Thousands of On-Line Observers is Just the Beginning

Nathan Moroney

HP Laboratories
HPL-2009-59

Keyword(s):
visual experiments, web-based tools, distributed tasks, on-line experiments

Abstract:

Web-based or on-line experiments are still a relatively new research topic. Laboratory or highly controlled experiments are, for many reasons, the preferred methodology for visual experiments. However recent experiments suggest that on-line experiments have some unique properties and advantages that may in some cases outweigh or offset their disadvantages. This paper will consider on-line experiments from both a general and a narrow perspective. Specifically a range of specific experiments, and on-line tools, will be considered from a broad vantage and possible themes or considerations for these experiments will be considered.

External Posting Date: March 21, 2009 [Fulltext] Approved for External Publication
Internal Posting Date: March 21, 2009 [Fulltext]

To be presented at the Human Vision and Electronic Imaging XIV Conference, San Jose, CA and to be published in SPIE Proceedings Volume 7240.

[1]Thousands of On-Line Observers is Just the Beginning

Nathan Moroney
Hewlett-Packard Laboratories, Palo Alto, CA, USA

ABSTRACT

Web-based or on-line experiments are still a relatively new research topic. Laboratory or highly controlled experiments are, for many reasons, the preferred methodology for visual experiments. However recent experiments suggest that on-line experiments have some unique properties and advantages that may in some cases outweigh or offset their disadvantages. This paper will consider on-line experiments from both a general and a narrow perspective. Specifically a range of specific experiments, and on-line tools, will be considered from a broad vantage and possible themes or considerations for these experiments will be considered.

Keywords: visual experiments, web-based tools, distributed tasks, on-line experiments

"Since the beginning, it was the same. The only difference, the crowds are bigger now." Elvis Presley

"The future belongs to crowds." Don Delillo in Mao II

1. INTRODUCTION

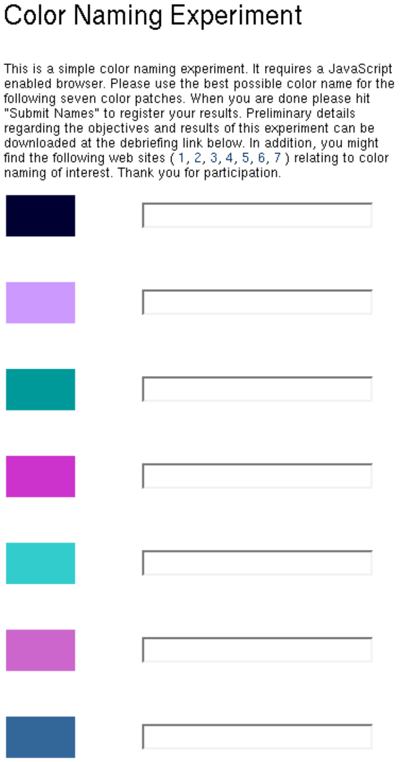
What is an on-line observer? In the context of this paper an on-line observer is an experimental participant that uses a communications network to provide data. This paper will be limited to visual tasks using some form of display but certainly the general principles discussed will apply to other modalities, such as audio. The thousands listed in the title is a direct reference to one of the primary advantages of using a communication network, such as the internet, to conduct experiments: the potential for a much larger number of observers. However in the case of online experiments it not just about how many observers – other factors are relevant.[1] This paper will begin by considering five ongoing experiments and three on-line tools. Certainly online experiments have enough issues without complicating the discussion with online tools but they will be considered because of the underlying implications of 'online participation'. Finally eight specific considerations will be covered with respect to web-based experiments and observers, most of which are not about the number of participants.

2. ON-LINE EXPERIMENTS

As examples of on-line experiments, consider the following five experiments. First is the unconstrained color naming experiment which consisted of seven color patches in which participants were instructed to provide the "best color names possible" for the patches and is shown in Figure 1. The patches were randomly selected from a six by six by six sampling in the device red, green and blue space. The colors were displayed on a white background and an optional comment box was provided. To date over 4,000 volunteers have provided over 30,000 color names. The result of the experiment is seven text strings and seven red, green and blue values per participant. The experiment is also being conducted in over 20 languages but English remains the language with the largest amount of participants. These results have been favorably compared to the laboratory results of Berlin and Kay, Sturges and Whitfield, and Boynton and Olson.[2,3]

[1]

