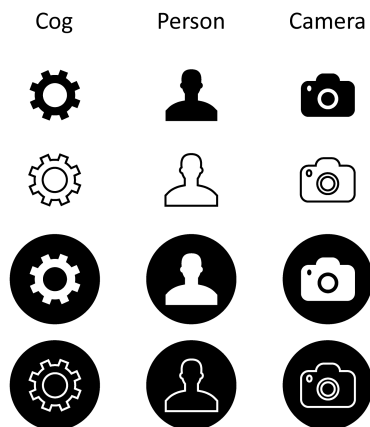
Apple iOS Icons

Trash Can	Cloud	Flag	Lock	Radio	Shopping Cart	Trophy	Tools

Android Icons

Microphone	Key	Magnifying Glass	Phone	Thumbs Up	Tags	Scissors	Star

Windows Phone Icons

Cog	Person	Camera
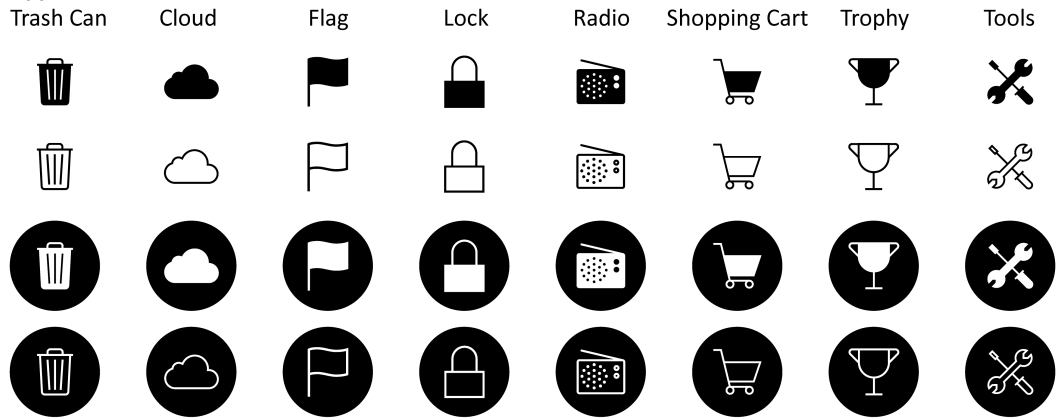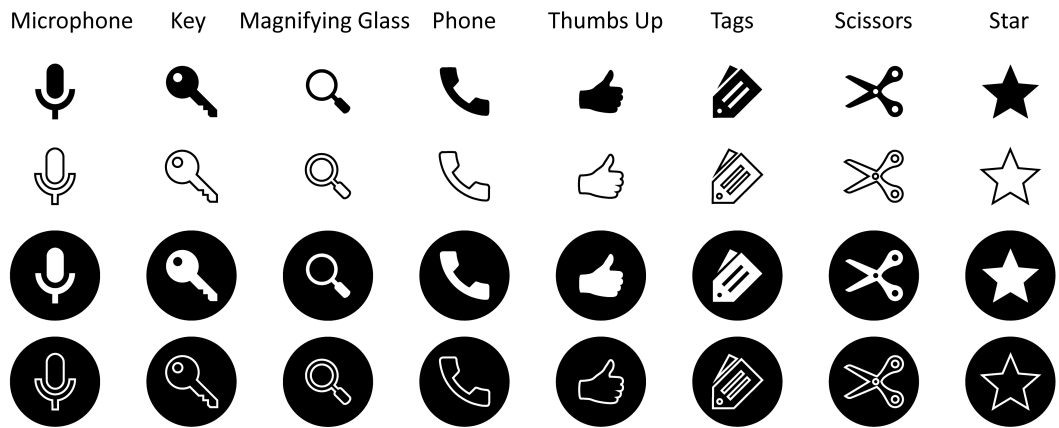
Johnson Icon

Speech Bubble

Figure 6. For each base icon form, four distinct style/color combinations were used. During the test, white icons were presented against a continuous black background, not within black circles as above.

To help minimize semantic factors of icon recognition, concrete pictorial representations were chosen and abstract or arbitrary icons (e.g. Copy, Share, Save) were avoided. The names given for each icon were concrete and unambiguous; for instance, Scissors was used instead of Cut and Trash Can was used instead of Trash.

In addition to the filled-in and outline versions of each icon's base form, two versions with inverted colors were created for each form (Figure 6). Altogether, a total of 80 distinct icons were used in the test (20 base icon forms × 2 styles × 2 color combinations). These variables will be referred to hereafter as "form," "style," and "color."

During the test, icons were displayed at a size of approximately 50 pixels wide (each icon form varied slightly) and optimized for sharpness on both standard and high pixel density displays. Although it is impossible to know each participant's screen size, pixel density, level of browser zoom, distance from the screen, and even eyesight, an icon size of 50 pixels should have fallen well within a perceived size of 0.7° – the minimum recommended by Lindberg & Näsänen (2003) – if it is assumed that the participant took the test at a comfortable distance with a readable zoom level.

Ideally, the quality and brightness of the display device, as well as the size and color of icons in the test, should be the same for every participant to avoid introducing their own effects on an experiment (Ling & Van Schaik, 2002; Näsänen, Karlsson, et al., 2001; Näsänen, Ojanpää, et al., 2001; Näsänen & Ojanpää, 2003; Ojanpää & Näsänen, 2003). Although this test was made freely available on the web and experienced via a variety of hardware and display settings, the overall aesthetic of the test was designed to be simple enough to translate well across any screen large enough to display it. Icons and text were

displayed only in black or white, minimizing the effects of poor display brightness or

contrast.

Test Design

The test environment was the user interface of a web application developed specifically for this test using Ruby on Rails and rendered using HTML, CSS, and Javascript. The test could be accessed by web browser at the following URL: http://icon-test.net. The application was hosted on Heroku, a cloud-based application hosting platform.

Using CSS media queries, the testing application sensed whether a browser viewport was large enough to display the entire test area without requiring scrolling (740 pixels in width by 700 pixels in height), and disabled interaction if the viewport was too small (or if it became too small due to resizing the window). This also had the practical effect of preventing participants from taking the test from most mobile devices, where the small display would have created difficulty in selecting icons accurately. Strictly speaking, however, the test was not disabled for touch devices, just for small screens.

The test was designed to perform consistently across a range of connection speeds. Before being allowed to begin the test, the image files for all 80 form-style-color icon variations were downloaded and cached in the participant's browser so that they could be rendered instantly at the beginning of each icon selection trial. CSS and Javascript were used to hide the icon array before each trial and then instantly reveal it on the user's command without requiring a request to the server. Javascript was used to record start and end timestamps and calculate a task time in milliseconds for each trial, which was sent to the server between trials. Calculating task times on the client side helped ensure their accuracy.

Although each participant was exposed to all 80 possible form-style-color variations over the course of the test, each test sequence consisted of only 24 icon recognition trials,

which kept the test time at around five minutes. For each icon-selection trial, each of the 20 icon forms (one for the icon matching the prompt and 19 distractors) were arrayed in a circular formation around a prompt word, positioned randomly and spaced evenly. For each trial, all icons were displayed with the same style (filled-in or outline) and color combination (black on white or white on black) (Appendix, Figures 4, 5).

Unknown to the participant, the results from the first four trials would be considered warm-ups and were not used for data analysis. Each of the first four warm-up trials tested for one of the 20 icon forms (selected at random without repeating) using one of the four style-color combinations (selected at random without repeating). This was to help the participant become comfortable with the testing procedure and with each of the four style-color combinations before the real trials began.

Each of the next 20 trials tested for one of the 20 icon forms (selected at random without repeating) using one of the four style-color combinations (ordered randomly but shown exactly five times each).

The background color of the test environment was white, so icons with black backgrounds were positioned within a black ring (Appendix, Figure 5). Containing an icon within this ring shape was seen as preferable to completely inverting the color of the background and prompt text, which might have been be jarring and disorienting to participants.

After all 24 trials were completed, the participants were instructed to fill out a brief close-ended questions to indicate their age range and the desktop and mobile operating systems that they are most comfortable using (Appendix, Figure 6). The operating system and browser that the participant used for the test were also recorded using a Ruby gem that

detected the participant's user agent (Vieira, n.d.). This data was used for exploratory

data analysis and to help troubleshoot any compatibility issues that might have occurred

with the testing application.

Partly as an incentive to complete the test, participants were shown a personalized

Results page following completion of the questionnaire (Appendix, Figure 7). This page

showed each participant his or her average speed and success rate, a percentile score for

the participant's speed compared to all other participants who had taken the test, two

ordered lists of the five icons that the participant selected the fastest the five selected

slowest, and an interactive, filterable bar chart that allowed the participant to view and

compare the speed and accuracy scores for him/herself, for the entire sample population,

and for subsets of the sample population like people between 41 and 60 years old or

people who are most comfortable using Windows. All participant data was anonymous.

This test design was granted approval by an Institutional Review Board at the University

of North Carolina at Chapel Hill in February 2014. To advertise the test to potential

participants, email and social media was used for a snowball sampling method. The

investigator's classmates, friends, family, and social media followers were targeted, as

were specific professionals in the fields of user experience, design, and usability who are

active on Twitter. At the end of the test, participants were asked to share the test with

others as well. This sampling method was effective in gathering a large sample size but

placed limits on the study's ability to make confident inferences about the

representativeness of the sample.

Procedure

The test sequence consisted of the following:

- A text-based orientation to the purpose and format of the test and an introduction

  to the 20 icons used in the test (Appendix, Figures 1, 2)

- A sequence of 24 icon-recognition trials (Appendix, Figures 3–5)

- A brief post-test questionnaire (Appendix, Figure 6)

- The display of the participant's test results (Appendix, Figure 7).

Each participant first arrived at the root URL of the web application, http://icon-test.net

(Appendix, Figure 1). The homepage included the study title and an information sheet

about the purpose and format of the test. A button labeled "Begin the Test" led the

participant to the next step of the test.

The next screen (Appendix, Figure 2) presented an icon orientation screen with the

following text:

> "Shown below are the 20 icons used in this test. Take as much time
> as you need now to familiarize yourself with each icon and its
> name. When you are ready to begin the test, click the button
> below."

This page displayed all 20 icon forms to be used in the test in both filled-in and outline

styles (all presented as black on a white background). A button labeled "Begin the First

Trial" led the participant to the next step of the test.

Next, the participant was led through a sequence of 24 icon recognition trials. Each trial

began with a dialogue box in the center of an otherwise blank screen (Appendix, Figure

3). The dialogue box contained a trial number (e.g. "Trial 1 of 24") and the following

instructions:

> "When you are ready, press Start to begin the timed trial. The
> name of an icon will appear in the center of the screen surrounded
> by 20 randomly positioned icons. Balancing speed and accuracy,
> select the icon that matches the name."

When the participant selected the Start button, a circular array of icons was immediately

displayed surrounding the name of one icon form in the center of the test area (Appendix,

Figures 4, 5). This prompt appeared just above where the Start button had been, so that it

was within the participant's field of visual fixation but not obscured by the participant's

cursor or finger. Once an icon was selected, the application recorded a task time and

whether or not the correct icon was chosen. Next, the pre-trial dialogue box for the next

trial appeared and the participant repeated the above process again for all 24 trials.

After completing 24 trials, the participant completed a brief questionnaire (Appendix,

Figure 6). After submitting this information, the participant was shown a screen with

results from the test (Appendix, Figure 7). The results page could be re-accessed by its

unique URL. Included on the page were buttons to help participants share the test with

friends via social media sites Facebook, Twitter, Tumblr, Pinterest, Reddit, and Google+.

