Bibliographic Displays in Web Catalogs: Does Conformity to Design Guidelines Correlate with User Performance?

Joan M. Cherry, Paul Muter, and Steve J. Szigeti

The present study investigated whether there is a correlation between user performance and compliance with screen-design guidelines found in the literature. Rather than test individual guidelines and their interactions, the authors took a more holistic approach and tested a compilation of guidelines. Nine bibliographic display formats were scored using a checklist of eighty-six guidelines. Twenty-seven participants completed ninety search tasks using the displays in a simulated Web environment. None of the correlations indicated that user performance was statistically significantly faster with greater conformity to guidelines. In some cases, user performance was actually significantly slower with greater conformity to guidelines. In a supplementary study, a different set of forty-three guidelines and the user performance data from the main study were used. Again, none of the correlations indicated that user performance was statistically significantly faster with greater conformity to guidelines.

Attempts to establish generalizations are ubiquitous in science and in many areas of human endeavor. It is well known that this enterprise can be extremely problematic in both applied and pure science.[1] In the area of human-computer interaction, establishing and evaluating generalizations in the form of interface-design guidelines are pervasive and difficult challenges, particularly because of the intractably large number of potential interactions among guidelines. Using bibliographic display formats from Web catalogs, the present study utilizes global evaluation by correlating user performance in a search task with conformity to a compilation of eighty-six guidelines (divided into four subsets).

The literature offers many design guidelines for the user interface, some of which cover all aspects of the user interface, some of which focus on one aspect of the user interface—e.g., screen design. Tullis, in chapters in two editions of the *Handbook of Human-Computer Interaction,* reviews the work in this area.[2] The earlier chapter provides a table describing the screen-design guidelines available at that time. He includes, for example, Galitz, whom he notes have several hundred guidelines addressing general screen design, and Smith and Mosier, whom he notes have about three hundred guidelines addressing the display of data.[3]

Earlier guidelines tended to be generic. More recently, guidelines have been developed for specific applications—e.g., Web sites for airline travel agencies, multimedia applications, e-commerce, children, bibliographic displays, and public-information kiosks.[4]

Although some of the guidelines in the literature are based on empirical evidence, many are based on expert opinion and have not been tested. Some of the research-based guidelines have been tested in isolation or in combination with only a few other guidelines. The National Cancer Institute (NCI) Web site, *Research-based Web Design and Usability Guidelines,* rates sixty guidelines on a scale of 0 to 5 based on the strength of the evidence.[5] The more valid the studies that directly support the guideline, the higher the rating. In interpreting the scores, the site advises that scores of 1, 2, or 3 suggest that "more evidence is needed to strengthen the designer's overall confidence in the validity of a guideline." Of the sixty guidelines on the site, forty-six (76.7 percent) fall into this group. In 2003, the United States Department of Health and Human Services Web site, *Research-based Web Design and Usability Guidelines,* rated 187 guidelines on a different five-point scale.[6] Eighty-two guidelines (43.9 percent) meet the criteria of having strong or medium research support. Another forty-eight guidelines (25.7 percent) are rated as having weak research support. Thus, there is some research support for 69.6 percent of the guidelines.

In addition to the issue of the validity of individual guidelines, there may be interactions among guidelines. An interaction occurs if the effect of a variable depends on the level of another variable—e.g., an interaction occurs if the usefulness of a guideline depends on whether some other guideline is being followed. A more severe problem is the potential for high-order interactions: The nature of a two-way interaction may depend on the level of a third variable, the nature of a three-way interaction may depend on the level of a fourth variable, and so on. Because of the combinatorial explosion, if there are more than a few variables the number of possible interactions becomes huge. As Cronbach stated: "Once we attend to interactions, we enter a hall of mirrors that extends to infinity."[7]

With a large set of guidelines, it is impractical to test all of the guidelines and all of the interactions, including high-order interactions. Muter suggested several approaches for handling the problem of intractable high-order interactions, including adapting optimizing algorithms such as Simplex, seeking "robustness in variation," re-construing the problem, and pruning the alternative space.[8] The present study utilizes another approach: global evaluation by

Joan M. Cherry (joan.cherry@utoronto.ca) is a Professor in the Faculty of Information Studies; **Paul Muter** (muter@psych.utoronto.ca) is an Assistant Professor in the Department of Psychology; and **Steve J. Szigeti** (szigeti@fis.utoronto.ca) is a doctoral student in the Faculty of Information Studies and the Knowledge Media Design Institute, all at the University of Toronto, Canada.

correlating user performance with conformity to a set of guidelines. Using this method, particular guidelines and interactions are not tested, but the set and subsets are tested globally, and some of the interactions, including high-order interactions, are captured. Bibliographic displays were scored using a compilation of guidelines, divided into four subsets, and the performance of users doing a set of search tasks using the displays was measured. An attempt was made to determine whether users find information more quickly on displays that receive high scores on checklists of screen-design guidelines.