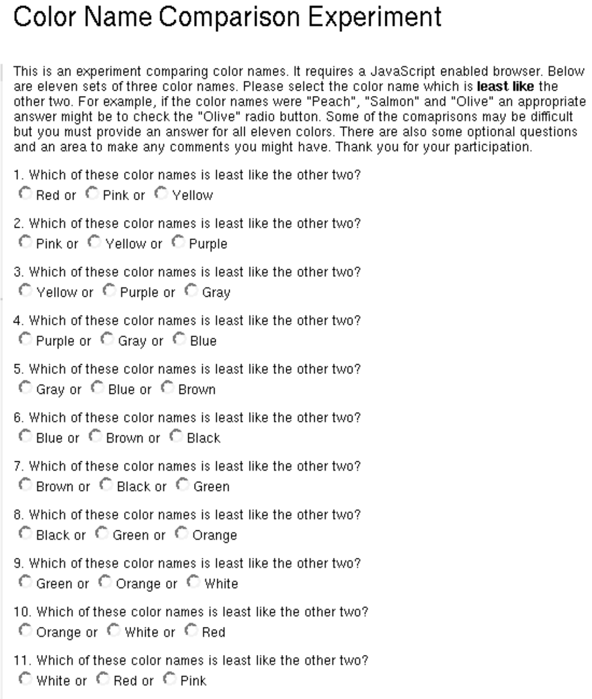
The second experiment, related to the first, is a non-repeating random walk multi-dimensional scaling of the text strings corresponding to the eleven basic color names.[4,5] In this case the full set of triadic comparisons is distributed over the participants. The resulting dissimilarity matrix is sent through a multi-dimensional scaling routine and the spatial arrangement of the names is computed. This experiment did not include colored patches and in this case the task is purely an abstract text only effort. The two dimensional arrangements of the color names are shown to follow a recognizable hue circle although the white is closer to yellow and black is closer to blue than is typically the case for opponent color spaces, such as CIELAB or CIECAM02. Figure 2 shows a screen capture of this experiment.



Color Naming Experiment

This is a simple color naming experiment. It requires a JavaScript enabled browser. Please use the best possible color name for the following seven color patches. When you are done please hit "Submit Names" to register your results. Preliminary details regarding the objectives and results of this experiment can be downloaded at the debriefing link below. In addition, you might find the following web sites (1, 2, 3, 4, 5, 6, 7) relating to color naming of interest. Thank you for participation.

Fig. 1. Unconstrained color naming experiment.



Color Name Comparison Experiment

This is an experiment comparing color names. It requires a JavaScript enabled browser. Below are eleven sets of three color names. Please select the color name which is **least like** the other two. For example, if the color names were "Peach", "Salmon" and "Olive" an appropriate answer might be to check the "Olive" radio button. Some of the comaprisons may be difficult but you must provide an answer for all eleven colors. There are also some optional questions and an area to make any comments you might have. Thank you for your participation.

