//

DID IT WORK?

EVALUATIVE RESEARCH METHODS FOR GRAPHIC DESIGN

Mike Zender, Bing Han, Oscar Fernández
University of Cincinnati
mike.zender@uc.edu

ABSTRACT

Did that design work? It's an evaluation question yet there appears to be little evaluative graphic design research to answer it. None of the 40 research methods covered in Brenda Laurel's book Design Research Methods and Perspectives described evaluative methods (Laurel, 2003). Yet designers should know, and clients will eventually demand to know, did that design work?

This paper reports on methods used for two related evaluative design research studies conducted at the University of Cincinnati. While the results of the studies were often definitive, the research methods were not. In response to issues uncovered during the execution of the research protocols, an additional study was conducted to explore not the subject of the original research, but the research methods used for it. The methodology of those studies is discussed here as a means of advancing evaluative research methods for graphic design, and improving the quality of design work.

Keywords: design research, evaluation, methods.

INTRODUCTION

Did it work? It's a simple question that few graphic designers are prepared to answer except in the crudest ways. 'Did it work' is often equated to 'did it sell.' Design output is measured monetarily. But did the design work, that is, 'did the design produce it's intended outcome' by non-monetary means such as functional response, attitudinal change, or cultural impact. This is a more challenging, and perhaps more significant question. A recent study found that an icon warning about tire inflation pressure (see Figure 1), mandated by law, was not understood by 60% of drivers: 46% could not even recognize it as a tire (Bird, 2011). Is design research prepared to respond to such performance issues? It appears not to be. A quick review of design research activity such as that described in Brenda Laurel's 2003 book features no methods that were evaluative.

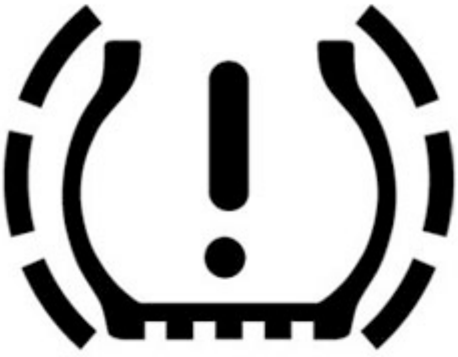


Figure 1.

Tire pressure warning icon

Evaluative research as used here is rigorous study to specifically answer questions of design performance. The question can be as casual as 'which symbol do you prefer?' or as significant as 'which symbol most quickly and effectively warns you not to cross the street?' Measures of outcome effectiveness, particularly outcomes that directly affect the user's well being as in the example just given, call for rigorous evaluative methods. Our studies suggest that while rigorous evaluative design research is possible there are many issues that still need to be addressed to develop a robust evaluative design research practice. These issues range from detailed concerns such as how to define success (metrics), gauge measurement (validity and reliability), and integrate analytical methods (statistics and language), to more expansive concerns such as selecting data types (qualitative vs. quantitative), research approaches (top-down vs. bottom-up), and proper levels to observe impact (micro vs. macro).

Consider levels of impact. Design impact can be evaluated at a macro level: did the iPhone impact the economy of rural farmers in Africa? At the macro level, evaluative design research can be considered as part cultural research. James Davidson Hunter (2010) in his book , "To Change the World" outlines eleven propositions relative to cultural change: seven about culture and four about cultural change. "PROPOSITION TWO: CULTURE IS A PRODUCT OF HISTORY" defines cultural meanings as accumulations of meaning over long periods of time. By this measure, analysis of effectiveness of a current design is clearly a long-term process. He describes overlapping networks of leaders, resources, institutions that provide the dynamics of cultural change (p. 44). Clearly, design operates within culture, and may be argued to have an impact on culture, but due to the complexity and scope of networks and factors Davidson describes, assigning design's role in cultural change is challenging.

Design's influence can also be considered at a micro level: did the symbol direct the person to the help desk? At a micro level, evaluative design research can measure changes in attitude, intentions, and behaviors. An HIV/AIDS prevention curriculum designed by a team led by the author has been proven to inform and change behavioral intentions in thousands of school children in South Africa. At this more micro level design's impact is more easily measured because the influencing forces are fewer and more localized.