The authors are aware of only two studies that have investigated conformity with a set of guidelines and user performance, and they both included only ten guidelines. D'Angelo and Twining measured the correlation between compliance with a set of ten standards (D'Angelo Standards) and user comprehension.[9] The D'Angelo Standards are in the form of principles for Web-page design, based on a review of the literature.[10] D'Angelo and Twining found a small correlation (.266) between number of standards met and user comprehension.[11] They do not report on statistical significance, but from the data provided in the paper it appears that the correlation is not significant. Gerhardt-Powals compared an interface designed according to ten cognitive engineering principles to two control interfaces and found that the cognitively engineered interface resulted in statistically significantly superior user performance.[12]

The guidelines used in the present study were based on a list compiled by Chan to evaluate displays of bibliographic records in online library catalogs.[13] The set of guidelines was broken down into four subsets. Participants in this study were given search tasks and clicked on the requested item on a bibliographic display. The main dependent variable of interest was response time.

Method

Participants

Twenty-seven participants were recruited through the University of Toronto Psychology 100 Subject Pool. Seventeen were female; ten were male. Most (twenty) were in the age group 17 to 24; three were in the age group 25 to 34 years, and four were in the age group 35 to 44. One had never used the Web; all others reported using the Web one or more hours per week. Participants received course credit.

Design

To control for the effects of fatigue, practice runs, and the like, the order of trials was determined by two orthogonal 9 x 9 Latin squares—one to select a display and one to select a book record. Each participant completed five consecutive search tasks—Author, Title, Call Number, Publisher, and Date—in a random order, with each display-book combination. (The order of the five search tasks was randomized each time.) This procedure was repeated, so that in total each participant did ninety tasks (9 displays x 5 tasks x 2 repetitions).

Materials and apparatus

The study used nine displays from library catalogs available on the Web. They were selected to represent a variety of systems and to illustrate the remarkable diversity in bibliographic displays in Web catalogs. The displays differed in the amount of information included, the structure of the display, employment of highlighting techniques, and use of graphical elements. Four examples of the nine displays are presented in figures 1a, 1b, 1c, and 1d. The displays were captured and presented in an interactive environment using Active Server Page (ASP) software. The look of the displays was retained, but hypertext links were deactivated.

Nine different book records were used to provide the content for the displays. Items selected were those that would be readily understood by most users—e.g., books by Saul Bellow, Norman Mailer, and John Updike.

The guidelines were based on a list compiled by Chan from a review of the literature in human-computer interaction and library science.[14] The list does not include guidelines about the process of design. Chan formatted the guidelines as a checklist for bibliographic displays in online catalogs. In work reported in 1996, Cherry and Cox modified the checklist for use with bibliographic displays in Web catalogs.[15] In a 1998 paper, Cherry reported on evaluations of bibliographic displays in catalogs of academic libraries, based on Chan's data for twelve OPACs and data for ten Web catalogs evaluated by Cherry and Cox using a modification of the 1996 checklist for Web catalogs.[16] The findings showed that, on average, displays in OPACs scored 58 percent and displays in Web catalogs scored 60 percent. The 1996 checklist of guidelines was modified by Herrero-Solana and De Moya-Anegón, who used it to explore the use of multivariate analysis in evaluating twenty-five Latin American catalogs.[17] For the present study four questions were removed that were considered less useful from the checklist used in Cherry's 1998 analysis.

The checklist consisted of four sections or subsets: Labels (these identify parts of the bibliographic description); Text (the display of the bibliographic, holdings/location, and circulation status information); Instructions (includes instructions to users, informational messages, and options available); and Layout (includes identification of the screen, the organization for the bibliographic

information, spacing, and consistency of information presentation). Items on the checklist were phrased as questions requiring Yes/No responses. Examples of the items are: Labels: "Are all fields/variables labeled?" Text: "Is the text in mixed case (upper and lowercase)?" Instructions: "Are instructional sentences or phrases simple, concise, clear, and free of typographical errors?" and Layout: "Is the width of the display no more than forty to sixty characters?"

The set used in the present study contained eighty-six guidelines in total, of which forty-eight were generic and could be applied to any application. Thirty-eight are specific and apply to bibliographic displays in Web catalogs.

The experiment was run on a Pentium computer with a seventeen-inch Sony color monitor with a standard keyboard and mouse.