1. Which of these color names is least like the other two?
○ Red or ○ Pink or ○ Yellow

2. Which of these color names is least like the other two?
○ Pink or ○ Yellow or ○ Purple

3. Which of these color names is least like the other two?
○ Yellow or ○ Purple or ○ Gray

4. Which of these color names is least like the other two?
○ Purple or ○ Gray or ○ Blue

5. Which of these color names is least like the other two?
○ Gray or ○ Blue or ○ Brown

6. Which of these color names is least like the other two?
○ Blue or ○ Brown or ○ Black

7. Which of these color names is least like the other two?
○ Brown or ○ Black or ○ Green

8. Which of these color names is least like the other two?
○ Black or ○ Green or ○ Orange

9. Which of these color names is least like the other two?
○ Green or ○ Orange or ○ White

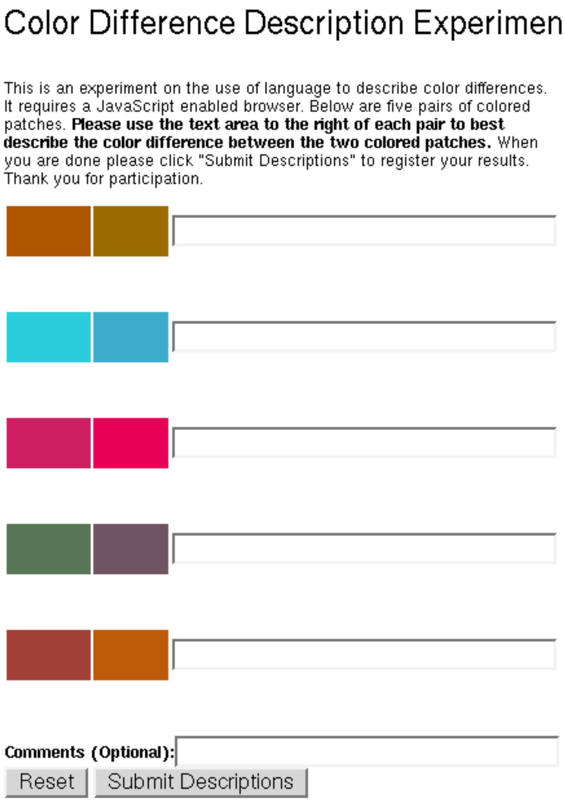
10. Which of these color names is least like the other two?
○ Orange or ○ White or ○ Red

11. Which of these color names is least like the other two?
○ White or ○ Red or ○ Pink

Fig. 2. Color name comparison experiment using triads of color names only generated using a non-repeating random walk.

Third is a color difference description experiment consisting of seven randomly generated but constrained pairs of color patches.[6,7] As can be seen in Figure 3 the task is then to provide a description of the color differences. This experiment is not limited to the six by six by six sampling of the first experiment but is based on a finer sampling of colors. However the task for participants is again unconstrained in that they are asked to 'best describe the differences' between the colors. There are few comparable laboratory results to compare to but again the frequency of words and the rough correspondence to the device red, green and blue values can be performed. The results show interesting similarities and differences to the first experiment – describing color differences is similar but not the same as naming colors. For instance the basic color names are considerably more frequent in difference description than they are in unconstrained naming. The term 'lime' is not a top 100 term for difference description but it is a top ten term for unconstrained color naming. Interestingly the term 'more' is six times more frequent than the term 'less'. Color differences would seem to be described as a relative judgment but one which is primarily about which color has more of a given feature or property.

Fourth is an unconstrained image quality evaluation experiment.[8] In this case one of X images is displayed and volunteers are asked to provide an unconstrained textual description of the quality of the image. A screen shot of this experiment is shown in Figure 4. In this case additional optional demographic questions were also asked, although note that these are also distributed in as much as no single participant is asked all of the optional demographic questions. Analysis of these results is more complex and again there are limited laboratory benchmarks. However the recurrent frequency analysis and concordance results are informative. Full results have not been previously published but it is interesting to note that one of the top ten lemmas for image quality descriptions is 'color' and other top terms are 'dark', 'contrast', 'composition', 'tones' and 'light'.



Color Difference Description Experiment

This is an experiment on the use of language to describe color differences. It requires a JavaScript enabled browser. Below are five pairs of colored patches. **Please use the text area to the right of each pair to best describe the color difference between the two colored patches.** When you are done please click "Submit Descriptions" to register your results. Thank you for participation.

Comments (Optional):

Fig. 3. Color difference description experiment.

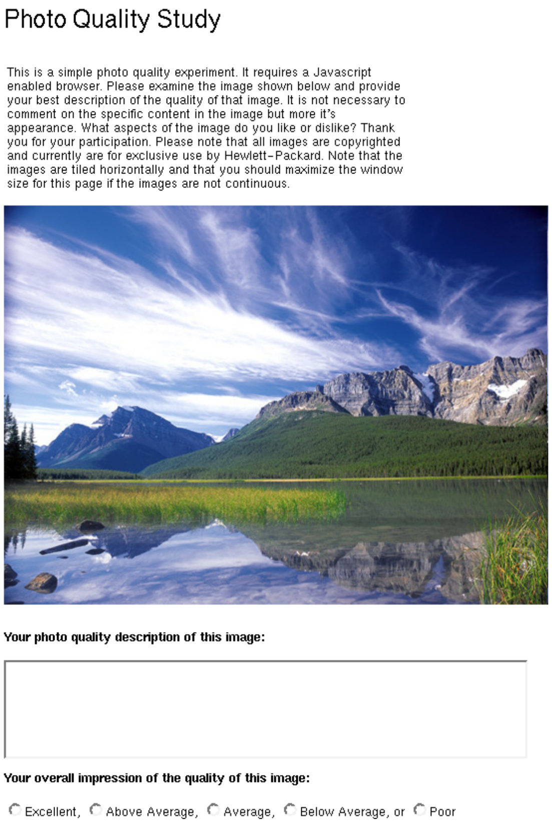


Photo Quality Study

This is a simple photo quality experiment. It requires a Javascript enabled browser. Please examine the image shown below and provide your best description of the quality of that image. It is not necessary to comment on the specific content in the image but more it's appearance. What aspects of the image do you like or dislike? Thank you for your participation. Please note that all images are copyrighted and currently are for exclusive use by Hewlett-Packard. Note that the images are tiled horizontally and that you should maximize the window size for this page if the images are not continuous.