The macro/micro question is but one issue of interest in evaluative design research. Each of the issues are interesting, but only a few can be addressed in this paper which focus primarily on evaluative research methods at the micro level: defining success; measurement; and analyzing. While the more expansive issue of how to integrate qualitative and quantitative data is touched upon briefly, the micro level issues focused on here are elemental and can be used as building blocks for evaluative studies at any level. While the methods described here are applied to graphic design, they could have wider application across both levels and design disciplines.

EVALUATIVE RESEARCH OVERVIEW

While design research is not well defined (Erlhoff & Marshall, 2008, p. 332), much of the research community in the United States defines research as 'systematic investigation that aims to produce generalizable knowledge.' (reference: Collaborative Institutional Training Initiative, providing research training for 1,130 institutions and over 1.3 million individuals, see: https://www.citiprogram.org/aboutus.asp?language=english) Evaluative research, though aimed at evaluating specific objects, can also produce generalizable results in the form of principles based on past performance to guide future actions. Evaluative research is defined in various ways by Friedman & Wyatt in their book "Evaluative Methods in Biomedical Informatics" (2006, p. 24). Freidman and Wyatt's book, aimed as it is toward evaluation of information resources in medicine, can be used to help define evaluative research for the communication in graphic design. The following definition, adapted from Friedman quoting House, is used to define evaluative design research in this paper: a process leading to a settled opinion that something about a design is the case, usually leading to a decision to act a certain way. This definition is useful in several ways. It describes a means to direct future actions. It stipulates that evaluative design research be integrated into the design process and inform design practice as it unfolds. It suggests that evaluative design research might be driven by problem framing and user need. Philosophically and epistemologically, this definition of evaluative design research is integrated into design as part of the on-going process of reflective practice and converting tacit to explicit knowledge.

Freidman and Wyatt's book cited above detailed key issues in evaluative research. Foremost of these was measurement. Reason suggests that in order to evaluate something you need a standard of measure by which to evaluate. All measurements have errors. To use a measure to guide action that measure must be both reliable and valid. Friedman states: "…reliability is the degree to which measurement is consistent or reproducible. A measurement that is reasonable reliable is measuring something. Validity is the degree to which the something is what the investigator wants to measure" (Friedman, p. 114). As a first step, evaluative research has to decide what to measure. This means identifying what is meaningful (validity). Evaluative research also has to define relevant features of the object being measured so that so that they can be manipulated. It must develop a definition of success, and the features of that definition as well. A fuzzy definition will likely be difficult to measure. Then evaluative research must establish a threshold of measurement that will qualify as success. How 'successful' does a design need to be to qualify as achieving the design's aim? Evaluative research must also confront who referees specific results. All measurement has an element of subjectivity. How to balance that subjectivity is critical to reliability. Design, with little experience in evaluation, has a vacuum in measurement standards (reliability). The studies described below have encountered many measurement challenges.

In addition to measurement issues are issues of data analysis. Simple techniques, such as forming categories and measuring averages, are sufficient to describe data. The most common statistical description is the arithmetic mean (the sum of items divided by the number of items). But to draw inferences from data more sophisticated statistical methods are needed. These statistical methods are largely foreign to designers and seldom used in design practice. Yet to infer from a singular circumstance: a sample, to a general situation: a population, requires statistical modeling. Design knowledge that remains particular is of limited use, whereas generalizable knowledge based on statistical inferences can guide future action. Incorporating statistical analysis into evaluative design research is challenging. However, while statistical analysis may be foreign to design, visual analysis is integral to design and an effective means of analyzing large bodies of data (Card, Mackinlay, & Schneiderman, 2003). The studies reported below explore how to derive important insights from data.

In the following we describe how we addressed evaluative research issues of measurement and data analysis in two evaluative studies and explored measurement stability in one measurement study.

TWO EVALUATIVE DESIGN STUDIES

This paper reports on the methods used for two evaluative design research studies: one measuring the comprehension of 54 universal healthcare symbols (see Figure 2) prior to their deployment in healthcare facilities in Tanzania, the other comparing comprehension of a subset of 4 of the 54 symbols that had varying contextual details; and one measurement study: 14 design students evaluated study responses to determine the effect of the number of evaluators on reliability of measurement.