Limitations

On one hand, making the icon recognition test freely available on the web had the advantages of device-independence and scalability to a large sample of participants. On the other hand, the web-based format introduced the study's most obvious weakness: the inability to control all aspects of the test environment. The use of multiple screen sizes and device pixel densities make it impossible to present icons at the same absolute size across devices. Color contrast, sharpness, and luminance, factors dependent in part on individual display device quality and settings, have been shown to influence reaction time in other studies of icon recognition (Ling & Van Schaik, 2002; Näsänen, Karlsson, et al., 2001; Näsänen, Ojanpää, et al., 2001; Näsänen & Ojanpää, 2003; Ojanpää & Näsänen, 2003). However, generally, perception of icons is quite resistant to moderate deterioration of image quality (Näsänen & Ojanpää, 2003).

The test design also relied to some extent on the good faith of participants. Participants were free to take the test multiple times (although the homepage information sheet asked them not to), which allowed participants to potentially learn and improve at the tasks over time. Like other studies that use self-reported questionnaires, this study relied on participants to report honest answers, but they may not have done so.

There is also a concern of content invalidity, that is, the study design does not measure all of the conditions that go into icon recognition. In real-life user interfaces, icon positions don't usually change. Coding meaning to the *location* of an icon is an aspect of icon search that has been purposely ignored for this study (Blankenberger & Hahn, 1991; Moyes, 1994) in order to isolate the effects of icon style and color.

RESULTS

The test was announced to potential participants on the afternoon of February 19th, 2014 and data collection continued for ten days, until March 1st, 2014. An initial spike in activity occurred on the first day in response to announcements made via UNC's School of Information and Library Science announcement listserv and via the investigator's Facebook and Twitter networks. By the end of the day, over 600 of the investigator's classmates, friends, followers, and indirect connections had completed the test. Traffic subsided in the following days but spiked again thanks to tweets from a several popular Twitter users, including usability professional and writer Steve Krug.

During the 10 days that the testing application collected data, 1,559 participants started the test and 1,260 completed it, for an overall completion rate of 80.8%.

Table 1 and Figure 7 show the number of participants who dropped off along each step of the test sequence. About 50% of participants who did not finish the test stopped after completing 0 or 1 trials. Trials from incomplete tests were not included in further data analysis.

Table 1. Participant drop-off

Step in Test Sequence	Completed	Drop-offs	% of total drop-offs
Begin Test	1559	-	-
Trial 1	1438	121	40.74%
Trial 2	1407	31	10.44%
Trial 3	1387	20	6.73%
Trial 4	1371	16	5.39%
Trial 5	1355	16	5.39%
Trial 6	1338	17	5.72%
Trial 7	1329	9	3.03%
Trial 8	1320	9	3.03%
Trial 9	1312	8	2.69%
Trial 10	1306	6	2.02%
Trial 11	1301	5	1.68%
Trial 12	1296	5	1.68%
Trial 13	1292	4	1.35%
Trial 14	1286	6	2.02%
Trial 15	1283	3	1.01%
Trial 16	1277	6	2.02%
Trial 17	1274	3	1.01%
Trial 18	1271	3	1.01%
Trial 19	1271	0	0.00%
Trial 20	1271	0	0.00%
Trial 21	1270	1	0.34%
Trial 22	1267	3	1.01%
Trial 23	1267	0	0.00%
Trial 24	1265	2	0.67%
Questionnaire	1260	5	1.01%
		299	100.00%
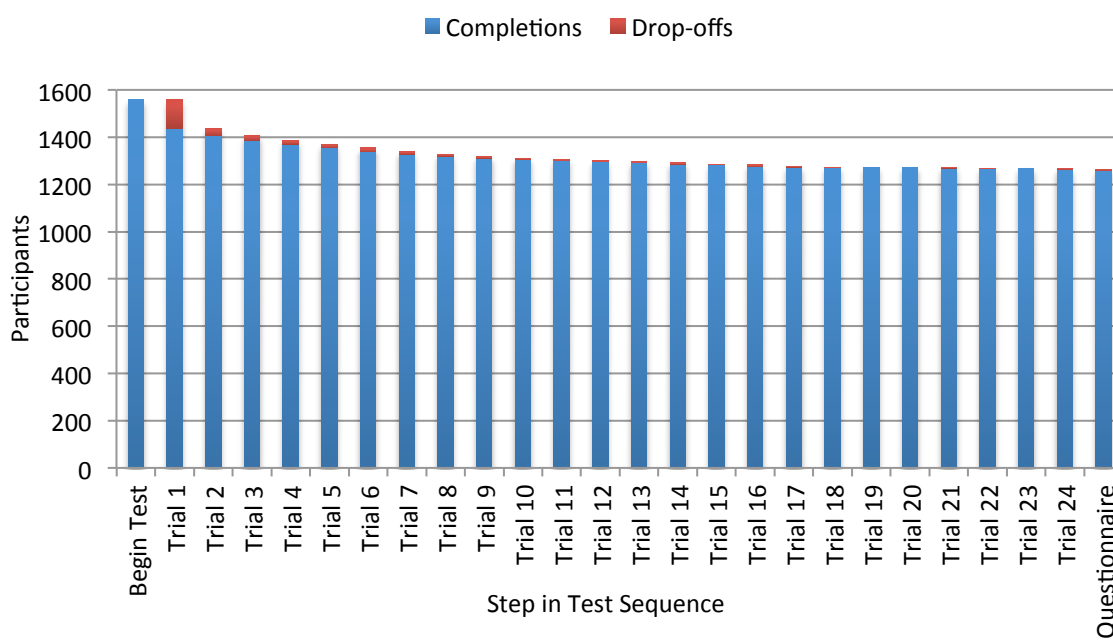
Figure 7. Participant drop-off

Demographics

The snowball sampling method used for this study was effective in gathering a large sample that skewed young, Apple-friendly, and tech-savvy. The investigator's friends and family made up one large segment of participants. Another large segment is thought to have been interface designers and other information professionals, with whom the test found a receptive audience on Twitter.

For each test, participants were asked to provide their age and the desktop and mobile operating systems with which they were most comfortable. The browser and operating system each participant used to take the test were also recorded automatically using a Ruby gem that detected the participant's user agent (Vieira, n.d.).

Table 2 and Figure 8 show the age of participants who completed the test. People aged between 26 and 40 years made up fully half of the sample, followed by people between 18 and 25, who made up one third. People older than 40 were firmly in the minority, and people older than 60 made up less than 2% of the sample.

Table 2. Participant age

Age	# Participants	% Participants
18-25	413	32.78%
26-40	638	50.63%
41-60	185	14.68%
>60	24	1.90%
TOTAL	1260	100.00%
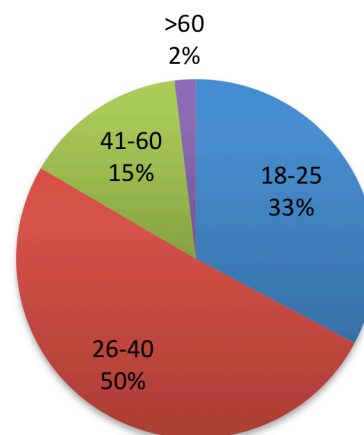
Figure 8. Participant age

Table 3 and Figure 9 show the operating systems participants used to take the test. About

95% of participants took the test using some version of Microsoft Windows, Mac OS-X,

or Linux, most likely from a desktop computer. About 5% of participants took the test

using iOS, presumably while using an iPad. This means that at least 5% of participants

completed the test using a touchscreen interface.

Desktop participants were about evenly split between Mac and Windows users, indicating

that Mac users were overrepresented in the sample, since overall market share for Mac

computers was estimated at only about 10% at the time of the test (w3schools.com,

2014).

Table 3. Operating systems used

OS	# Participants	% Participants
Mac	595	47.22%
Windows	562	44.60%
iOS	66	5.24%
Linux	33	2.62%
Other	4	0.32%
TOTAL	1260	100.00%
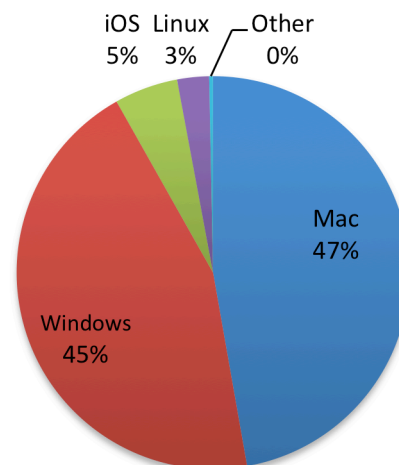
Figure 9. Operating systems used

Table 4 and Figure 10 show the web browsers that participants used to take the test. A

solid majority of participants (60%) used Google Chrome. It is unclear if the two tests

using Chrome for Android occurred from mobile devices, since both of these tests also

came from Linux-based operating systems.

Table 4. Browsers used

Browser	# Participants	% Participants
Chrome	753	59.76%
Firefox	267	21.19%
Safari	113	8.97%
Safari for iPad	66	5.24%
Internet Explorer	56	4.44%
Opera	3	0.24%
Chrome for Android	2	0.16%
TOTAL	1260	100.00%
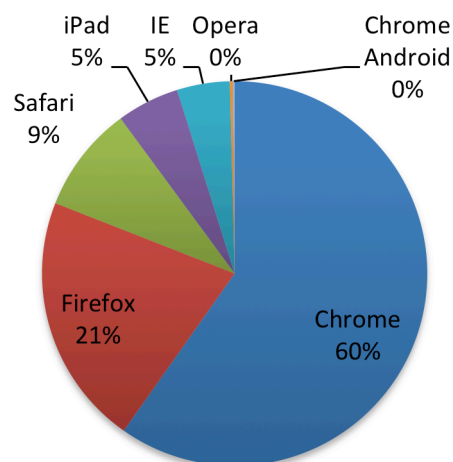
Figure 10. Browsers used

The desktop operating systems that participants said they were most comfortable using are shown in Table 5 and Figure 11. The responses map fairly closely to the distribution of operating systems that participants actually did use for the test. The data again show an overrepresentation of Mac users in the sample.

Table 5. "The desktop operating system that you are most comfortable using"

Desktop OS	# Participants	% Participants
Mac	648	51.43%
Windows	556	44.13%
Linux	39	3.10%
I don't know	15	1.19%
Other	2	0.16%
TOTAL	1260	100.00%
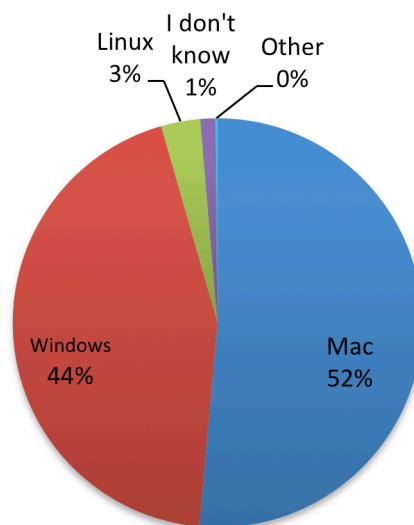
Figure 11. "The desktop operating system that you are most comfortable using"

Table 6 and Figure 12 show the mobile operating systems that participants said they were most comfortable using. iOS users made up a strong majority (60%), outnumbering Android users 2 to 1. This shows that Apple users in general, not just Mac users, were overrepresented in the sample.