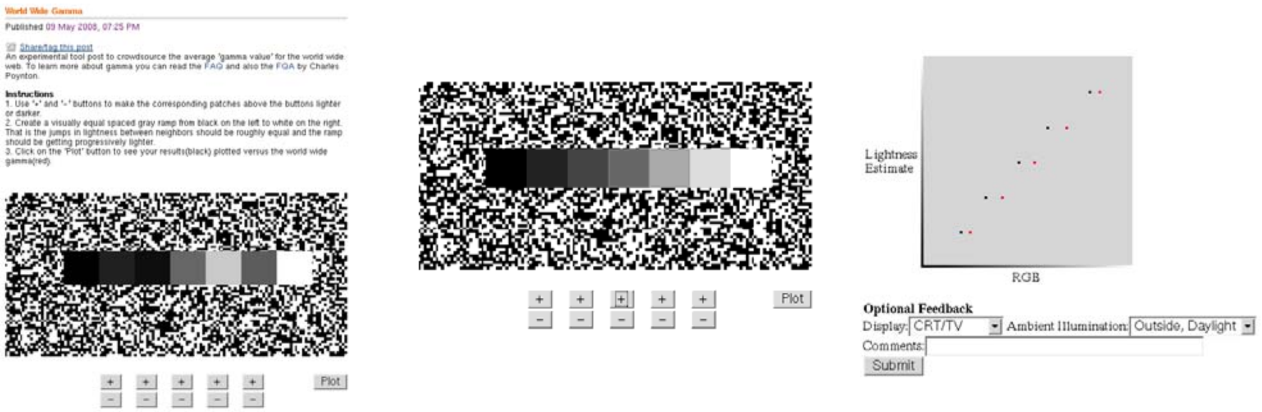
Your photo quality description of this image:

Your overall impression of the quality of this image:

○ Excellent, ○ Above Average, ○ Average, ○ Below Average, or ○ Poor

Fig. 4. Image quality description experiment.

Finally is the initial effort to infer an aggregate or overall average tone curve for displays on the internet. This experiment is called the "world wide gamma"[9] and consists of a white and black anchor on a black and white pattern. Between the anchors are six variable lightness patches which participants can adjust to create an approximately uniform lightness step ramp. Previous laboratory results[10] show that this spatial arrangement and task of lightness partitioning to be an efficient means to infer an observer's lightness scale. Given a known result for an idealized display and observer it is then possible to compare the average results to the ideal. A screen shot of an initial screen view with a randomized ramp is shown in Figure 5(a) and the result of performing the visual task is shown in Figure 5(b). This experiment differs from the previous four in that it is purely an adjustment task but again the potential to reach hundreds of participates is essential. Preliminary results for this experiment are shown in Figure 5(c).



World Wide Gamma

Published 09 May 2008, 07:25 PM

Share/tag this post

An experimental tool post to crowdsource the average 'gamma value' for the world wide web. To learn more about gamma you can read the FAQ and also the FGA by Charles Poynton.

Instructions

1. Use '+' and '-' buttons to make the corresponding patches above the buttons lighter or darker.
2. Create a visually equal spaced gray ramp from black on the left to white on the right. That is the jumps in lightness between neighbors should be roughly equal and the ramp should be getting progressively lighter.
3. Click on the 'Plot' button to see your results(black) plotted versus the world wide gamma(red).