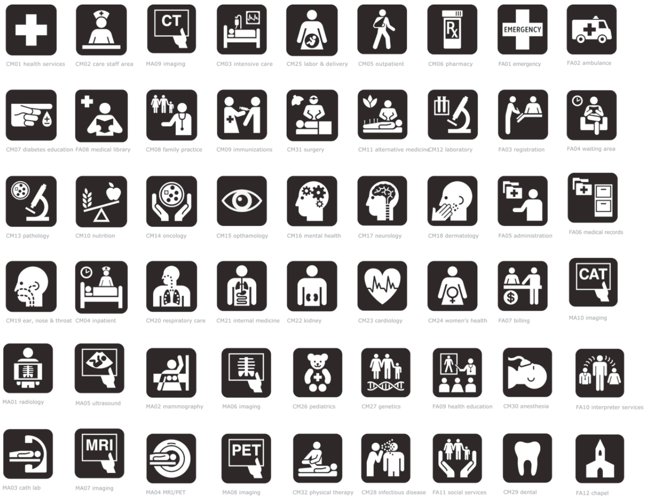


Figure 2.

54 icons of the Universal Healthcare Symbol System (UHCS), developed by both professional and student designers under guidance of Hablamos Juntos and SEGD.

Both evaluative studies focused on symbol comprehension. Both studies followed the ISO defined comprehension test protocol (ISO, 2007), with minor exceptions such as the number of symbols per page. The fundamental research question for the first study was: will the 54 universal medical symbols (referred to as icons in what follows) work in Tanzania? We hypothesized that there were two primary drivers of correct icon comprehension: medical literacy (knowledge of medical subject matter) and cultural perspective. Our research question then became: to what degree do medical literacy and cultural perspective affect the interpretation of the 54 medical icons? Our question contained two key issues: 1. medical literacy and 2. cultural perspective. We therefore designed a comparative open-ended comprehension study to be conducted in both Tanzania and in the United States.

To evaluate the effect of medical literacy on comprehension the test in each country was divided evenly into two cohorts: those with standard and those with advanced medical literacy. We defined 'standard' medical literacy as that of anyone without advanced medical training or education. We reasoned that icons understood by high medically literate but not by standard medically literate failed due to medical knowledge, not culture. To measure the effect of cultural perspective on comprehension we tested in two countries. We reasoned that icons understood by high medically literate in the USA but not by high medically literate in Tanzania might have failed more due to culture than difference in medical knowledge. The first comprehension study measured comprehension of a sample standard population (N = 20) in rural Tanzania compared to comprehension of the same icons in the same location by a sample medically literate population (N = 20). The corresponding USA study involved a similar sample standard population (N = 45) and sample medically literate population (N = 45). Each subject in both countries was shown the same icon survey that asked two questions: what do you think this icon means, and what would you do in response to it? The hypotheses for this study was that many of these 'universal' icons would not be adequately comprehended in Tanzania (ISO standard = 85% correct). The study was done to support our application questions: which icons work in Tanzania, why or why not, and what specifically should we do to make icons that will work?

The fundamental research question of the second study was: how much will changes in symbol details affect interpretation of an icon. In this context a single visualized object such as a book is a symbol, while several symbol objects contained together: a book, a shelf, a red cross, a man holding a book, form an icon that represents a referent such as "medical library" (see Figure 3). This study compared comprehension of alternate versions of 4 of the 54 universal healthcare symbols noted above. Half of the subjects (N = 55) responded to icons with an added contextual symbol in the icon, while half of the subjects (N = 55) responded to icons without the added symbols or with a less specific added symbol. The hypothesis of the second study was that fewer and/or less specific contextual symbol content would reduce comprehension.

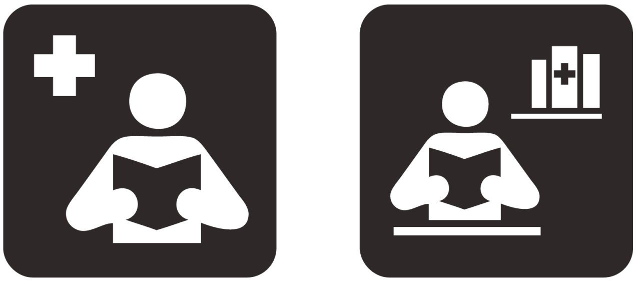


Figure 3.