Table 6. "The mobile device operating system that you are most comfortable using"

Mobile OS	# Participants	% Participants
iPhone/iOS	762	60.48%
Android	384	30.48%
I don't use a mobile device	51	4.05%
Other	31	2.46%
Windows Mobile	14	1.11%
I don't know	13	1.03%
BlackBerry	5	0.40%
TOTAL	1260	100.00%
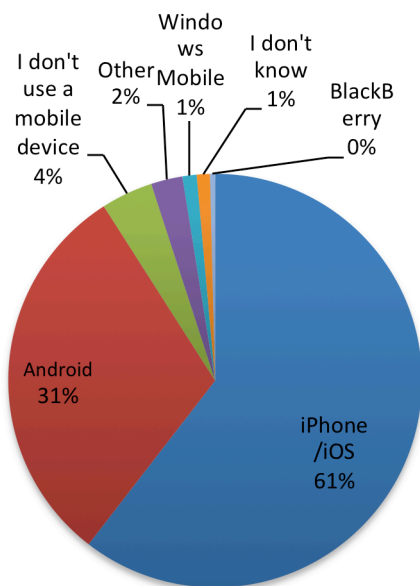
Figure 12. "The mobile device operating system that you are most comfortable using"

It should be noted again that because this test was available on the open web and did not collect any personally identifying information, some participants may have taken the test more than once or provided fictitious questionnaire responses.

Removing Failed Tests

Before further analysis, the data were examined to identify any tests that should be removed due to a high number of task failures (trials where the participant chose an icon that did not match the prompt) or due to other indications that the participant did not understand the test instructions, was purposely subverting the results, or seemed to lack a sufficient level of care or attention.

Table 7 shows the distribution of completed tests by the number of task failures in each test. More than 86% of the total sample (1,092 participants) selected every icon correctly in the test. Most participants who failed at least one task (168 participants) failed only one task (140 participants) with only a handful of participants failing more than one task (28 participants). This indicates that most participants had no trouble remembering or identifying the icons by name and making their selections without error.

Table 7. Task failures per participant

# Task Failures	# Participants	% Participants
0	1092	86.67%
1	140	11.11%
2	18	1.43%
3	4	0.32%
4	1	0.08%
6	2	0.16%
10	1	0.08%
18	1	0.08%
22	1	0.08%
	1260	100.00%

Further examination of tests with a high number of task failures (i.e. more than five)

revealed four tests that should be considered "failed tests":

- In one test, the participant failed 18 trials and had very short task times, as if

    he/she was attempting to make selections as fast as possible regardless of whether

    they were correct.

- In one test, the participant failed 22 trials, with the same behavior as above.

- In one test, the participant failed 10 trials, including the first six in a row.

- In one test, the participant failed the final 6 trials, with task times for two of those

    failures that were longer than 2 minutes.

These four tests and their associated trials were removed from the dataset. The resulting

sample is shown in Table 8.

Table 8. Task failures per participant with failed tests removed

# Task Failures	# Participants	% Participants
0	1092	86.94%
1	140	11.15%
2	18	1.43%
3	4	0.32%
4	1	0.08%
6	1	0.08%
	1256	100.00%

Removing Warm-up Trials

Although participants were not aware of it, the first four trials in each test were designed to serve as warm-up trials that would be discarded before data analysis. These four trials showed the participant all four icon style-color variations, ordered randomly with a random, non-repeating icon form, and were intended to diminish the effects of participants encountering the test procedure or a particular icon aesthetic for the first time. After removing the first four trials from each test, the total number of trials was 25,120 (1,256 tests * 20 trials), as shown in Table 9.

Table 9. Trials with failed tests and warm-ups removed

Icon Aesthetic	# of Trials	% of Trials
All	25120	100.00%
Filled-in	12560	50.00%
Outline	12560	50.00%
Black on white	12560	50.00%
White on black	12560	50.00%
Filled-in, black on white	6280	25.00%
Outline, black on white	6280	25.00%
Filled-in, white on black	6280	25.00%
Outline, white on black	6280	25.00%

Task Success

Task Success by Icon Style and Color

The prompted icon was misidentified in about one half of one percent of all trials (0.55%). There was no statistically significant effect of either icon style, χ^2 (3, N = 25120) = 0.36, p = .55, or icon color, χ^2 (3, N = 25120) = 1.88, p = .17, on task success. Averaged across all 20 icon forms, participants' ability to select the correct icon was not associated with whether the icon was filled-in, outline, black, white, or some combination of these. Table 10 shows task success and failure rates for all style-color variations.

Table 10. Task success by icon style and color

	Icon Aesthetic	# Task Failures	Success Rate	Failure Rate
	All	137	99.45%	0.55%
	Filled-in	73	99.42%	0.58%
	Outline	64	99.49%	0.51%
	Black on white	77	99.39%	0.61%
	White on black	60	99.52%	0.48%
	Filled-in, black on white	45	99.28%	0.72%
	Outline, black on white	32	99.49%	0.51%
	Filled-in, white on black	28	99.55%	0.45%
	Outline, white on black	32	99.49%	0.51%

Task Success by Icon Style and Color for Individual Icon Forms

When each icon form was analyzed individually, only one showed a statistically significant effect of either style or color on task success. The Lock icon form was selected correctly 98.2% of the time when shown as white on a black background, and 96.2% of the time when shown as black on a white background, χ^2 (3, $N = 1256$) = 4.86, p = .028. Expressed another way, the Lock icon was misidentified 2% of the time when white-on-black and 4% of the time when black-on-white.

Task Success by Trial Sequence Order

Figure 13 and Table 11 show task success by the order a target icon was presented in the test sequence. Possible sequence numbers are 5–24 because the first four warm-up trials were removed. The order a target icon was presented to the participant had no significant effect on its likelihood of success, χ^2 (19, $N = 25120$) = 21.5, p = .312.
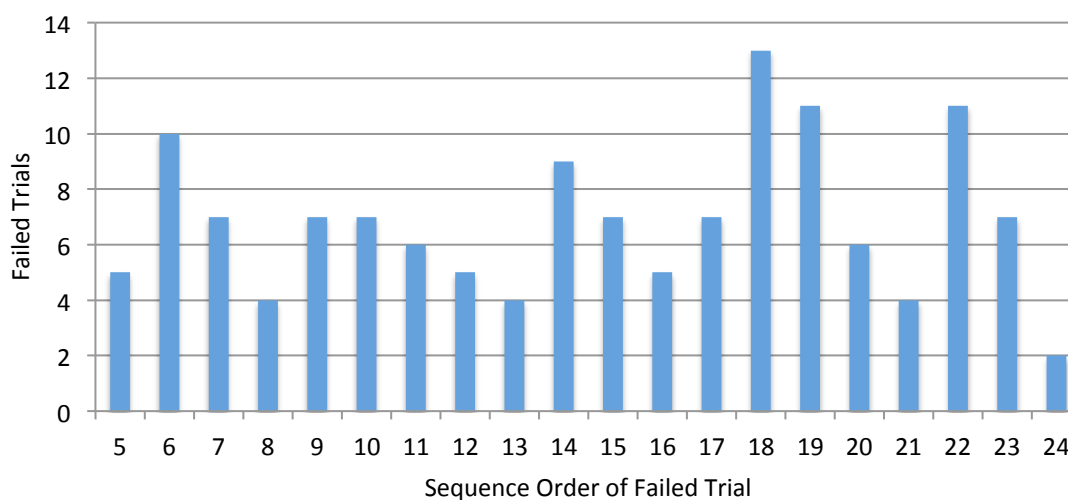

Figure 13. Task failure by trial sequence order

Table 11. Task failure by trial sequence order

Sequence Order	# Failures
5	5
6	10
7	7
8	4
9	7
10	7
11	6
12	5
13	4
14	9
15	7
16	5
17	7
18	13
19	11
20	6
21	4
22	11
23	7
24	2
	137

Task Success by Icon Form

Icon form had a statistically significant effect on task success, χ^2 (19, $N = 25120$) = 183, p < .0001. Table 12 and Figure 14 show task success by the base form of the target icon. While two icon forms were never misidentified in all 1,256 tests (Shopping Cart, Person), five others were the source of 73% of all failures (Lock, Cog, Speech Bubble, Tools, and Tags) and one alone was responsible for more than a quarter of all failures (Lock). Although there was considerable variation in the number of failures for each icon form, overall failure rates were quite low. This is a validation that the icons used for this test

were generally easy to remember and identify by name. Only four icons had greater than

a one percent probability of being misidentified, and the worst performing icon form

(Lock) was only misidentified 3% of the time on average.

Table 12. Task failure by icon form

Icon name	# Failures	Failure Rate
Shopping Cart	0	0.00%
Person	0	0.00%
Trash Can	1	0.08%
Flag	1	0.08%
Camera	1	0.08%
Thumbs Up	1	0.08%
Phone	2	0.16%
Star	2	0.16%
Radio	3	0.24%
Trophy	3	0.24%
Magnifying Glass	3	0.24%
Key	3	0.24%
Microphone	5	0.40%
Scissors	5	0.40%
Cloud	7	0.56%
Tags	12	0.96%
Tools	14	1.11%
Speech Bubble	14	1.11%
Cog	24	1.91%
Lock	36	2.87%
	137	
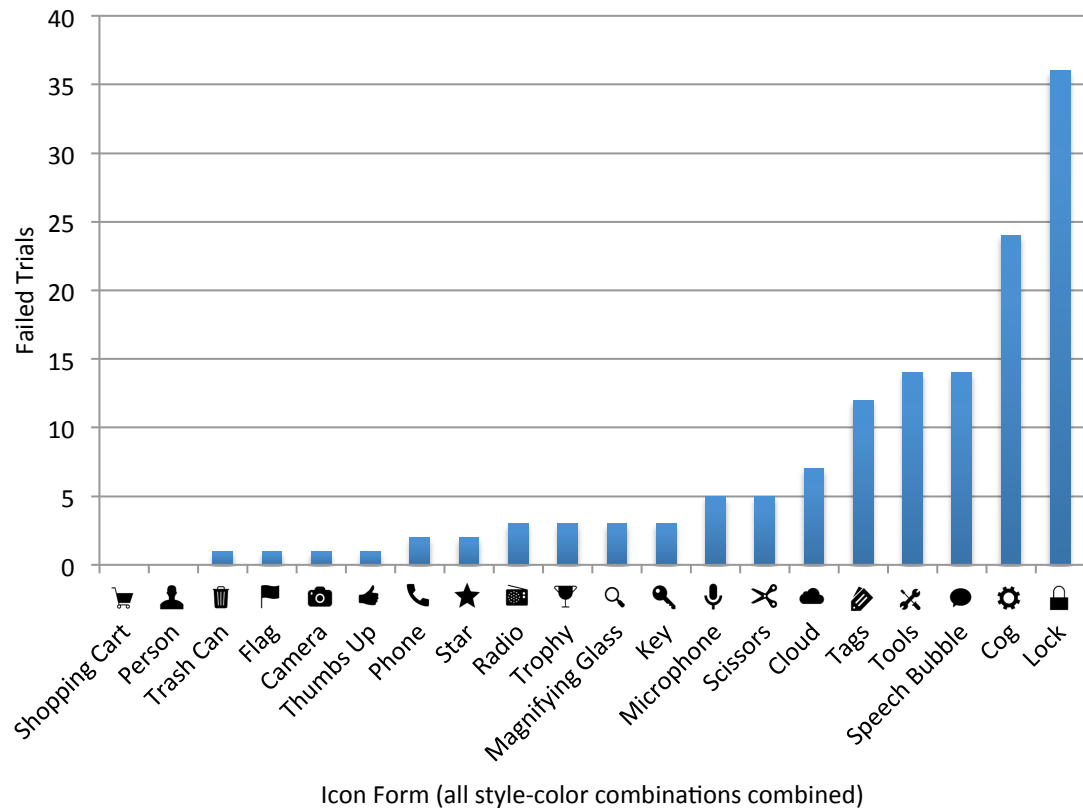
Figure 14. Task failure by icon form

Task Time

Removing Outliers

A small number of trials had unusually long task times. For instance, the maximum task time was 179,753 milliseconds, or almost 3 minutes. Very long task times suggest that a participant was not continuously engaged with the test during that time; a participant may have begun a trial but was then distracted by something outside of the testing application. Because at least some of these outliers represent invalid task time data, they should be removed from the dataset. Since there is no way to know for sure which task times represent continuous engagement and which ones do not, a reasonable threshold for removing outliers was needed.