Optional Feedback

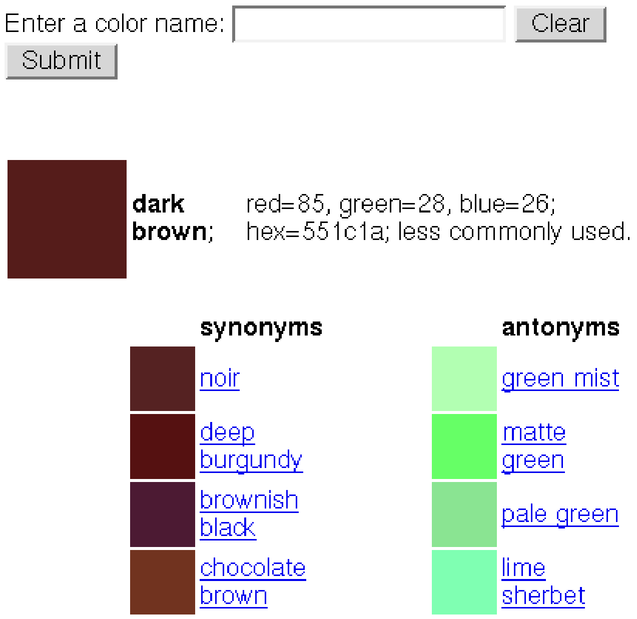
Display: CRT/TV Ambient Illumination: Outside, Daylight

Comments:

Fig. 5(a) on left showing the world wide gamma, 5(b) in the center showing adjusted ramp and 5(c) the graphical results for the current display in black and the average of over 600 displays in red.

3. ON-LINE TOOLS

On-line tools are those dedicated applications that are based directly or indirectly on the results of the on-line experiments. The tools are included in this paper to complement the discussion of the experiments and to transition to the considerations section. Specifically, one of the tools is both a tool and an experiment – it is useful for a specific task and it generates specific experimental data. The first tool is shown in Figure 6 and is the on-line color thesaurus.[11,12] This tool allows a user to query a pre-formatted version of the unconstrained color naming data. If the color name is found then the corresponding red, green and blue values are returned, along with synonyms and antonyms. This tool provides a simple and direct means to search for colors by name. However this tool cannot exist without the large scale color naming database generated from experiment one. To date over 140,000 color names have been served by this on-line color thesaurus.



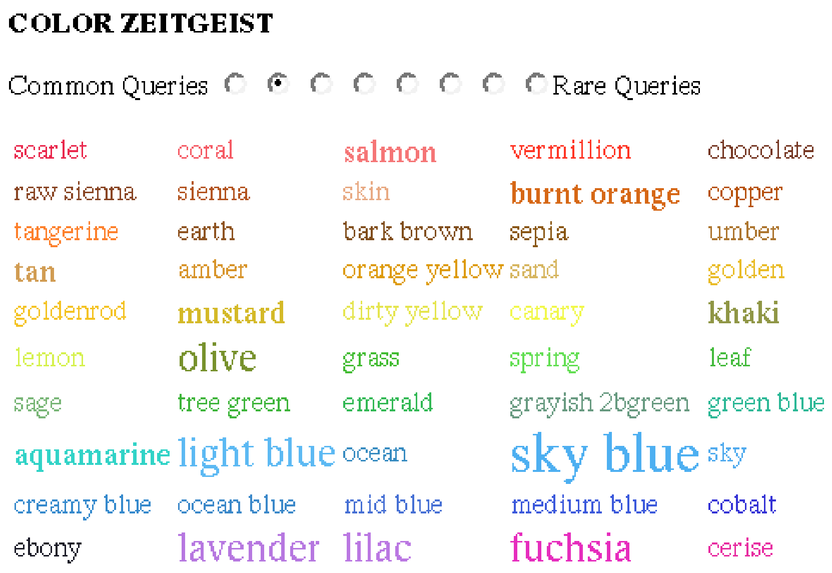
dark brown; red=85, green=28, blue=26; hex=551c1a; less commonly used.

synonyms	antonyms
noir	green mist
deep burgundy	matte green
brownish black	pale green
chocolate brown	lime sherbet

Figure 6. The online color thesaurus showing the results for the query 'dark brown'.

The second tool is a color zeitgeist and is shown in Figure 7. This tool is based on the data generated by the first tool or the on-line color thesaurus.[13] Specifically, the color names shown in the zeitgeist are the most frequently queried color names for the thesaurus. The user can now visually explore a large number of color names in a simple hue sorted lexical cloud. This tool is based on the first tool and simply demonstrates how tools can create useful data. In this case the data is the current most sought color names, which is not the same as the underlying color naming database. Note that this tool is also directly linked to the color thesaurus in that when a specific color name is clicked the specific results shown are the color values, synonyms and antonyms for that name.