"medical library" icons: icon without added symbols on the left, icon with added symbols on the right, the icon on the right was comprehended significantly better

Our experience with the evaluative studies led us to question our measurement methods. Specifically, we were concerned about the reliability of our evaluation, involving as it did subjective decisions of individuals. We hypothesized that increasing the number of evaluators would reduce individual subjectivity, increase scoring accuracy, and improve reliability. We conducted a measurement study to analyze this. 14 design graduate students and 1 faculty, each with expertise in symbol design, independently scored 10 of the evaluative research surveys. Their scoring results were analyzed by comparing the mean scores and variances for each icon by different numbers of evaluators: 1, 3, 5, 10 and 15 evaluators. This study was done to answer our application question: how many evaluators are needed for reliability? We found that increasing the number of evaluators did not increase the stability of the results. In fact, except for the score of an individual icon, we found little variance between 3, 5 and 14 evaluators. Our finding on stability of a small expert panel is consistent with findings for similar panels elsewhere (Akins, Tolson, & Cole, 2005).

The result of our measurement study was instructive: a small panel of scorers was reliable. The results of both evaluative studies were definitive: many icons failed to be adequately comprehended; some icons failed for identifiable cultural reasons; medical literacy did increase comprehension of some icons; the icons with added objects or more specific objects did improve comprehension. But the process was not definitive. The methods used in the study revealed many issues in evaluative design research that need further development, described in the following section.

EVALUATIVE METHODS: PROBLEMS AND QUESTIONS

DEFINING SUCCESS: DEFINITION OF ICON REFERENTS

As noted above, evaluative studies seek to measure something. The research question for both evaluative studies focused on issues affecting icon comprehension. Consequently, our objects were icons and our measurement instrument was comprehension. Our first task in measurement was to develop a definition of comprehension success for each icon. This seemed simple because each icon was based on a definite concept or idea or object called a referent. Successful icons communicate their referents. However, what appeared to be simple proved otherwise. While scoring will be discussed in more detail in the next section, complications arose when applying definitions during scoring because few correct answers used the exact referent word. Many subjects used synonyms of the referent in their answers. This was further complicated because many referents are more than one word. Of the 54 health care icons, 33 of the referents were a single-word: outpatient, pharmacy, surgery, laboratory, ambulance for example; while 21 referents contained two or more words: medical library, diabetes education, mental health, medical records, physical therapy for example. The 110 responses to the added symbol ‘medical library’ icon were typical: 13 answers were precise to the referent: “medical library” or “library of medicine,” but most correct responses included synonym combinations that were judged correct, including: “medical books to read,” “reading books about diseases,” “information about medicine here,” and “library/bookstore medical books.”

As noted previously, each answer had two parts: definition, “what does it mean,” and action, “what would you do.” The ISO protocol calls for these two questions, recognizing the importance of both understanding and action in defining success. While consultation with medical professionals helped clarify the referent definitions, the intended action was elusive. Is the intended use of the “ambulance” icon to warn passers-by to stay away from a speeding vehicle, or to alert those in need to where an emergency vehicle might be found? As a consequence, is the “ambulance” icon intended to indicate a vehicle, a drop-off space, or a parking spot? Our scorers considered both definition and action when scoring but found that while the two-part answers added content to guide decision, that greater content added complexity to scoring and increased the need to exercise judgment.

On the issue of defining success, we observed that matching answers to referents required expert judgment both about the acceptability of synonyms and in the balance of meaning and action. We hypothesize that the greater clarity of both the definitions and actions of success at the start of a project will not only guide design development more precisely but will reduce the number of difficult judgment decisions during scoring. However well defined the criteria, we believe that some judgments are inevitable but that natural language processing (NLP) software algorithms could be developed that increase the control over decisions, and ultimately the reliability of, scoring qualitative answers. NLP software already exists (LIWC2007 was used for this study) that can be adapted to this task.