Removing all times beyond three standard deviations of the mean is an accepted method for removing outliers in task time data (Albert, Tullis, & Tedesco, 2009). However, because only 99.1% of task times occurred within three standard deviations of the mean, a slightly less restrictive threshold was desired. Table 13 shows that the mean and median task time are quite stable across various outlier thresholds. A more conservative threshold of 99.7% was chosen because it eliminated the very long, clearly invalid task times without removing many moderately long task times that may represent valid data. Using this threshold, 75 trials with task times longer than 13,803 milliseconds (13.8 seconds) were removed from the dataset.

Table 13. Task time outlier thresholds

Threshold	Mean (ms)	Median (ms)	Std Dev (ms)	Max time (ms)	Count	# Removed
No outliers removed	3064	2599	2413	179753	25120	0
Minus top 5	3043	2599	1757	36471	25115	5
99.7%	3000	2597	1534	13803	25045	75
3 Std Devs (99.1%)	2947	2590	1381	10290	24894	226

Removing Failed Trials

In this study, task time is a measure of how long it took the participant to read the prompt word, identify the correct icon from the array, and select that icon. In trials that resulted in an *incorrect* selection, participants made a cognitive or physical error that had an impact on task time independent of the impact of the test variables (form, style, color). In order to isolate the effects of these variables, failed trials were removed from the dataset before task time analysis began.

After removing outliers, the dataset contained 125 failed trials. Removing these failed trials resulted in a dataset of 24,920 trials. The distribution of task times is shown in Figure 15. In the pictured box plot, the whiskers represent the minimum and maximum task times (927 ms; 13,790 ms), outside edges of the box represent the 25th and 75th percentiles (1,982 ms; 3,502 ms), the box midline represents the median (2,594 ms), and the hash mark inside the box represents the mean (2,989 ms).
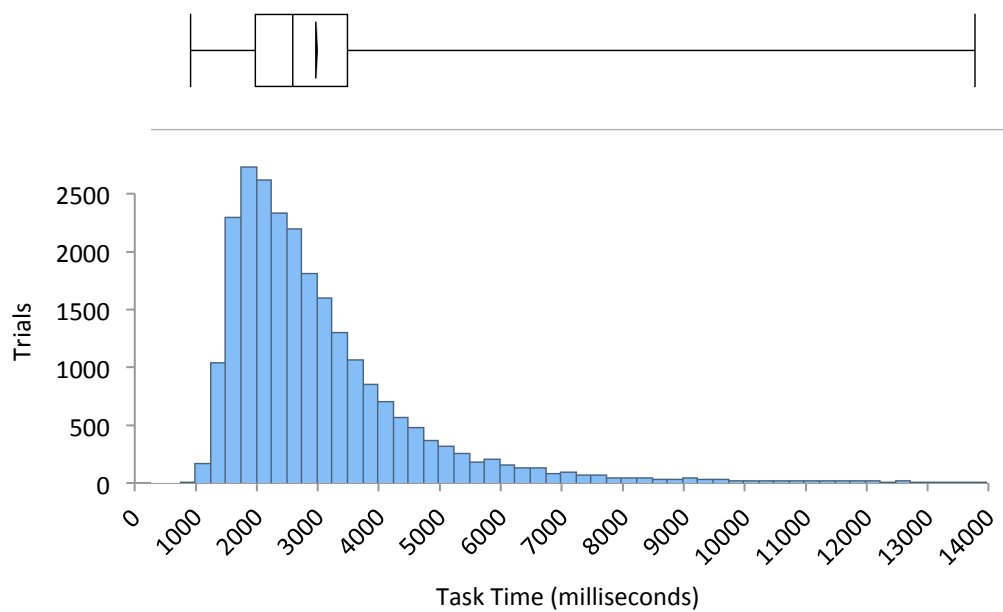
Figure 15. Distribution of task times

Task Time by Icon Style and Color

A 2×2 ANOVA with style (filled-in, outline) and color (black on white background, white on black background) revealed statistically significant main effects of both style, $F(3, 24916) = 27.7$, $p < .0001$, and color, $F(3, 24916) = 17.5$, $p < .0001$, and a significant interaction of style and color, $F(3, 24916) = 15.3$, $p < .0001$. The mean and median task times for each style-color variation are shown in Table 14.

Table 14. Task time by icon style and color

Icon Aesthetic	Mean (ms)	Median (ms)
All	2989	2594
Filled-in	2938	2554
Outline	3039	2633
Black on white	2949	2555
White on black	3039	2638
Filled-in, black on white	2936	2541
Outline, black on white	2961	2571
Filled-in, white on black	2941	2572
Outline, white on black	3116	2720

The interaction of style and color is shown in Figure 16. Post-hoc analyses using Tukey's HSD showed that outline, white on black icons were selected significantly slower than all three other style-color combinations ($p < .0001$ in each comparison) by an average duration of 170 milliseconds. No other pairwise comparisons of style-color combinations were statistically significant.
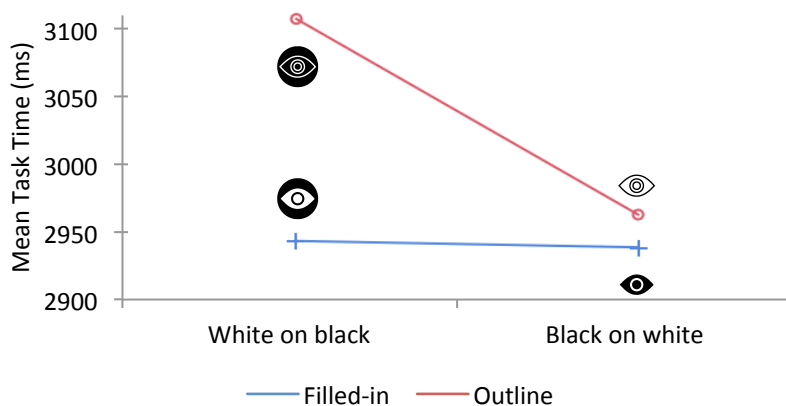
Figure 16. Task time interaction of style and color

Task Time by Icon Style and Color for Individual Icon Forms

A 2×2 ANOVA with style (filled-in, outline) and color (black on white background, white on black background) was performed for the task time of each individual icon form (forms 1–20). This revealed unique effects of style and color for each icon form.

Six icon forms showed no statistically significant effects of either style or color: Flag, Shopping Cart, Trophy, Cog, Microphone, and Star. These six icon forms performed equally well in all four style-color variations.

Three icon forms showed a statistically significant interaction of style and color: Thumbs Up, $F(3, 1248) = 18.2$, $p < .0001$; Cloud, $F(3, 1245) = 9.98$, $p = .002$; and Speech Bubble, $F(3, 1233) = 9.89$, $p = .002$. Figure 19 shows these interactions.
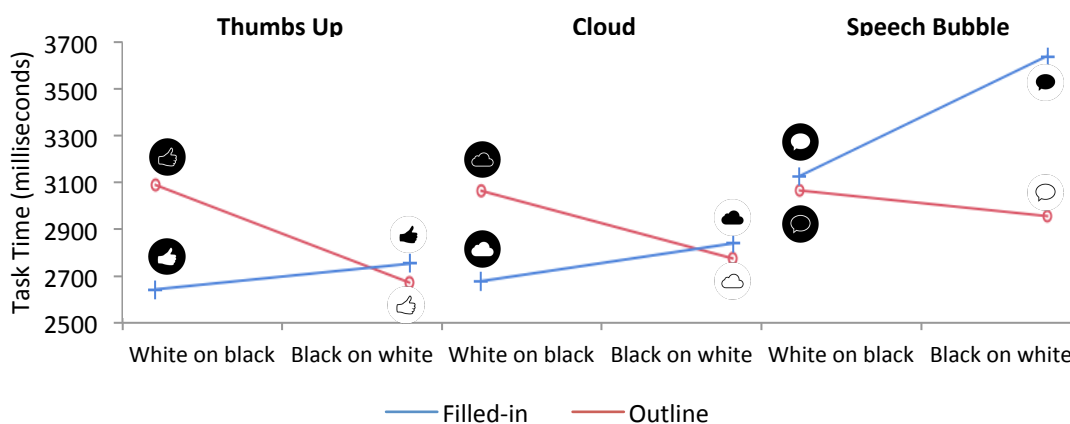
Figure 19. Task time interaction of style and color for individual icon forms

The interaction of style and color seen for Thumbs Up closely resembled the interaction seen for the combined icon set: outline, white on black icons were selected significantly slower than filled-in, black on white ($p = .0005$), outline, black on white ($p < .0001$), and filled-in, white on black icons ($p < .0001$) by an average duration of 406 milliseconds. No

other pairwise comparisons of style-color variations for Thumbs Up were statistically significant.

For the Cloud icon form, outline, white on black icons were selected significantly slower than outline, black on white (p = .019) and filled-in, white on black icons, (p = .001) by an average duration of 344 milliseconds. No other pairwise comparisons of style-color combinations for Cloud were statistically significant.

For the Speech Bubble icon form, filled-in, black on white icons were selected significantly slower than outline, black on white (p < .0001), filled-in, white on black (p = .001), and outline, white on black icons (p = .0003) by an average duration of 573 milliseconds. No other pairwise comparisons of style-color combinations for Speech Bubble were statistically significant.

Eleven icon forms showed a statistically significant main effect of style (filled-in, outline) on task time, as shown in Table 15 and Figure 17. Of these, eight were selected more quickly when shown in a filled-in style (Lock, Magnifying Glass, Phone, Tools, Person, Thumbs Up, Cloud, Scissors) and three were selected more quickly when shown in an outline style (Trash Can, Speech Bubble, Key). The effect size of style was small for most of these eleven icon forms, although five showed differences of 300 milliseconds or more. The filled-in style of the Lock icon form was selected almost three-quarters of a second faster than the outline version, and the filled-in style of the Magnifying Glass icon form was selected almost half a second faster than its outline counterpart.