The third and final tool for discussion is the Italian color thesaurus.[14] This tool is of interest in that while data collection for color naming has been ongoing in over 20 languages, this was the first time non-English results were generated, formatted and provided online. A sample screen shot for this is shown in Figure 8(a). In this case the underlying Italian color naming database is smaller than the English color naming database. Therefore a small but significant modification was made to the tool. In the cases where a color name is missing a small 'color remote', shown in Figure 8(b), is provided so that the user may adjust a colored patch to create a corresponding red, green and blue display value for their color query. In this way the underlying Italian color naming database is efficiently populated through instance-based harvesting. An additional small but significant modification is the optional ranking of the color names in the Italian color naming database. This allows an efficient validation of specific color names for possible outliers. The Italian color thesaurus is primarily a tool, but it is also two experiments. Considered broadly, on-line experiments extend to adaptive, highly distributed tools.

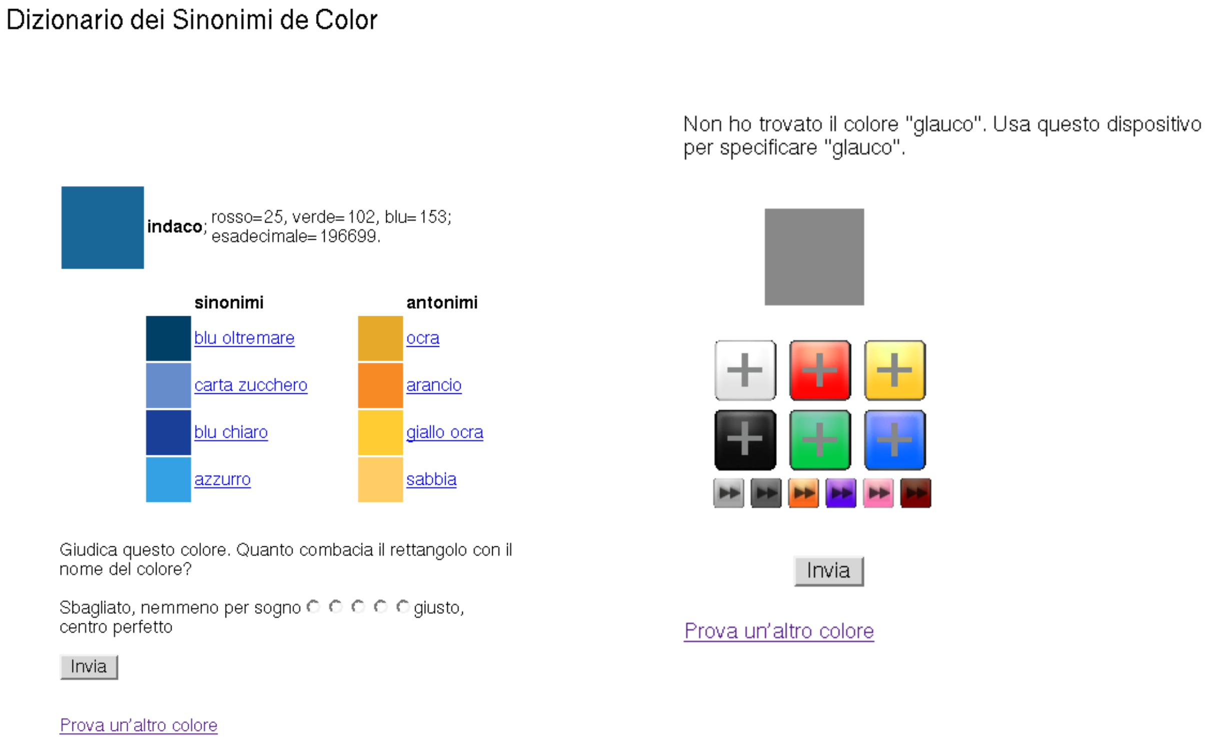


COLOR ZEITGEIST

Common Queries ○ ● ○ ○ ○ ○ ○ ○ Rare Queries

scarlet coral salmon vermillion chocolate
raw sienna sienna skin burnt orange copper
tangerine earth bark brown sepia umber
tan amber orange yellow sand golden
goldenrod mustard dirty yellow canary khaki
lemon olive grass spring leaf
sage tree green emerald grayish 2bgreen green blue
aquamarine light blue ocean sky blue sky
creamy blue ocean blue mid blue medium blue cobalt
ebony lavender lilac fuchsia cerise

Fig. 7. The online color zeitgeist showing a hue sorted lexical cloud for fairly common color name queries.



Dizionario dei Sinonimi de Color

indaco; rosso=25, verde=102, blu=153; esadecimale=196699.

sinonimi	antonimi
blu oltremare	ocra
carta zucchero	arancio
blu chiaro	giallo ocra
azzurro	sabbia

Giudica questo colore. Quanto combacia il rettangolo con il nome del colore?