SCORING ANSWERS: CONVERTING QUALITATIVE DATA TO QUANTITATIVE CATEGORIES

Subjects in our comprehension studies responded to open-ended questions by writing answers in their own words. Because this original data was qualitative it had to be converted, or ‘scored,’ to determine if each answer was correct. All measurement has an element of subjectivity, scoring qualitative data has more room for subjectivity than most, so control of that subjectivity was critical to reliability. To improve reliability in our studies, each evaluator was given an identical scoring sheet with evaluator instructions, referent definitions, and acceptable synonyms. The directions included:

- all equivalent (synonyms) descriptions of a symbol are correct

- a paraphrase the implies the meaning without naming it is correct E.G. “a series of dots” instead of “conveyor belt” or “man” instead of “patient”
- naming an object not intended to be represented is incorrect e.g. ‘hand’ instead of “glove”
- PLUS indicates a response must include both words of a multi-word referent
- OR: alternate correct answers, either way is correct

Each icon referent listed acceptable synonyms. For example, for the icon referent “medical library” the scoring sheet required:

FA08: Medical Library
medical or health or healthcare or hospital or clinic or doctor’s office or care/care center, etc.
PLUS
library or books or book collection or reading room/area or information place/source, etc.

Note the inclusion of, “etc.” which allowed room for expert judgment. The many icon referents that used two or more words created the possibility of a subject using one word: “library” for the “medical library” icon, while omitting the other word: “medical.” We defined answers that used one word of a two part referent as “partially correct” answers. Evaluators assigned one of four possible scores to each icon answer:

c = correct = matches the scoring sheet
i = incorrect = does not match the scoring sheet, blank, or ‘don’t know’
p = partial = part of the answer matches the scoring sheet
f = fatal = the answer would create critical confusion in the subject with possible harmful results (thinking a bathroom sign is an exit)

As noted above, in our measurement study we found that a small number of scorers was stable. Even so, while most icons received consistent scores from all evaluators, a minority of icons received a variety of different scores. Our analysis of these showed that diverse scores trended in a definite direction, for example: correct + partially correct; or incorrect + partially correct; or incorrect + fatal. By comparing scores of these ‘difficult to score’ icons we were able to define a trend in a specific direction, suggesting whether ‘partially correct’ scores for a particular icon could be more accurately included with the correct, or the incorrect, answers.

On the issue of scoring, we found that while a scoring sheet and definitive instructions can contribute to scoring stability, that a small number of scorers contributed in ways beyond stability. We found that whenever a diversity of scored answers were given for an icon that this was a clear indicator that the icon was problematic. In the future, design objects with diverse scores could be immediately pulled for the visual and textual analysis described below to identify the source of the problem.

ANALYZING ANSWERS: COMPARING QUALITATIVE AND QUANTITATIVE RESULTS

We used a variety of standard statistical instruments to analyze scored answers. We compared and evaluated the arithmetic means of scores for icons from the standard and highly medically literate samples in both countries. Variance and standard deviations were calculated. Significance testing for close mean scores was done. These techniques gave clear results for some icons, such as these for data from Tanzania: “ambulance” 90% correct for standard literacy and 95% correct for high medical literacy; “ultrasound” 15% correct for standard literacy and 70% correct for high medical literacy. We concluded from this that the “ambulance” icon is comprehended regardless of culture or medical literacy, while the “ultrasound” icon works cross culturally but only for those with high medical literacy.

To further our understanding of icon performance, we also conducted a visual analysis of the correct scores. A bar graph was helpful but limited in that the meaning of each bar was abstracted. We designed a visualization of all 54 icons using each icon to represent the mean % correct by the standard (blue), or high medically literate (pink) cohorts (see Figure 4). This visual approach provided simultaneous visual comparison of the scores of all 54 icons of both cohorts in the context of the icon symbols. Analysis of the visualization revealed patterns: most icons using body parts succeeded (CM15, CM17); most icons for medical imaging failed (MA02, MA03, MA07); microscopes are a kind of medical instrument that is widely recognized (CM12, CM13).