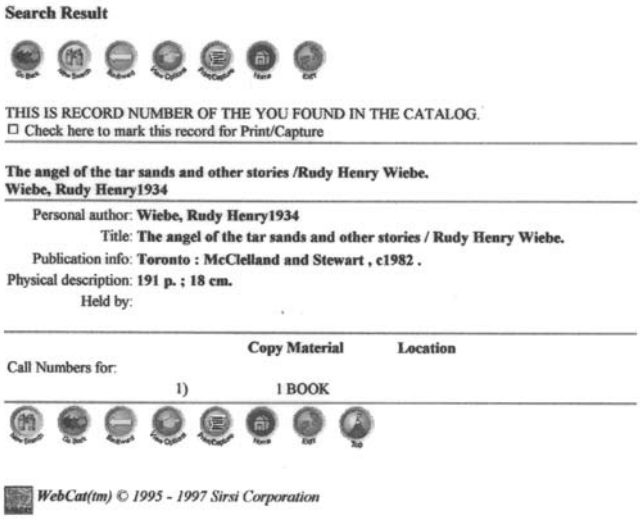

Figure 1a. Example of display

Figure 1b. Example of display

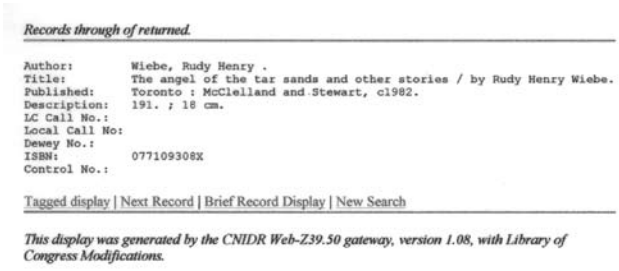

Figure 1c. Example of display

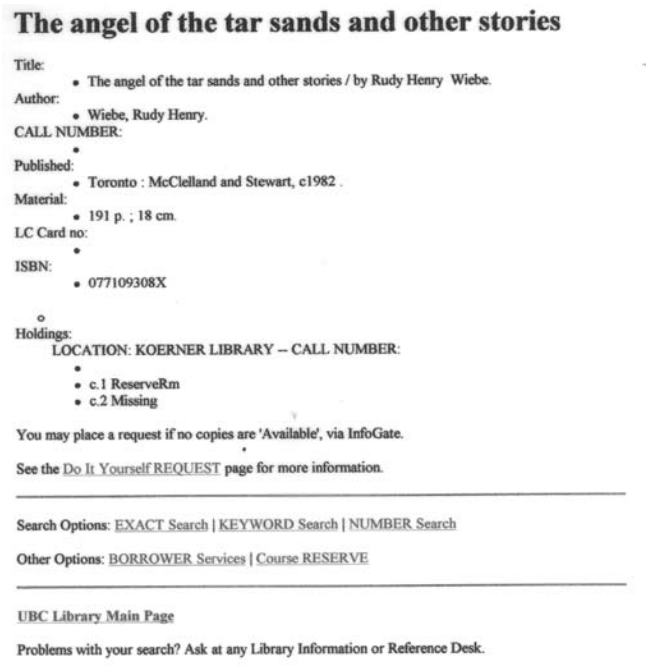

Figure 1d. Example of display

Procedure

Participants were tested individually. Five practice trials with a display and book record not used in the experiment familiarized the participant with the tasks and software.

At the beginning of a trial, the message "When ready, click" appeared on the screen. When the participant clicked on the mouse, a bibliographic display appeared along with a message at the top of the screen indicating whether the participant should click on the author, title, call number, publisher, or date of publication—e.g., "Current task: Author." Participants clicked on what they thought was the correct answer. If they clicked on any other area, the display was shown again. An incorrect click was not defined as an error—in effect, percent correct was always 100—but an incorrect click would of course add to the response time. The software recorded the time to successfully complete each search, the identification for the display and the book record, and the search-task type. When a participant completed the five search tasks for a display, a message was shown indicating the average response time on that set of tasks.

When participants completed the ninety search tasks, they were asked to rank the nine displays according to their preference. For this task, a set of laminated color printouts of the displays was provided. Participants ranked the displays, assigning a rank of 1 to the display that they preferred most, and 9 to the one they preferred least. They were also asked to complete a short background questionnaire. The entire session took less than forty-five minutes.

Scoring the displays on screen design guidelines

The authors' experience has indicated that judging whether a guideline is met can be problematic: evaluators sometimes differ in their judgments. In this study, three evaluators assessed each of the nine displays independently. If there was any disagreement amongst the evaluators' responses for a given question for a given display, that question was not used in the computation of the percentage score for that display. (A guideline regarding screen density was evaluated by only one evaluator because it was very time-consuming.) The total number of questions used to assess each display was eighty-six. The number of questions on which the evaluators disagreed ranged from twelve to thirty across the nine displays. All questions on which the three evaluators agreed for a given display were used in the calculation of the percentage score for that display. Hence the percentage scores for the displays are based on a variable set and number of questions—from fifty-six to seventy-four. The subset of questions on which the three evaluators agreed for all nine displays was small—twenty-two questions.

Results

With regard to conformity to the guidelines, in addition to the overall scores for each display, which ranged from 42 percent to 65 percent, the percentage score was calculated for each subset of the checklist (Labels, Text, Instructions, and Layout).

The time to successfully complete each search task was recorded to the nearest millisecond. (For some unknown reason, six of the 2,430 response times recorded [27 x 90] were 0 milliseconds. The program was written in such a way that the response-time buffer was cleared at the time of stimulus presentation, in case the participant clicked just before this time. These trials were treated as missing values in the calculation of the means.) Six mean response times were calculated: Author, Title, Call Number, Publisher, Date, and the sum of the five response times, called All Tasks. The mean of All Tasks response times ranged from 13,671 milliseconds to 21,599 milliseconds for the nine formats. The nine display formats differed significantly on this variable according to an analysis of variance, $F(8, 477) = 17.1$, $p < .001$.