Sbagliato, nemmeno per sogno ○ ○ ○ ○ ○ giusto, centro perfetto

Invia

Prova un'altro colore

Non ho trovato il colore "glauco". Usa questo dispositivo per specificare "glauco".

Invia

Prova un'altro colore

Fig. 8(a) on the left showing the online color thesaurus in Italian and the ‘color remote’ for instance-based harvesting of missing color names.

4. CONSIDERATIONS: LABORATORY VERSUS ON-LINE

4.1 1. Lots of Observers

Certainly a key benefit of online experiments and tools is the ability to reach a larger, potentially more diverse population of observers and participants. The difference is significant and is in fact an increase of more than one order of magnitude. The early years of strictly laboratory experiments for the author were limited to tens of participants. In comparison, the online experiments and tools have been used by hundreds, thousands and tens of thousands of people. It is worth noting that this increase in the number of participants has not required substantially more effort.

4.2 Distributed Design

A second consideration implicit in the discussion of the experiments is the use of a distributed design versus an exhaustive design. That is for on-line experiments no single observer completes the entire evaluation or scaling of all the stimuli. This was initially part of the design for two reasons. First was the possibility that for increasingly longer and more difficult tasks, the volunteer participation would drop. Second was the benefit of minimizing the potential influence of any single participant. Previous analysis of the online color naming experiment in English showed a roughly 4% disruptive participants rate. A distributed design allows both an efficient division of labor and limits systematic global deviations. The trade-off is one of breadth over depth. An exhaustive laboratory experiment provides a more comprehensive understanding of the task over a larger range of stimuli, but for a smaller pool of observers.

4.3 Ambiguity

A third consideration is one of minimal constraints versus precision definition. Many laboratory experiments require systematic control of stimuli, presentation and viewing. This precise definition allows a targeted and controlled evaluation of a specific phenomena or mechanism. This is likely a result of the historical focus of psychometric techniques for threshold and detection experiments. However for categorical evaluation, naming tasks or natural language processes these rigid approaches may be less critical. In the case of color naming for example, it would seem that to have a detailed and aggregate understanding of robust naming of colors that the inclusion of real world sources of variability is in fact representative of underlying cognitive processes in real world environments. An average person rarely has a need to perform a threshold or detection task but frequently uses natural language for lexical categorization – and they generally do so outside of a laboratory setting. A majority of the online experiments have provided as general as possible instructions and as a result minimal constraints.

4.4 Hypotheses versus Training

A forth consideration of online experiments is a shift in the post-processing and analysis of the results from hypothesis testing given noise to adversarial machine learning given noise. Laboratory experiments are often focused on demonstrating statistically significant differences for detection or thresholds or the derivation of perceptual attribute correlates. These experiments are focused on untangling underlying processes and interdependencies in a statistically sound manner. In contrast the online color naming experiment is not related to these types of analyses. However, there are multiple applications of the underlying color naming database and these tend to be suited for providing ground truths for machine learning. Specifically the data from the online color naming experiment can be used for the formulation and testing of machine color naming algorithms.[15] Note however that the sources of noise in the two general cases are different, in a laboratory experiment noise may be due to the difficulty of the task or observer fatigue. For an online experiment there is noise throughout, from displays to hardware to observers to malicious participants.

4.5 Simplicity

The fifth consideration is speculative and likely an author preference but for online experiment or tool there would seem to be a case for simplistic infrastructure versus the specialized infrastructure of a laboratory. This includes task design and software infrastructure. While the latest technology will always provide the largest feature set the trade-off is fewer potential participants and the temptation to add task complexity through features. The original JavaScript for the color naming experiment is still running as is and its simplistic nature has provided implicit constraints on the task and even the data collection. In comparison laboratory experiments often make used of highly specialized devices and software. This will likely be a persistent gap but with advance consideration this may be a way to evaluate which tasks or experiments are best suited for a laboratory experiment versus an online experiment.

4.6 Global and Open-ended

A sixth consideration is a shift in scale for experimental design. Laboratory experiments by their very nature are generally performed according to a fixed schedule in a limited geography. Data is collected, generally within an academic or industrial setting, in order to test a specific hypothesis and then the experiment is concluded. Data collection is also generally limited to a one, or with some fieldwork a few number, of locations. There is limited benefit and often significant expense to continuing to collect data beyond a limited timeframe and geography. For online experiments though such as the image quality description experiment there is no single specific hypothesis to be tested. The resulting data becomes more nuanced and detailed the longer and more broadly the experiment is carried out. This means that experiments can now be global and indefinite.