Table 15. Icon forms with significant main effect of icon style

Icon Form	Task Time Advantage for Filled-In Style (ms)	Task Time Advantage for Outline Style (ms)	ANOVA	p
Lock	749		$F(3, 1205) = 36.3$	< .0001
Magnifying Glass	492		$F(3, 1245) = 34.3$	< .0001
Speech Bubble		362	$F(3, 1233) = 14.5$	< .0001
Phone	325		$F(3, 1241) = 8.55$	.0035
Trash Can		308	$F(3, 1250) = 27$	< .0001
Tools	269		$F(3, 1237) = 11.6$	.0007
Key		205	$F(3, 1248) = 5.7$	.017
Person	189		$F(3, 1251) = 13.3$	.0003
Thumbs Up	187		$F(3, 1248) = 8.99$	.003
Cloud	159		$F(3, 1245) = 4.89$	.027
Scissors	140		$F(3, 1246) = 8.11$	.005
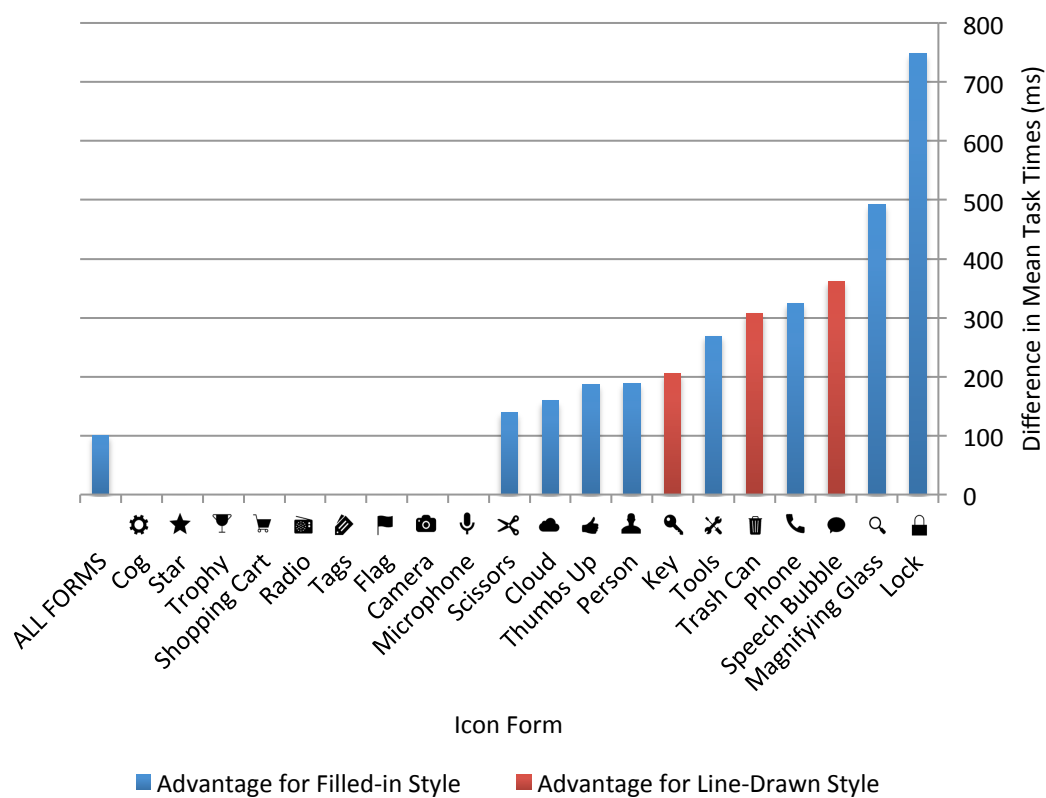
Figure 17. Statistically significant task time differences attributable to icon style

Eight icon forms showed a statistically significant main effect of color (black on white background, white on black background) on task time, as shown in Table 16 and Figure 18. Of these, seven icon forms were selected faster when shown as black on white (Tags, Magnifying Glass, Radio, Camera, Scissors, Thumbs Up, Person) and one was selected faster when shown as white on black (Speech Bubble). The effect size of icon color was generally smaller than that of style, with no differences greater than 300 milliseconds.

Table 16. Icon forms with significant main effect of icon color

Icon Form	Task Time Advantage for Black on White (ms)	Task Time Advantage for White on Black (ms)	ANOVA	p
Tags	286		F(3, 1231) = 7.22	.007
Magnifying Glass	282		F(3, 1245) = 11.2	.0008
Radio	204		F(3, 1249) = 9.13	.003
Speech Bubble		204	F(3, 1233) = 4.43	.036
Camera	198		F(3, 1248) = 7.26	.007
Scissors	172		F(3, 1246) = 10.4	.001
Thumbs Up	156		F(3, 1248) = 6.38	.012
Person	116		F(3, 1251) = 4.97	.026
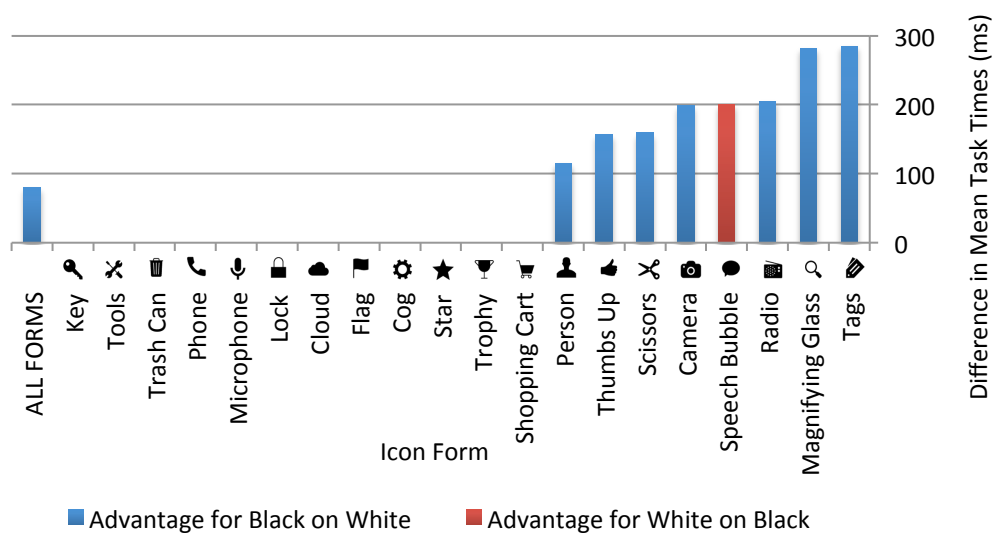
Figure 18. Statistically significant task time differences attributable to icon color

Task Time by Trial Sequence Order

Analysis of variance showed that the order in which an icon was displayed in the test had no effect on task time, $F(19, 24900) = 0.71$, $p = .81$.

Task Time by Icon Form

Analysis of variance showed a main effect of icon form (forms 1–20) on task time, $F(19, 24900) = 95.3$, $p < .0001$. The effect size of icon form was much greater than either style or color; the difference in mean task times between the best-performing icon form (Person, averaged between all style-color combinations) and the worst-performing icon form (Lock, averaged between all style-color combinations) was 1,565 milliseconds (1.6 seconds). The mean and median of each individual icon form are shown in Figure 19.
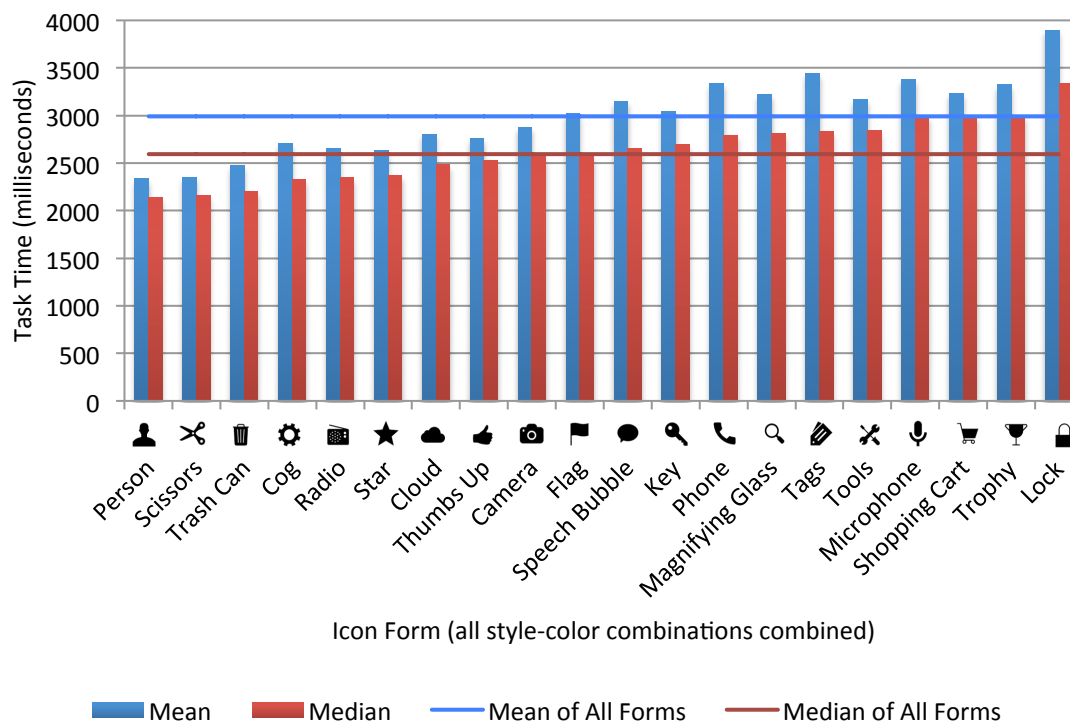
Figure 19. Task time by icon form

DISCUSSION

The primary goal of this study was to test whether "flat," single-color icons would be

selected more quickly and accurately when presented in a filled-in style than in an outline

style. The results of this study show that this is not necessarily true.

Each icon selection task in this experiment can be thought of as having occurred in two

phases: a cognitive phase, in which the participant read and interpreted the prompt,

visually scanned the icons in some manner, and made judgments about their correctness

until one was identified as correct; and a physical phase, in which the participant selected

the icon identified as correct using a mouse or by touch. These two phases map roughly

to the Goals and Operators phases of the classic GOMS (Goals, Operators, Methods,

Selection Rules) model of human-computer interaction put forth by Card, Moran, and

Newell (1986), which conceptualized task performance as a sequence of distinct

cognitive and physical phases.

Once a participant identified an icon as the correct match, the icon's form, style, and

color should not have influenced the time required to physically select it. In fact, the

experiment was designed to control for factors that might influence the duration of the

physical phase for each trial (each icon and its clickable area were the same size; the

participant's cursor or finger began in the central position where the Start button had

been) and when averaged across all trials (the position of the target icon was randomly

assigned for each trial). This suggests that differences in task times were the result of

differences in the duration of the cognitive phase. Since cognition occurs in time, the

duration of cognition serves as a measure of the amount of cognitive effort involved. In other words, longer task times indicate greater cognitive effort.

So what effect did icon style have on task time, and by implication, on cognitive effort? When averaged across the 20 icons chosen for this study, the effect of an icon's style was manifested only through interaction with its color. When icons were displayed in black against a white background, neither style had an impact on task times. (Similarly, color had no effect on the task time of filled-in icons.) However, when an outline style was combined with a white-on-black color scheme, task times increased by a small but significant duration (an average of 170 milliseconds longer than each of the other variations).

When examined at the level of individual icons, however, the results are less straightforward. For almost half of the individual icon forms (9 of 20), neither style nor color had a significant effect on task time. Only two icon forms (Thumbs Up, Cloud) exhibited a style-color interaction similar to that shown in the icon set as a whole. The only other icon form with a significant style-color interaction (Speech Bubble) actually showed a *reverse* effect: only filled-in, black-on-white icons were selected slower than the other variations. The results of Speech Bubble constitute a special case with a possible explanation. Speech bubbles are commonly presented in white and/or with an outline. A filled-in, black speech bubble lacks both of these identifying characteristics, perhaps contributing to slower recognition speed.

When main effects of style were observed in individual icon forms, they did not favor one style consistently. Among the eleven icons with significant main effects of style, eight were selected faster when presented in a filled-in style, but three were selected

faster in an outline style. Although an icon's style was shown to effect its selection time more often and with a generally larger effect size than its color, given the variability shown by individual icon forms, it is unclear whether a set of 20 different icons would produce the same aggregate results at all.

There is also a question of whether or not these small differences in task times reach the level of practical significance. Would the 170-millisecond disadvantage for outline, white-on-black icons have a meaningful impact on people's experience of using a digital interface? At the level of a single session of use, for many applications the answer is probably no. Adding 170 milliseconds to a 3,000-millisecond task amounts to just a 6% increase in its already short duration. For some individual icon styles, the influence of style may be more perceptible. The most extreme effect of icon style observed in this study, a 750-millisecond advantage for the Lock icon when shown in a filled-in style, amounts to a 25% decrease in task time. Although this icon form was an outlier in terms of both failure rate and task time, a task time advantage approaching one second does suggest that icon style can have a meaningful impact for some icons, if not for all icons generally.