The correlations between response times and guidelines-conformance scores are presented in table 1. It is important to note that a high correlation between response time and conformity to guidelines indicates a low correlation between user performance (speed) and conformity to guidelines. Row 1 of table 1 contains correlations between the total guidelines score and response times; Column 1 contains correlations between All Tasks (the sum of the five response times) and guidelines scores. Of course, the correlations in table 1 are not all independent of each other.

Only five of the thirty correlations in table 1 are significant at the .05 level, and they all indicate slower response times with higher conformity to guidelines. Of the six correlations in table 1 indicating faster response times with higher conformity to guidelines, none approaches statistical significance. The upper left-hand cell of table 1 indicates that the overall correlation between total scores on the guidelines and the mean response time across all search tasks (All Tasks) was 0.469 ($df = 7$, $p = 0.203$)—i.e., conformity to the overall checklist was correlated with slower overall response times, though this correlation did not approach statistical significance.

Figure 2 shows a scatter plot of the main independent variable, overall score on the checklist of guidelines, and the main dependent variable, the sum of the response times for the five tasks (All Tasks). Figure 3 shows a scatter plot for the highest obtained correlation: between score on the overall checklist of guidelines and the time to complete the Title search task. Visual inspection suggests patterns consistent with table 1: no correlation in figure 2, and slower search times with higher guidelines scores in figure 3.

Finally, correlations were computed between preference and response times (All Tasks response times and five

specific-task response times) and between preference and conformity to guidelines (overall guidelines four subsets of guidelines). None of the eleven correlations approached statistical significance.

Table 1. Correlations between scores on the checklist of screen design guidelines and time to complete search tasks: Pearson Correlation (Sig. - 2-tailed); N=9 all cells

	All tasks	Author	Title	Call #	Publisher	Year
Total score:	.469	.401	.870	.547	.035	.247
	(.203)	(.285)	(.002)	(.127)	(.930)	(.522)
Labels:	.722	.757	.312	.601	.400	.669
	(.028)	(.018)	(.413)	(.087)	(.286)	(.049)
Text:	-.260	-.002	.595	-.191	-.412	-.288
	(.500)	(.997)	(.091)	(.623)	(.271)	(.452)
Instructions:	.422	.442	.712	.566	.026	.126
	(.258)	(.234)	(.032)	(.112)	(.947)	(.748)
Layout:	.602	-.102	.383	.624	.492	.367
	(.086	(.794)	(.308)	(.073)	(.179)	(.332)

Supplementary Study

To further validate the results of the main study, it was decided to score the interfaces against a different set of guidelines based on the 2003 U.S. Department of Health and Human Services *Research-based Web Design and Usability Guidelines*. This set consists of 187 guidelines and includes a rating for each guideline based on strength of research evidence for that guideline. The present study started with eighty-two guidelines that were rated as having either moderate or strong research support, as the definitions of both of these include "cumulative research-based evidence."[18] Compliance with guidelines that address the *process* of design can only be judged during the design process, or via access to the interface designers. Since this review process did not allow for that, a total of nine process-focused guidelines were discarded. This set of seventy-three guidelines was then compared with the sixty-guideline 2001 NCI set, *Research-based Web Design and Usability Guidelines*, intending to add any outstanding NCI guidelines supported by strong research evidence to the existing list of seventy-three. However, all of the strongly supported NCI guidelines were already represented in the original seventy-three. Finally, the guidelines in the ISO 9241, *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)*, part 11 (*Guidance on Usability*), part 12 (*Presentation of Information*), and part 14 (*Menu Dialogues*) were compared to the existing set of seventy-three, with the intention that any prescriptive guideline in the ISO set that was not already included in the original seventy-three would be added.[19] Again, there were none. The seventy-three guidelines were organized into three thematic groups: (1) layout (the organization of textual and graphic material on the screen), (2) interaction (which included navigation or any element with which the user would interact), and (3) text and readability.

All of the guidelines used were written in a manner allowing readers room for interpretation. The authors explicitly stated that they were not writing rules, but rather, guidelines, and recognized that their application must allow for a level of flexibility.[20] This ambiguity creates problems in terms of assessing displays. In this study, two evaluators independently assessed the nine displays.

The first evaluator applied all seventy-three guidelines and found thirty to be nonapplicable to the specific types of interfaces considered. The second evaluator applied the shortened list of forty-three guidelines. Following the independent evaluations, the two evaluators compared assessments. The initial rate of agreement between the two assessments ranged from 49 percent to 70 percent across the nine displays. In cases where there was disagreement, the evaluators discussed their rationale for the assessment in order to achieve consensus.