4.7 Usage as Data

The seventh consideration is the fuzzy boundary between online experiments and online tools. If usage is data then when does a tool stop and an experiment begin? Online visual experiments are sufficiently novel and recent that it is difficult to consider even consider even less rigorous and implicit means of collecting experimental data. If a visual task is useful, educational or even enjoyable is the resulting data necessarily less useful? In the extreme consider an online video game which systematically modulates color differences in order to derive target threshold values. Each step away from the laboratory is a further step away from familiar scientific methods towards potentially larger and more diverse data.

4.8 Mutual Bootstrapping

The final consideration is the capacity for feedback for online experiments and tools. Classical psychometrics is generally performed in an “open loop” manner or in a decontextualized setting. A staircase technique may provide fine adaptive adjustment to a stimulus but this is limited to the determination of a threshold. Likewise completing a scaling task may provide a better familiarity with the defined correlate. In comparison an online experiment or tool can provide alternative forms of feedback and bootstrapping, such as instance based harvesting of the Italian color thesaurus or learning-through-use. For example while the author has performed roughly a half dozen laboratory experiments that made some use directly or indirectly of chroma scales it was not until reviewing the over 30,000 color names that the finding that none of the names included the sub-string ‘chroma’ was made. Likewise one year into the online color naming experiment the author became confident using the color name ‘chartreuse’. While this learning-through-use potentially complicates things it is an intriguing implication of online experiments and tools. Perhaps human observers and machines can mutually bootstrap.

5. CONCLUSIONS

This paper was attempted to provide some additional considerations for online experiments and specifically considered the advantage of a larger number of observers as one of many considerations. The specifics of experiments and a summary of the results were provided for five online experiments and three online tools. For the last tool the suggestion is made that with instance-based harvesting the line separating an experiment and a tool becomes fuzzy. Finally eight specific considerations for online experiments were listed and discussed. These considerations are in summary form scale, distributed design, ambiguity, hypotheses testing versus training, simplicity, global and open-ended nature of the experiments, usage as data and finally mutual bootstrapping.

REFERENCES

[1] Beretta, G. and Moroney, N., “Cognitive Aspects of Color”, HP Labs Technical Report, HPL-2008-94, (2008).

[2] Moroney, N., “Unconstrained web-based color naming experiment”, Proceedings of SPIE - The International Society for Optical Engineering **5008**, pp. 36-46 (2003).

[3] http://www.hpl.hp.com/personal/Nathan_Moroney/color-name-hpl.html

[4] Moroney, N. and Tastl, T., “Multidimensional scaling with non-repeating random paths”, Proceedings of SPIE - The International Society for Optical Engineering **5668**, pp. 20-27 (2005).

[5] http://www.hpl.hp.com/personal/Nathan_Moroney/colorname-similarity.html

[6] Moroney, N., “Color differences without probit analysis”, Proceedings of SPIE - The International Society for Optical Engineering **6494** (2007).

[7] http://www.hpl.hp.com/personal/Nathan_Moroney/cdd-hpl.html

[8] http://www.hpl.hp.com/personal/Nathan_Moroney/photo-quality/photo-quality.html

[9] http://www.communities.hp.com/online/blogs/mostly_color/archive/2008/05/09/HPPost6333.aspx , world wide gamma.

[10] Moroney, N., "Factors affecting lightness partitioning", Proceedings of SPIE - The International Society for Optical Engineering **4663**, pp. 35-42 (2002).

[11] Moroney, N., *The Color Thesaurus*, June 2008 edition, magcloud.com. (2008).

[12] http://www.communities.hp.com/online/blogs/mostly_color/archive/2007/10/30/HPPost4914.aspx , online color thesaurus.

[13] http://www.communities.hp.com/online/blogs/mostly_color/archive/2008/02/25/HPPost5795.aspx , online color zeitgeist.

[14] http://www.communities.hp.com/online/blogs/mostly_color/archive/2008/07/30/un-dizionario-dei-sinonimi-dei-colori.aspx , online color thesaurus in Italian.

[15] Moroney, N., Obrador, P. and Beretta, G., “Lexical Image Processing”, Proceedings of IS&T/SID 16th Color Imaging Conference (2008).