Johnson (2013) argued that outline icons may have a long-term impact on users' satisfaction with an interface, as accumulating cognitive effort causes users to "tire of the design and decide they don't like it." This proposition is difficult to test, given the many possible factors that might contribute to users' dissatisfaction with an interface. However, as this study showed, icon style did not contribute to a higher rate of selection errors, one potential source of frustration. Furthermore, since icons are often used in static positions in an interface and since people tend to map the meaning of an icon to its position rather

than reinterpreting it during every use, it is reasonable to assume that whatever extra cognitive effort an icon's appearance contributes is only experienced the first several times a user interacts with an interface.

Given the evidence, it is impossible to declare a "winner" between filled-in and outline styles for flat, monochromatic icons. For the 20 icons used in this study, the effect of style was (mostly) small and inconsistent. These results suggest that designers should not view outline icons as inherently less usable than their filled-in counterparts. As mentioned earlier, the very distinction between what constitutes a filled-in or outline icon is somewhat fluid, a fact that undermines prescriptive arguments about binary style variations in icons.

The results of this study suggest that alternating between two style variations is generally an acceptable way to indicate the state of flat, single-color icons, especially since this practice may have an accessibility benefit for users with color blindness. As with any other method for showing the state of an interface element, however, designers should take care to provide sufficient contextual clues about which items are selected and which are unselected. Using a redundant visual cue, such as showing selected icons in a different style *and* color (as Apple does), will help users differentiate the two styles and reinforce their semantic distinction.

Perhaps most importantly for icon designers, this study showed that an icon's form has a considerably larger effect on task time and success rate than either its style or color. Despite attempting to limit the influence of semantic factors in the icon recognition test, these factors proved more influential than the visual factors under study. Although visual factors were relatively easy to isolate and measure, icon style is but one component of the

complex, multifaceted activity of icon design. As a final example about the complexity of icon design, Table 17 compares the ranking of each icon form by its median task time to its ranking by task success.

Table 17. Icon forms ranked by task time and task success

Icon name	Rank by Median Task Time (1 = fastest)	Rank by Task Success (1 = highest success rate)
Person	1	2
Scissors	2	14
Trash Can	3	4
Cog	4	19
Radio	5	10
Star	6	9
Cloud	7	15
Thumbs Up	8	7
Camera	9	6
Flag	10	5
Speech Bubble	11	17
Key	12	12
Phone	13	3
Magnifying Glass	14	11
Tags	15	16
Tools	16	18
Microphone	17	13
Shopping Cart	18	1
Trophy	19	8
Lock	20	20

The table shows that some icon forms had among the fastest task times *and* the highest rates of successful selection. For instance, Person and Trash Can were selected both quickly and accurately. We can speculate that this is because these icons were familiar to participants, not similar to any other icons in the set, and labeled in a way that made sense to participants.

In contrast, some icon forms took participants longer to identify and were more likely to be misidentified, like Lock, Tools, and Microphone. Lock was especially bad, perhaps because of its overly simple stylized form, which participants may have misidentified (say, as a handbag) or simply been baffled by. It is also possible that some participants chose the conceptually related Key icon instead. (Regrettably, the testing application was not designed to record this information.)

Additionally, some icon forms took a longer time to identify but were eventually identified correctly, like Shopping Cart and Phone. Perhaps participants were looking for the image of a smartphone instead of the shape of the increasingly antiquated telephone handset. Unlike in the case of Lock and Key, there were no obvious conceptual siblings for these icons, so they were eventually identified successfully.

Finally, some icons were selected quickly but misidentified, like Scissors and Cog. Perhaps in both cases, participants quickly selected the Tools icon, which was conceptually related to Cog and visually similar to Scissors.

As this example shows, designing icons is a complex exercise involving a host of creative decisions about each icon's concreteness, uniqueness, simplicity, familiarity, clarity of labeling, cultural factors, and other considerations. Given this complexity, the desire for simple, binary style rules that have a predictable influence on icon usability is understandably attractive, but perhaps unrealistic.

CONCLUSION

This study sought to determine whether flat, single-color icons could be more quickly and accurately recognized by users when presented in a filled-in style or an outline style. A software application was built which allowed participants to take a 5-minute test measuring their speed and accuracy in selecting prompted icons from among an array of distractor icons, where icons where shown in either the filled-in or outline style and either as black on a white background or white on a black background. The test was made available on the open web and 1,260 participants completed the test.

Averaged across the 20 distinct icon forms used in the test, icon style and color had no effect on task success or task time independently, but had a statistically significant interaction effect on task time. Outline icons shown in white against a black background led to task times about 170 milliseconds slower than the other style-color variations, which showed no significant differences in comparison to each other. However, for individual icon forms, the effect of icon style varied from no effect (9 icon forms) to a task time advantage for filled-in icons (8 icon forms) to a task time advantage for outline icons (3 icon forms). Except for a few exceptions, most effect sizes were small. This study concluded that a filled-in icon style is not objectively better than the outline style, and that the form of an icon has a greater influence on its usability than do either style or color.