Results of supplementary study

As with the initial study, in addition to the overall scores for each display, the percentage score was calculated for each subset of the checklist (Labels, Interaction, and Text and Readability). It is worth noting that the overall scores witnessed higher compliance to this second set of guidelines, ranging from 68 percent to 89 percent. The correlations between response times and guidelines-conformance scores are presented in table 2. Again, it is important to note that a high correlation between response time and conformity to guidelines indicates a low correlation between user performance (speed) and conformity to guidelines. Row 1 of table 2 contains correlations between the total guidelines score and response times; column 1 contains correlations between All Tasks (the sum of the five response times) and guidelines scores. Of course, the correlations in table 2 are not all independent of each other.

Only one of the twenty-four correlations in table 2

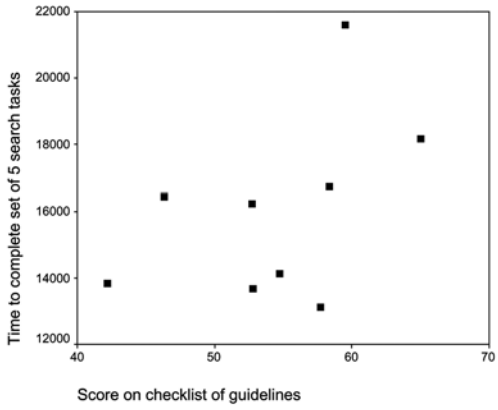

Figure 2. Scatter plot for overall score on checklist of screen design guidelines and time to complete set of five search tasks

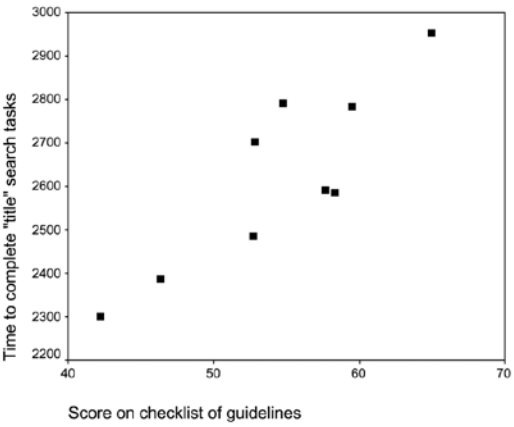

Figure 3. Scatter plot for overall score on checklist of screen design guidelines and time to complete "Title" search tasks

is significant at the .05 level, and it indicates a slower response time with higher conformity to guidelines. Of the ten correlations in table 2 indicating faster response times with higher conformity to guidelines, none approaches statistical significance. The upper left-hand cell of table 2 indicates that the overall correlation between total scores on the guidelines and the mean response time across all search tasks (All Tasks) was 0.292 ($p = 0.445$)—i.e., conformity to the overall checklist was correlated with slower overall response times, though this correlation did not approach statistical significance.

Figure 4 shows a scatter plot of the main independent variable, overall score on the checklist of guidelines, and the main dependent variable, the sum of the response times for the five tasks (All Tasks). Figure 5 shows a scatter plot for the highest-obtained correlation: between score on the Text and Readability category of guidelines and the time to complete the Title search task. Visual inspection suggests patterns consistent with table 2: no correlation in figure 4, and slower search times with higher guidelines scores in figure 5.

Discussion

In the present experiment and the supplementary study, none of the correlations indicating faster user performance with greater conformity to guidelines approached statistical significance. In some cases, user performance was actually significantly slower with greater conformity to guidelines—i.e., in some cases, there was a negative correlation between user performance and conformity to guidelines.

The authors are aware of no other study indicating a negative correlation between user performance and conformity to interface design guidelines. Some researchers would not be surprised at a finding of zero correlation between user performance and conformity to guidelines, but a negative correlation is somewhat puzzling. A negative correlation implies that there is something wrong somewhere—perhaps incorrect underlying theories or an incorrect body of assumptions. Such a negative correlation is not without precedent in applied science. In the field of medicine, before the turn of the twentieth century, seeing a doctor actually decreased the chances of improving health.[21] Presumably, medical guidelines of the time were negatively correlated with successful practice, and the negative correlation implies not just worthlessness, but medical theories or beliefs that were actually incorrect and harmful.

The boundary conditions of the present findings are unknown. The present findings may be specific to the tasks employed—fairly simple search tasks. The findings may apply only to situations in which the user is switching formats frequently, as opposed to situations in which each user is using only one format. (A between-subjects design would test this possibility.) The findings may be specific to the two sets of guidelines used. With sets of ten guidelines, D'Angelo and Twining and Gerhardt-Powals found positive correlations between user performance and conformity to guidelines (though apparently not statistically significantly in the former study).[22] The guidelines used in the authors' main study and supplementary study tended to be more detailed than in the other two studies. Detailed guidelines are sometimes seen as advantageous, since developers who use guidelines need to be able to interpret the guidelines in order to implement them. However, perhaps following a large number of detailed

guidelines reduces the amount of personal judgment used and results in less effective designs. (Designers of the nine displays used in the present study would not have been using either of the sets of guidelines used in our studies but may have been using some of the sources from which our guidelines were extracted.) As noted by Cheepen in discussing guidelines for voice dialogues, sometimes a designer's experience may be more valuable than a particular guideline.[23]