But analyzing means, statistically or visually, still left the results unclear for some icons. The icons for “MRI” and “outpatient” both scored poorly in

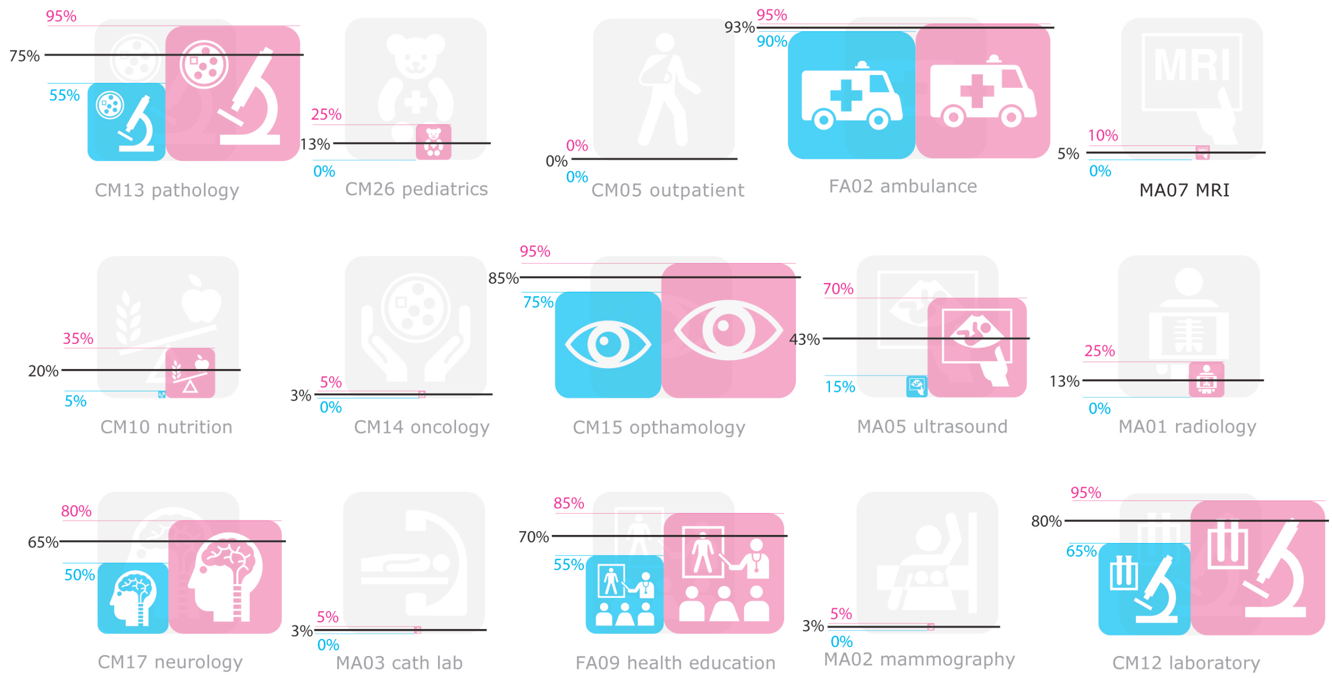


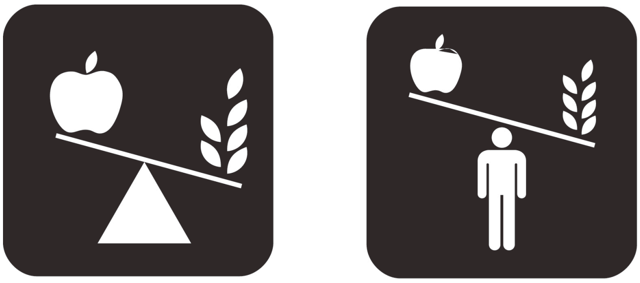
Figure 4.

visual analysis of 54 icon comparison study in rural Tanzania, 15 shown here, gray is the icon representing 100% correct, blue is mean % correct for standard medical literacy, pink is mean % correct for advanced medical literacy

.