ACKNOWLEDGEMENTS

APPENDIX
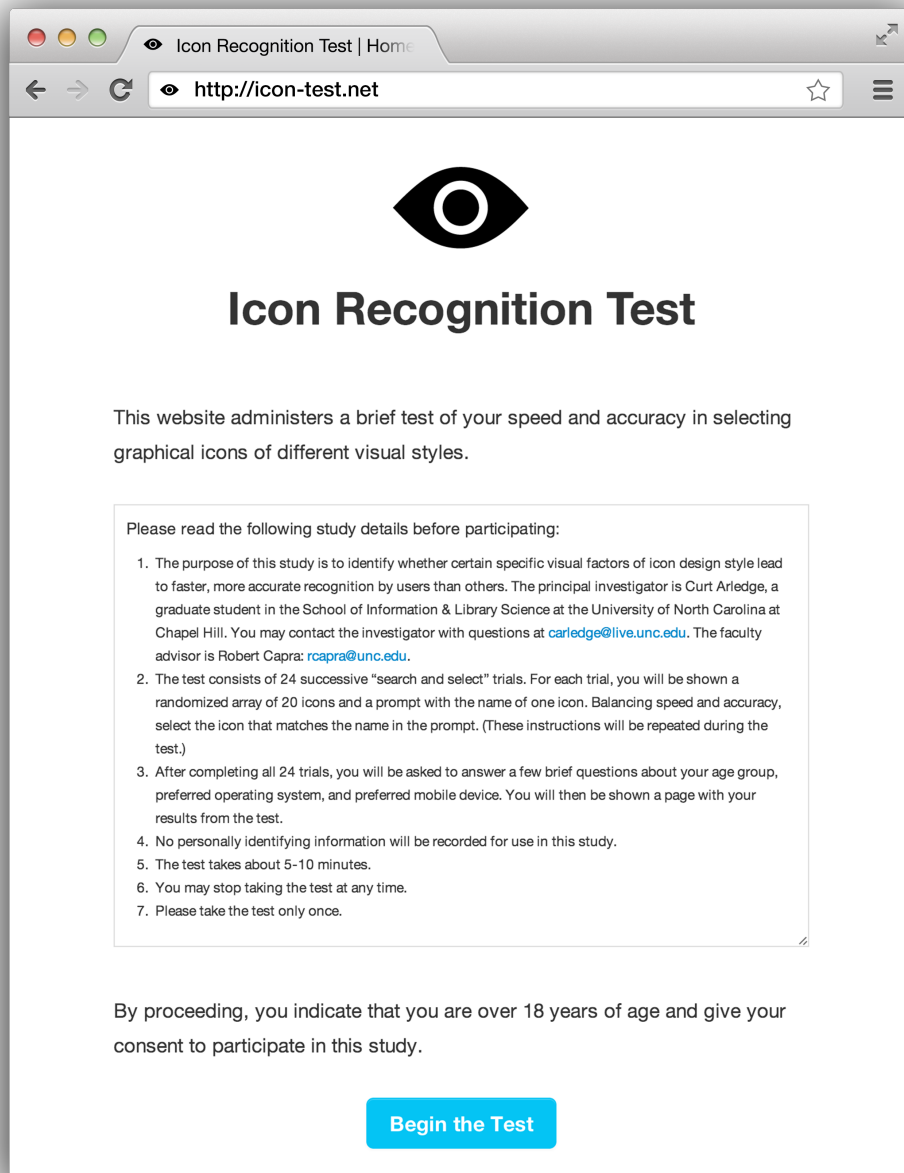
Figure 1. The home screen at http://icon-test.net.
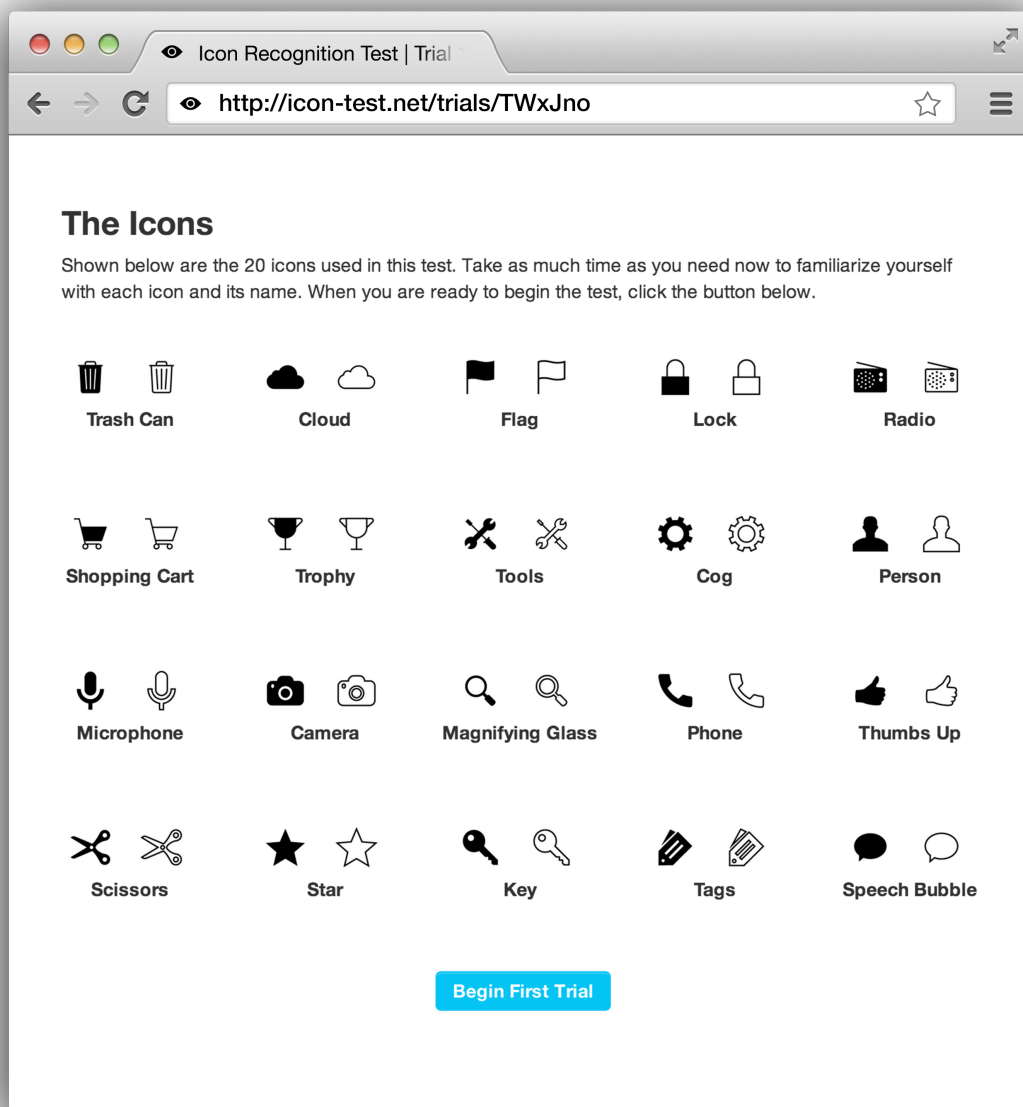
Figure 2. Before beginning the test, the participant is shown all 20 icons with their labels.
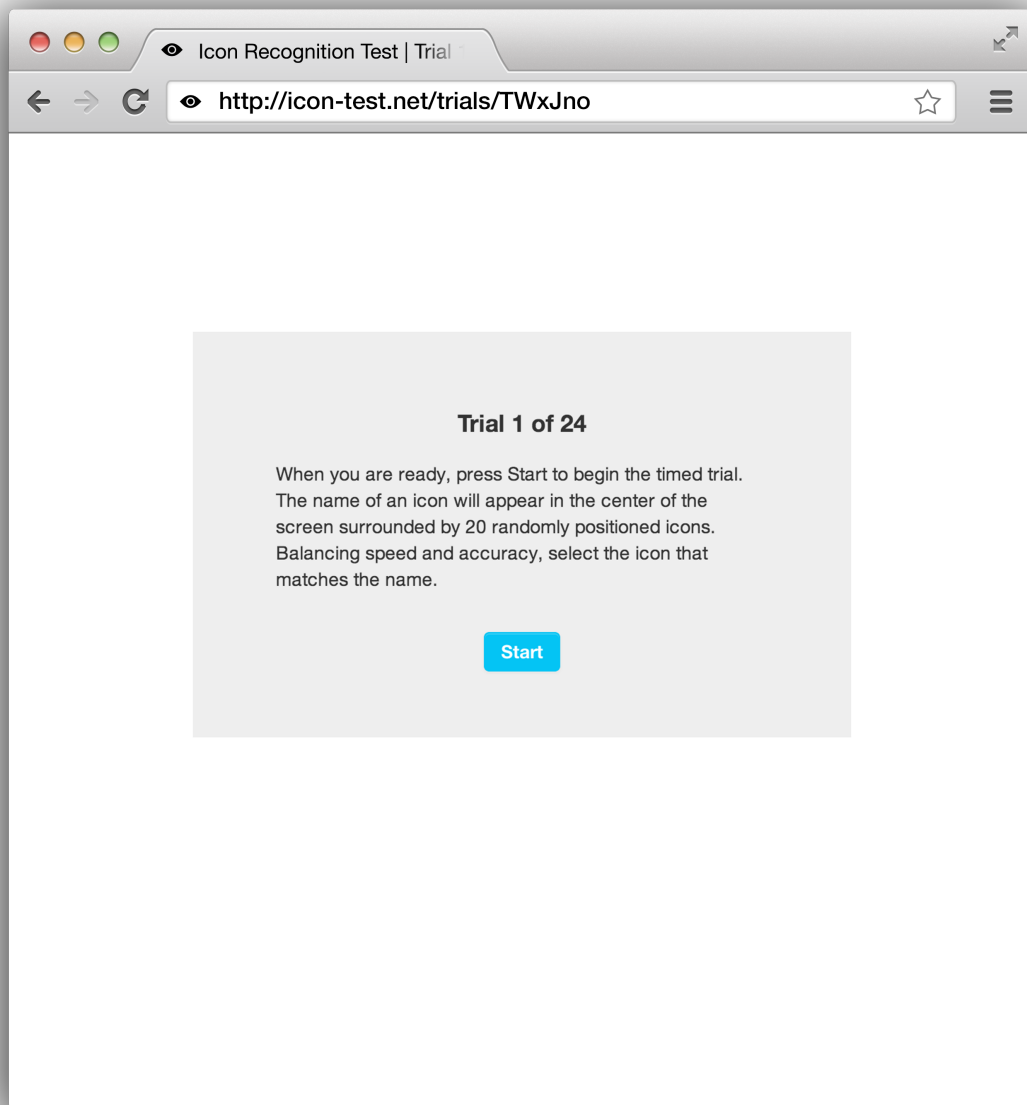
Figure 3. Each trial begins with the pictured message and current trial number.
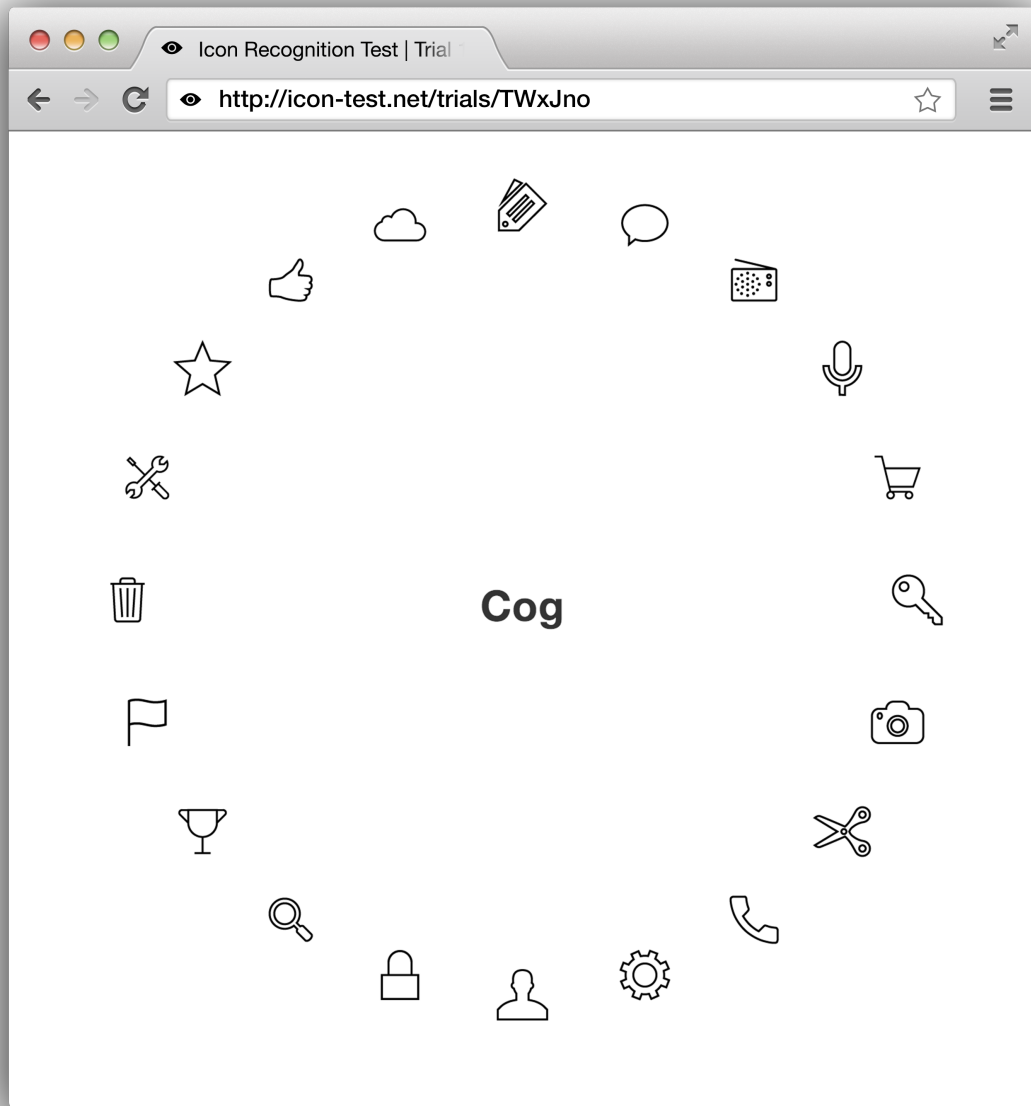
Figure 4. After pressing Start, an array of icons appears and the timer begins. When the participant selects an icon, the timer ends and the next trial begins with a message like that in Figure 3.
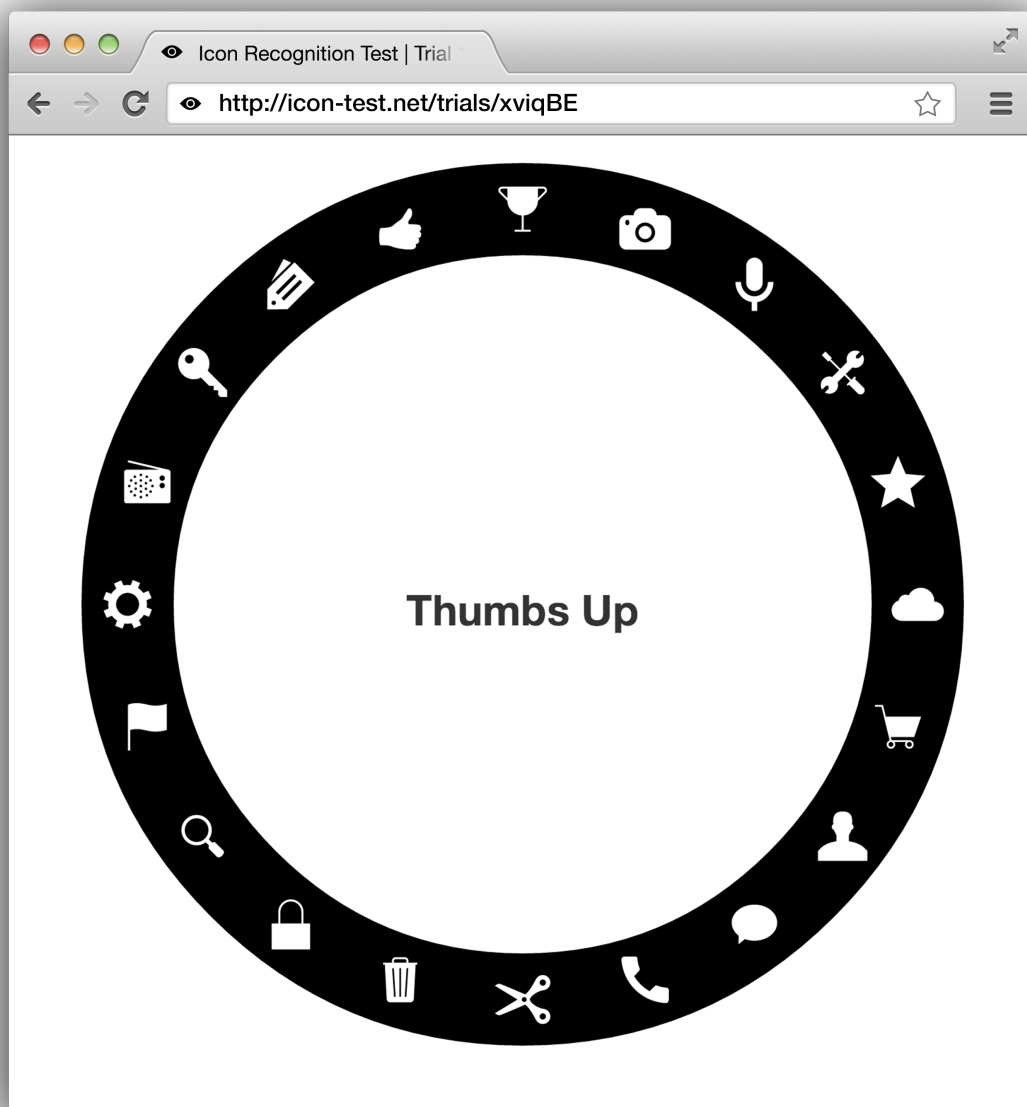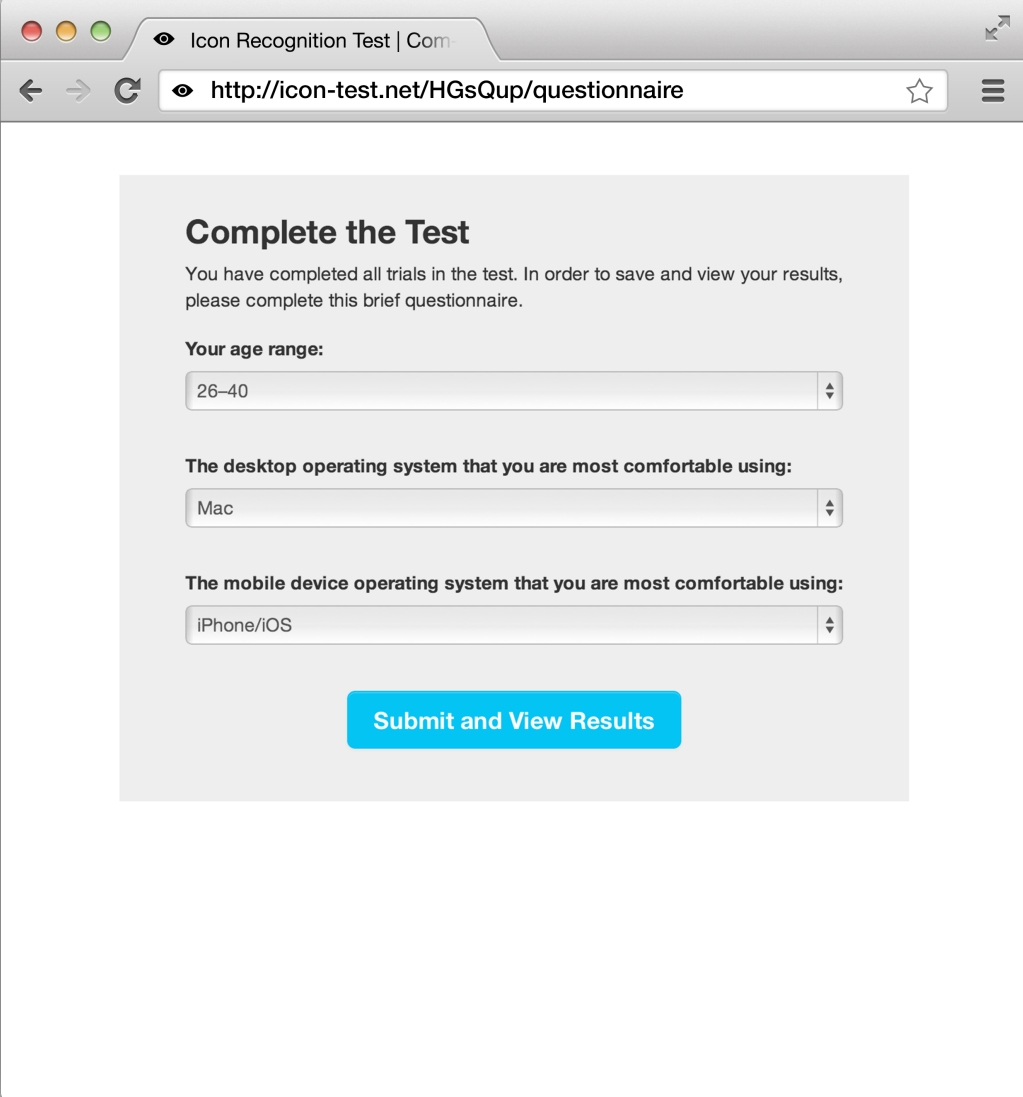
Figure 5. For trials where white icons are displayed, they are contained within a black ring.