The lack of agreement in interpreting the guidelines was an unexpected but interesting factor revealed during the collection of data in both the main study and the supplementary study. While a higher rate of agreement had been expected, the differences raised an important point in the use of guidelines. If guidelines intentionally leave room for interpretation, what factor does expert opinion and experience play in design? In the main study, the number of guidelines on which the evaluators disagreed ranged from 14 percent to 35 percent across the nine displays. In the supplementary study, both evaluators had experience in interface design through a number of different roles in the design process (both academic and professional). This meant the evaluators' interpretations of the guidelines were informed by previous experience. The initial level of disagreement ranged from 30 percent to 51 percent across the nine displays. While it was possible to quickly reach consensus

Table 2. Correlations between scores on subset of the U.S. Dept. of Health and Human Services (2003) *Research–based Web Design and Usability Guidelines* and time to complete search tasks: Pearson Correlation (Sig. - 2-tailed); N=9 all cells

	All tasks	Author	Title	Call #	Publisher	Year
Total score:	.292	.201	.080	-.004	.345	.499
	(.445)	(.604)	(.839)	(.992)	(.363)	(.172)
Layout:	-.308	-.264	-.512	-.332	.046	-.294
	(.420)	(.492)	(.159)	(.383)	(.906)	(.442)
Text:	.087	-.051	.712	-.059	-.095	-.259
	(.824)	(.895)	(.032)	(.879)	(.808)	(.500)
Interaction:	.638	.603	.055	.439	.547	.625
	(.065)	(.085)	(.887)	(.238)	(.128)	(.072)

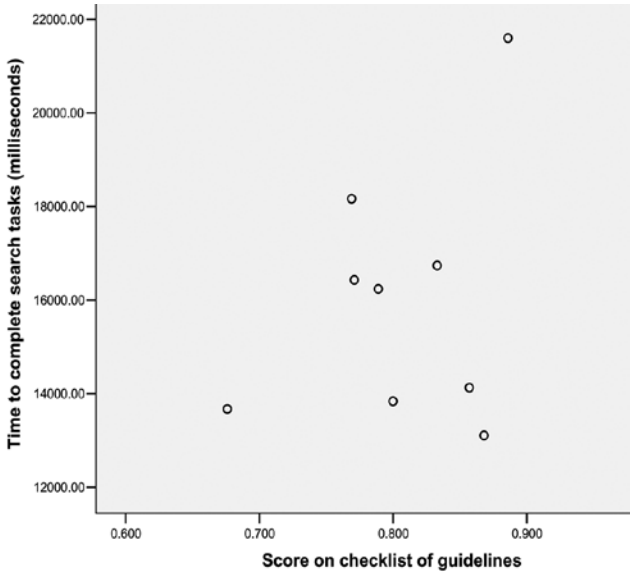

Figure 4. Scatter plot for subset of U.S. Department of Health and Human Services (2003) *Research–based Web Design and Usability Guidelines* conformance score and total time to complete five search tasks

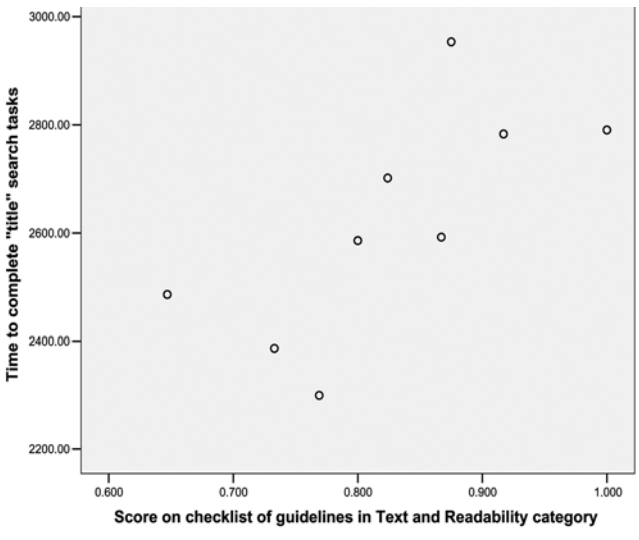

Figure 5. Scatter plot for Text and Readability category of U.S. Department of Health and Human Services (2003) *Research–based Web Design and Usability Guidelines* and time to complete "Title" search tasks

on a number of assessments (because both evaluators recognized the high degree of subjectivity that is involved in design), it also led to longer discussions regarding the intentions of the guideline authors. A majority of the differences involved lack of guideline clarity (where one evaluator had indicated a meet-or-fail score, while another felt the guideline was either unclear or not applicable). Does this imply that guidelines can best be applied by committees or groups of designers? The dynamic of such groups would add another complex variable to understanding the relationship between guideline conformity and user performance.