Tanzania: "MRI" 0% and 10% correct standard vs. high medical literacy, and "outpatient" 0% and 0% correct standard vs. high medical literacy, but we were unsure why. To help answer this, we analyzed the raw, un-scored written answers to make word counts, find synonyms, even to note positive vs. negative words. For the poorly understood "MRI" icon, text analysis revealed that most answers from the standard medical literacy sample were some variant of "don't know," suggesting complete confusion, while many answers form the medically literate sample referred to a medical research institute, suggesting confusion with a local organization and failure to interpret "MRI" as an acronym for "Magnetic Resonance Imaging."

Figure 5.



alternate icons for nutrition from icon added symbol study

Textual analysis also enlightened our study of the effect of adding symbol details to an icon. There several subjects answered that the nutrition icon with the triangle was making a quantitative statement about the ideal quantity of grain compared to fruits in a diet with answers such as: "apples healthier than wheat," and "wheat plays a bigger part in diet than fruit." The literal scale with the triangle in this case led subjects astray, whereas the metaphorical human figure forming the fulcrum of the scale evoked more correct responses for "nutrition" (see Figure 5).

We also combined textual analysis with visual and statistical methods. From this we observed that when users are confused about meaning they tend to default to a literal description. For the "outpatient" icon many incorrect answers had some variant of "man walking with a broken arm," a literal description of the icon symbols. Setting aside for a moment how such an answer should be scored, it was very useful for designers making design revisions to know what concept was being generated in users' minds. We also combined textual and visual analysis by visually coding text answers with different colors and textures: correct = solid green or green texture; incorrect = pink texture; and fatal = pink solid. Table 1 shows the answers for “medical library” (see Figure 3 right side). It’s easy to see that the main wrong interpretations are “waiting room” and “help/information desk”.

score	Icon 2
c	medical texts available here
i	d k
i	waiting room
p	p - library - be quiet
i	waiting room
i	d k
c	library of medicine
p	p - reading - while waiting
c	reading medical books for giving prescription
c	medical books - read
c	medical books to read
p	p - info avail medical trmt
i	private room for prayer - ignore it
c	medical literature - encour. patients to learn
c	medical books avail. For anyone - ask if I could read
c	research medical conditions - what precautions to look out for
c	medical library - med info
p	library - read book
i	there is a medical counselor in t hospita - ask specialty of educationl
i	information - ask person behind desk for information
c	medical research - medical library
c	information about medicine here
c	medical reference - book
c	research facilities - study
c	reading area - book to read
c	reading book about diseases
c	library for medical books -learn more about my health condition
c	medical books
c	reading medical books
p	p - read your reports -results get analyzed
i	a university related to medicine - if interested in studying
c	medical library
p	p - learn knowledge - study harder
c	medical library - read-learn about different medical terms
c	find information on health
c	medical archives - look for my records
c	medical books - research
c	medical learning library - find medical information
c	health/medical library- for information
c	medical books/literature

Table 1.

visualization of text answers to “medical library” icon

On the issue of analysis, we observed that statistical data combined with visual and textual analysis can bring to light important findings. Visual analysis alone revealed things that were not apparent in the statistical data: that microscopes were well understood for example. Textual analysis, particularly of failed icons, revealed important insights to guide design revision. Overall, one approach alone was much less revealing than all three together. We found this particularly true of designers, more schooled in visualization than statistics. We hypothesize that designers leading evaluative research will continue to integrate statistical, visual, and textual methods and that over time specific combined analytical methods will be validated, and perhaps integrated into software.

OTHER ISSUES

INTEGRATION OF MIXED METHODS

As you can see from above, our evaluative studies followed an integrated mixed methods research paradigm. We collected qualitative data, converted it to quantitative data, then used both forms of data together to direct our conclusions.

ACTING ON RESULTS

Research that claims to demonstrate something needs a threshold of proof. Typically a “p” value of <.05 is considered significant, with some domains requiring the more stringent p <.01. Design is a domain that seldom involves life and death issues. What is a proper standard of proof for design, who and how should determine threshold of proof for design, and how to establish this are questions to which we have no answer.

WHEN TO APPLY EVALUATIVE METHODS

It might have been reasonably assumed by readers that evaluative design research methods occur after the design has been developed, but this is far from the truth as the studies cited here show. Our evaluative methods were deployed as part of the design process. Even when evaluative research has been done by the author to measure the effectiveness of a final product being used in the field, as in the previously mentioned HIV/AIDS curriculum, the evaluative results will be used to guide the design of revised HIV/AIDS curriculum, and the principles identified in the evaluative studied can be used for any similar design project.