Figure 6. After all 24 trials have been completed, the participant is presented with a blank questionnaire (shown here with mock input).

Figure 7. Finally, the participant is shown results from his/her test and an interactive data visualization representing the results from everyone who has taken the test.

BIBLIOGRAPHY

Albert, W., Tullis, T., & Tedesco, D. (2009). *Beyond the Usability Lab: Conducting Large-scale Online User Experience Studies*. Morgan Kaufmann.

Andrews, A. (n.d.). Modern UI Icons. Retrieved February 18, 2014, from http://modernuiicons.com

Apple. (2013). iOS Human Interface Guidelines. *Apple Developer*. Retrieved February 18, 2014, from https://developer.apple.com/library/safari/documentation/UserExperience/Conceptual/MobileHIG/BarIcons.html

Bernsen, N. O. (1994). Foundations of Multimodal Representations: a Taxonomy of Representational Modalities. *Interacting with Computers*, *6*(4), 347–371.

Blankenberger, S., & Hahn, K. (1991). Effects of icon design on human-computer interaction. *International Journal of Man-Machine Studies*, *35*(3), 363–377.

Borodin, P. (2013, April 15). Flat Icons (PSD). *Dribbble*. Retrieved February 18, 2014, from http://dribbble.com/shots/1029199-Flat-icons-PSD-3-Dribbble-invites

Byrne, M. D. (1993). Using Icons to Find Documents: Simplicity Is Critical. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*. ACM.

Card, S., Moran, T. P., & Newell, A. (1986). *The Psychology of Human-Computer Interaction*. CRC Press.

Evers, V., Kukulska-Hulme, A., & Jones, A. (1999). Cross-cultural understanding of interface design: A cross-cultural analysis of icon recognition. *IWIPS*.

Flaticon. (n.d.). *Flaticon*. Retrieved February 18, 2014, from http://www.flaticon.com

Gandy, D. (n.d.). Font Awesome. Retrieved February 18, 2014, from http://fortawesome.github.io/Font-Awesome

Garcia, M., Badre, A. N., & Stasko, J. T. (1994). Development and Validation of Icons Varying in Their Abstractness. *Interacting with Computers*, *6*(2), 191–211.

Gittins, D. (1986). Icon-based Human-computer Interaction. *International Journal of Man-Machine Studies*, *24*, 519–543.

Google. (n.d.). Iconography. *Android Developers*. Retrieved February 18, 2014, from http://developer.android.com/design/style/iconography.html

Holloway, J. B., & Bailey, J. H. (1996). Don't use a product's developers for icon testing. In *Conference Companion on Human Factors in Computing Systems*. ACM.

Huang, K.-C. (2007). Effects of computer icons and figure/background area ratios and color combinations on visual search performance on an LCD monitor. *Displays*, *29*(3), 237–242.

Huang, K.-C., & Chiu, T.-L. (2007). Visual search performance on an LCD monitor: effects of color combination on figure and icon background, shape of icon, and line width of icon width of icon border. *Perceptual and Motor Skills*, *104*(2), 562–574.

Johnson, A. (2013, August). Hollow Icons. *Medium*. Retrieved from https://medium.com/design-ux/a93647e5a44b

Kacmar, C. J., & Carey, J. M. (1991). Assessing the usability of icons in user interfaces. *Behaviour & Information Technology*, *10*(6), 443–457.

Kholmatova, A. (2013, September 17). Optimizing UI Icons for Faster Recognition. *Boxes and Arrows*. Retrieved from http://boxesandarrows.com/optimizing-ui-icons-for-faster-recognition

Lin, R., & Kreifeldt, J. G. (1992). Understanding the Image Functions for Icon Design. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 341–345).

Lindberg, T., & Näsänen, R. (2003). The effect of icon spacing and size on the speed of icon processing in the human visual system. *Displays*, *24*(3), 111–120.

Ling, J., & Van Schaik, P. (2002). The effect of text and background colour on visual search of Web pages. *Displays*, *23*(5), 223–230.

Microsoft. (2013, November 16). Icons. *Windows Developer Center - Windows Store Apps*. Retrieved from http://msdn.microsoft.com/library/windows/apps/dn439350.aspx

Moyes, J. (1994). When Users Do and Don't Rely on Icon Shape. In *Conference companion on Human factors in computing systems*. ACM.

Näsänen, R., Karlsson, J., & Ojanpää, H. (2001). Display quality and the speed of visual letter search. *Displays*, *22*(4), 107–113.

Näsänen, R., & Ojanpää, H. (2003). Effect of image contrast and sharpness on visual search for computer icons. *Displays*, *24*(3), 137–144.

Näsänen, R., Ojanpää, H., & Kojo, I. (2001). Effect of stimulus contrast on performance and eye movements in visual search. *Vision Research*, *41*(14), 1817–1824.

Ojanpää, H., & Näsänen, R. (2003). Effects of luminance and colour contrast on the search of information on display devices. *Displays*, *24*(4), 167–178.

Otto, M., Thornton, J., & Bootstrap Contributors. (n.d.). Bootstrap Components. *Bootstrap*. Retrieved February 18, 2014, from http://getbootstrap.com/components

Passini, S., Strazzari, F., & Borghi, A. (2008). Icon-function Relationship in Toolbar Icons. *Displays*, *29*, 521–525.

Pixeden. (2013a, June 13). Tab Bar Icons IOS 7. *Pixeden*. Retrieved February 18, 2014, from http://www.pixeden.com/media-icons/tab-bar-icons-ios-7

Pixeden. (2013b, July 12). Tab Bar Icons IOS 7 Vol 2. *Pixeden*. Retrieved February 18, 2014, from http://www.pixeden.com/media-icons/tab-bar-icons-ios-7-vol2

Pixeden. (2013c, September 6). Tab Bar Icons IOS Vol 3. *Pixeden*. Retrieved February 18, 2014, from http://www.pixeden.com/media-icons/tab-bar-icons-ios-7-vol3

Rogers, Y. (1989). Icons at the Interface: Their Usefulness. *Interacting with Computers*, *1*(1), 105–117.

Schröder, S., & Ziefle, M. (2008). Effects of icon concreteness and complexity on semantic transparency: Younger vs. older users. In *Lecture Notes in Computer Science* (pp. 90–97).

Sharp, H., Rogers, Y., & Preece, J. (2007). *Interaction Design: Beyond Human-Computer Interaction*. Wiley.

Solomon, B. (2013, August 21). Hollow Icons? A Hollow Argument. *The Fox Is Black*. Retrieved from http://www.thefoxisblack.com/2013/08/21/hollow-icons-hollow-argument/

The Noun Project. (n.d.). *The Noun Project*. Retrieved February 18, 2014, from

> http://thenounproject.com

UX Magazine Staff. (2013, December 20). The Top UX Predictions for 2014. *UX*

> *Magazine*. Retrieved from http://uxmag.com/articles/the-top-ux-predictions-for-
>
> 2014

Vieira, N. (n.d.). browser. *GitHub*. Retrieved March 31, 2014, from

> https://github.com/fnando/browser

w3schools.com. (2014, February). OS Platform Statistics. *w3schools.com*. Retrieved

> April 1, 2014, from http://www.w3schools.com/browsers/browsers_os.asp

Wong, C. (2013, August 27). How Does the Usability and Cognitive Load of Hollow

> Icons Compare to Filled Icons? *Quora*. Retrieved from
>
> https://www.quora.com/Usability/How-does-the-usability-and-cognitive-load-of-
>
> hollow-icons-compare-to-filled-icons/answer/Caesar-Wong?srid=t7YV&share=1

Wroblewski, L. (2013). Favorites: Which Icon Is Easier to Understand? *Polar Polls*.

> Retrieved from http://polarb.com/polls/119730

Zurb Inc. (n.d.). Foundation Icon Fonts. Retrieved February 18, 2014, from

> http://zurb.com/playground/foundation-icons