Future research should test other tasks and other sets of guidelines to confirm or refute the findings of the present study. There should also be investigation of other potential predictors of display effectiveness. For example, would the ratings of usability experts or graphic designers for a set of bibliographic displays be positively correlated with user performance? Crawford, in response to a paper presenting findings from an evaluation of bibliographic displays using a previous version of the checklist of guidelines used in the main study, commented that the design of bibliographic displays still reflects art, not science.[24] Several researchers have discussed aesthetics and user interface design. Reed et al. noted the need to extend our understanding of the role of aesthetic elements in the context of user-interface guidelines and standards.[25] Ngo, Teo, and Byrne discussed fourteen aesthetic measures for graphic displays.[26] Norman discussed these ideas in "Emotions and Design: Attractive Things Work Better."[27] Tractinsky, Katz, and Ikar found strong correlations between perceived aesthetic appeal and perceived usability.[28]

Most empirical studies of guidelines have looked at one variable only or, at the most, a small number of variables. The opposite extreme would be to do a study that examines a large number of variables factorially. For example, assuming eighty-six yes/no guidelines for bibliographic displays, it would be theoretically possible to do a factorial experiment testing all possible combinations of yes/no—2 to the 86th power. In such an experiment, all two-way interactions and higher interactions could be assessed, but such an experiment is not feasible. What the authors have done is somewhere between these two extremes. This study has the disadvantage that we cannot say anything about any individual guideline, but it has the advantage that it captures some of the interactions, including high-order interactions.

Despite the present results, the authors are not recommending abandoning the search for guidelines in interface design. At a minimum, the use of guidelines may increase consistency across interfaces, which may be helpful. However, in some research domains, particularly when huge numbers of potential interactions result in extreme complexity, it may be advisable to allocate resources to means other than attempting to establish guidelines, such as expert review, relying on tradition, letting natural selection take its course, utilizing the intuitions of designers, and observing user-interaction. Indeed, in pure and applied research in general, perhaps more resources should be allocated to means other than searching for explicit generalizations. Future research may better indicate when to attempt to establish generalizations and when to use other methods.

Acknowledgements

This work was supported by a Social Sciences and Humanities Research Council General Research Grant awarded by the Faculty of Information Studies, University of Toronto, and by the Natural Sciences and Engineering Research Council of Canada. The authors wish to thank Mark Dykeman and Gerry Oxford who developed the software for the experiment; Donna Chan, Joan Bartlett, and Margaret English, who scored the displays with the first set of guidelines; Everton Lewis, who conducted the experimental sessions; M. Max Evans, who helped score the displays with the supplementary set of guidelines; and Robert L. Duchnicky, Jonathan L. Freedman, Bruce Oddson, Tarjin Rahman, and Paul W. Smith for helpful comments.

References and notes

1. See, for example, A. Chapanis, "Some Generalizations About Generalization," *Human Factors* 30, no. 3 (1988): 253–67.

2. T. S. Tullis, "Screen Design," in *Handbook of Human-Computer Interaction,* ed. M. Helander (Amsterdam: Elsevier, 1988), 377–411; T. S. Tullis, "Screen Design," in *Handbook of Human-Computer Interaction,* 2d ed., eds. M. Helander, T. K. Landauer, and P. Prabhu (Amsterdam: Elsevier, 1997), 503–31.

3. W. O. Galitz, *Handbook of Screen Format Design,* 2d ed. (Wellesley Hills, Mass.: QED Information Sciences, 1985); S. L. Smith and J. N. Mosier, *Guidelines for Designing User Interface Software*, Technical Report ESD-TR-86-278 (Hanscom Air Force Base, Mass.: USAF Electronic Systems Division, 1986).

4. C. Chariton and M. Choi, "User Interface Guidelines for Enhancing the Usability of Airline Travel Agency e-Commerce Web Sites," *CHI '02 Extended Abstracts on Human Factors in Computing Systems,* Apr. 20–25, 2002 (Minneapolis, Minn.: ACM Press), 676–77, http://portal.acm.org/citation.cfm?doid=506443.506541 (accessed Dec. 28, 2005); M. G. Wadlow, "The Andrew System; The Role of Human Interface Guidelines in the Design of Multimedia Applications," *Current Psychology: Research and Reviews* 9 (Summer 1990): 181–91; J. Kim and J. Lee, "Critical Design Factors for Successful e-Commerce Systems," *Behaviour and Information Technology* 21, no. 3 (2002): 185–99; S. Giltuz and J. Nielsen, *Usability of Web Sites for Children:*

70 Design Guidelines (Fremont, Calif.: Nielsen Norman Group, 2002); Juliana Chan, "Evaluation of Formats Used to Display Bibliographic Records in OPACs in Canadian Academic and Public Libraries," Master of Information Science Research Project Report (University of Toronto: Faculty of Information Studies, 1995); M. C. Maquire, "A Review of User-Interface Design Guidelines for Public Information Kiosk Systems," *International Journal of Human-Computer Studies* 50, no. 3 (1999): 263–86.