CONCLUSION

In our evaluative studies we encountered many issues. We had to identify what was meaningful (validity) and how to measure it. Then we had to establish a threshold of measurement that would qualify as success. We also had to confront methods for scoring including who referees specific results and the reliability of our evaluators. We had to develop rigorous analytical methods suitable for designers to use.

On the issue of defining success, we observed that clear definitions of success are needed but difficult to obtain. Other fields (such as biomedical informatics, (Friedman & Wyatt, 2006)) may deploy separate measurement studies to ascertain the reliability of measurement criteria before starting an evaluative study. Designers may adapt this method, but we did not use it here. We recognize that designers often redefine success as the design process moves iteratively through various proposed solutions and problem frames. Reframing the problem can change the criteria for success in meaningful ways, making measurement a moving target. However, we hypothesize that a focus on measurement criteria as they change may help the design process as it evolves by clarifying the evolution of definitions of success as they occur. This could identify a trajectory of the design process, leading perhaps into unexplored territory. We believe automation using natural language processing (NLP) software algorithms could facilitate flexibility and control over measurement definitions in evaluative stiudies, but have yet to prove this.

On the issue of scoring, we found that a small number of scorers contributed stability and helped identify when an icon was problematic through the diversity of scorer responses.

On the issue of analysis, we observed that detailed knowledge about a design's failures provided insight and creative stimulus for further design development: failures were the most stimulating (most successful?!?) study objects! We successfully integrated visual and textual analysis to reveal things that were not apparent in the statistical data, again providing not only insight but also food for further creative development. We hypothesize that designers leading evaluative research will continue to integrate statistical, visual, and textual methods and that over time specific combined analytical methods will be validated, and perhaps integrated into software.

Some may argue that everything is subjective making measurement and precision an illusive chasing after the wind. The response to this hypothetical objection is that it is an objective-based subjective opinion founded on worldview more than fact. Issues of performance and accuracy matter both to the function and to the societal value of design.

We believe we have pushed the envelope on each of several evaluative issues, but are far from establishing standards. We are certain however that evaluative design research even in its infancy is valuable both as a follow-up measure of success and, perhaps more importantly, as a part of the design process leading up to a more effective design.

REFERENCES

Akins, R. i. B., Tolson, H., & Cole, B. R. (2005). Stability of response characteristics of a Delphi Panel: application of bootstrap data expansion. *BMC Medical Research Methodology, 5*(37).

Bird, C. (2011). Do You Know What This Symbol Means? Retrieved April, 29, 2011, 2011, from http://autos.yahoo.com/articles/autos_content_landing_pages/1498/do-you-know-what-this-symbol-means

Card, S. K., Mackinlay, J. D., & Schneiderman, B. (2003). *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA: Morgan Kaufman.

Erlhoff, M., & Marshall, T. E. (2008). *Design Dictionary: Perspectives on Design Terminology*. Basel Boston Berlin: Birkhäuser.

Friedman, C. P., & Wyatt, J. C. (2006). *Evaluation Methods in Biomedical Informatics*. New York, NY: Springer.

Hunter, J. D. (2010). *To Change the World: The Irony, Tragedy, and POssibility of Christianity in the Late Modren World*. Oxford: Oxford University Press.

Graphical symbols - Test methods Part 1: Methods for testing comprehensibility, 9186-1 C.F.R. (2007).

Laurel, B.-E. (2003). *Design Research Methods and Perspectives* (Vol. Cambridge, MA): MIT Press.

ACKNOWLEDGEMENT

Special thanks to graduate design students in course 23DSGN752 – Research Methods 2, for their participation in the measurement study and administration of the USA portion of the medical literacy comprehension study: Gabe Botkins, Wanqi Chai, Suyuan Fan, Thomas Gilmore, Doug Hovekamp, Do Young Lee, Sichun Li, Marcey Mankosa, Feifei Pang, Patricia Poarch, Tsung-Yu Shih, Jianzhou Song, Rachel Stuckey, and Yingxue Zhao.