5. National Cancer Institute, *Research-based Web Design and Usability Guidelines* (2001), www.usability.gov/guidelines/index.html (accessed Dec. 28, 2005).

6. U.S. Department of Health and Human Services, *Research-based Web Design and Usability Guidelines* (2003), http://usability.gov/pdfs/guidelines.html (accessed Dec. 28, 2005).

7. L. J. Cronbach, "Beyond the Two Disciplines of Scientific Psychology," *American Psychologist* 30, no. 2 (1975): 116–27.

8. P. Muter, "Interface Design and Optimization of Reading of Continuous Text," in *Cognitive Aspects of Electronic Text Processing,* eds. H. van Oostendorp and S. de Mul (Norwood, N.J.: Ablex, 1996), 161–80; J. A. Nelder and R. Mead, "A SIMPLEX Method for Function Minimization," *Computer Journal* 7, no. 4 (1965): 308–13; T. K. Landauer, "Research Methods in Human-Computer Interaction," in *Handbook of Human-Computer Interaction*, ed. M. Helander (Amsterdam: Elsevier, 1988), 905–28; R. N. Shepard, "Toward a Universal Law of Generalization for Psychological Science," *Science* 237 (Sept. 11, 1987): 1317–323.

9. J. D. D'Angelo and J. Twining, "Comprehension by Clicks: D'Angelo Standards for Web Page Design, and Time, Comprehension, and Preference," *Information Technology and Libraries* 19, no. 3 (2000): 125–35.

10. J. D. D'Angelo and S. K. Little, "Successful Web Pages: What are They and Do They Exist?" *Information Technology and Libraries* 17, no. 2 (1998): 71–81.

11. D'Angelo and Twining, "Comprehension by Clicks."

12. J. Gerhardt-Powals, "Cognitive Engineering Principles for Enhancing Human-Computer Performance," *International Journal of Human-Computer Interaction* 8, no. 2 (1996): 189–211.

13. Chan, "Evaluation of Formats."

14. Ibid.

15. Joan M. Cherry and Joseph P. Cox, "World Wide Web Displays of Bibliographic Records: An Evaluation," *Proceedings of the 24th Annual Conference of the Canadian Association for Information Science* (Toronto, Ontario: Canadian Association for Information Science, 1996), 101–14.

16. Joan M. Cherry, "Bibliographic Displays in OPACs and Web Catalogs: How Well Do They Comply with Display Guidelines?" *Information Technology and Libraries* 17, no. 3 (1998): 124–37; Cherry and Cox, "World Wide Web Displays of Bibliographic Records."

17. V. Herrero-Solana and F. De Moya-Anegón, "Bibliographic Displays of Web-Based OPACs: Multivariate Analysis Applied to Latin-American Catalogs," *Libri* 51 (June 2001): 75–85.

18. U.S. Department of Health and Human Services, *Research-based Web Design and Usability Guidelines*, xxi.

19. International Organization for Standardization, *ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)—Part 11: Guidance on Usability* (Geneva, Switzerland: International Organization for Standardization, 1998); International Organization for Standardization, *ISO 9241-12: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)—Part 12: Presentation of Information* (Geneva, Switzerland: International Organization for Standardization, 1997); International Organization for Standardization, *ISO 9241-14: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)—Part 14: Menu Dialogues* (Geneva, Switzerland: International Organization for Standardization, 1997).

20. U.S. Department of Health and Human Services, *Research-based Web Design and Usability Guidelines*.

21. Ivan Illich, *Limits to Medicine: Medical Nemesis: The Expropriation of Health* (Harmondsworth, N.Y.: Penguin, 1976).

22. D'Angelo and Twining, "Comprehension by Clicks"; Gerhardt-Powals, "Cognitive Engineering Principles."

23. C. Cheepen, "Guidelines for Dialogue Design—What is Our Approach? Working Design Guidelines for Advanced Voice Dialogues Project. Paper 3," (1996), www.soc.surrey.ac.uk/research/reports/voice-dialogues/wp3.html (accessed Dec. 29, 2005).

24. W. Crawford, "Webcats and Checklists: Some Cautionary Notes," *Information Technology and Libraries* 18, no. 2, (1999): 100–03; Cherry, "Bibliographic Displays in OPACs and Web Catalogs."

25. P. Reed et al., "User Interface Guidelines and Standards: Progress, Issues, and Prospects," *Interacting with Computers* 12, no. 1 (1999): 119–42.

26. D. C. L. Ngo, L. S. Teo, and J. G. Byrne, "Formalizing Guidelines for the Design of Screen Layouts," *Displays* 21, no. 1 (2000): 3–15.

27. D. A. Norman, "Emotion and Design: Attractive Things Work Better," *Interactions* 9, no. 4 (2002): 36–42.

28. N. Tractinsky, A. S. Katz, D. Ikar, "What is Beautiful is Usable," *Interacting with Computers* 13, no. 2 (2000): 127–45.