

# Classification of Library Resources by Subject on the Library Website: Is There an Optimal Number of Subject Labels?

Mathew J. Miles and  
Scott J. Bergstrom

*The number of labels used to organize resources by subject varies greatly among library websites. Some librarians choose very short lists of labels while others choose much longer lists. We conducted a study with 120 students and staff to try to answer the following question: What is the effect of the number of labels in a list on response time to research questions? What we found is that response time increases gradually as the number of the items in the list grow until the list size reaches approximately fifty items. At that point, response time increases significantly. No association between response time and relevance was found.*

It is clear that academic librarians face a daunting task drawing users to their library's Web presence. "Nearly three-quarters (73%) of college students say they use the Internet more than the library, while only 9% said they use the library more than the Internet for information searching."<sup>1</sup> Improving the usability of the library websites therefore should be a primary concern for librarians. One feature common to most library websites is a list of resources organized by subject. Libraries seem to use similar subject labels in their categorization of resources. However, the number of subject labels varies greatly. Some use as few as five subject labels while others use more than one hundred. In this study we address the following question: What is the effect of the number of subject labels in a list on response times to research questions?

## Literature review

McGillis and Toms conducted a performance test in which users were asked to find a database by navigating through a library website. They found that participants "had difficulties in choosing from the categories on the home page and, subsequently, in figuring out which database to select."<sup>2</sup>

A review of relevant research literature yielded a number of theses and dissertations in which the authors compared the usability of different library websites. Jeng in particular analyzed a great deal of the usability testing published concerning the digital library. The following are some of the points she summarized that were highly relevant to our study:

- User "lostness": Users did not understand the structure of the digital library.
- Ambiguity of terminology: Problems with wording accounted for 36 percent of usability problems.
- Finding periodical articles and subject-specific databases was a challenge for users.<sup>3</sup>

A significant body of research not specific to libraries provides a useful context for the present research. Miller's landmark study regarding the capacity of human short-term memory showed as a rule that the span of immediate memory is about  $7 \pm 2$  items.<sup>4</sup> Sometimes this finding is misapplied to suggest that menus with more than nine subject labels should never be used on a webpage. Subsequent research has shown that "chunking," which is the process of organizing items into "a collection of elements having strong associations with one another, but weak associations with elements within other chunks,"<sup>5</sup> allows human short-term memory to handle a far larger set of items at a time.

Larson and Czerwinski provide important insights into menuing structures. For example, increasing the depth (the number of levels) of a menu harms search performance on the Web. They also state that "as you increase breadth and/or depth, reaction time, error rates, and perceived complexity will all increase."<sup>6</sup> However, they concluded that a "medium condition of breadth and depth outperformed the broadest, shallow web structure overall."<sup>7</sup> This finding is somewhat contrary to a previous study by Snowberry, Parkinson, and Sisson, who found that when testing structures of 2<sup>6</sup>, 4<sup>3</sup>, 8<sup>2</sup>, 64<sup>1</sup> (2<sup>6</sup> means two menu items per level, six levels deep), the 64<sup>1</sup> structure grouped into categories proved to be advantageous in both speed and accuracy.<sup>8</sup> Larson and Czerwinski recommended that "as a general principle, the depth of a tree structure should be minimized by providing broad menus of up to eight or nine items each."<sup>9</sup>

Zaphiris also corroborated that previous research concerning depth and breadth of the tree structure was true for the Web. The deeper the tree structure, the slower the user performance.<sup>10</sup> He also found that response times for expandable menus are on average 50 percent longer than sequential menus.<sup>11</sup> Both the research and current practices are clear concerning the efficacy of hierarchical menu structures. Thus it was not a focus of our research. The focus instead was on a single-level menu and how the number and characteristics of subject labels would affect search response times.

## Background

In preparation for this study, library subject lists were collected from a set of thirty library websites in the United

---

Mathew J. Miles (milesm@byui.edu) is Systems Librarian and Scott J. Bergstrom (bergstroms@byui.edu) is Director of Institutional Research at Brigham Young University-Idaho in Rexburg.

---

States, Canada, and the United Kingdom. We selected twelve lists from these websites that were representative of the entire group and that varied in size from small to large. To render some of these lists more usable, we made slight modifications. There were many similarities between label names.

## Research design

Participants were randomly assigned to one of twelve experimental groups. Each experimental group would be shown one of the twelve lists that were selected for use in this study. Roughly 90 percent of the participants were students. The remaining 10 percent of the participants were full-time employees who worked in these same departments. The twelve lists ranged in number of labels from five to seventy-two:

- Group A: 5 subject labels
- Group B: 9 subject labels
- Group C: 9 subject labels
- Group D: 23 subject labels
- Group E: 6 subject labels
- Group F: 7 subject labels
- Group G: 12 subject labels
- Group H: 9 subject labels
- Group I: 35 subject labels
- Group J: 28 subject labels
- Group K: 49 subject labels
- Group L: 72 subject labels

Each participant was asked to select a subject label from a list in response to eleven different research questions. The questions are listed below:

1. Which category would most likely have information about modern graphical design?
2. Which category would most likely have information about the Aztec Empire of ancient Mexico?
3. Which category would most likely have information about the effects of standardized testing on high school classroom teaching?
4. Which category would most likely have information on skateboarding?
5. Which category would most likely have information on repetitive stress injuries?
6. Which category would most likely have information about the French Revolution?
7. Which category would most likely have information concerning Walmart's marketing strategy?
8. Which category would most likely have information on the reintroduction of wolves into Yellowstone Park?

9. Which category would most likely have information about the effects of increased use of nuclear power on the price of natural gas?
10. Which category would most likely have information on the Electoral College?
11. Which category would most likely have information on the philosopher Emmanuel Kant?

The questions were designed to represent a variety of subject areas that library patrons might pursue. Each subject list was printed on a white sheet of paper in alphabetical order in a single column, or double columns when needed. We did not attempt to test the subject lists in the context of any Web design. We were more interested in observing the effect of the number of labels in a list on response time independent of any Web design. Each participant was asked the same eleven questions in the same order. The order of questions was fixed because we were not interested in testing for the effect of order and wanted a uniform treatment, thereby not introducing extraneous variance into the results.

For each question, the participant was asked to select a label from the subject list under which they would expect to find a resource that would best provide information to answer the question. Participants were also instructed to select only a single label, even if they could think of more than one label as a possible answer. Participants were encouraged to ask for clarification if they did not fully understand the question being asked. Recording of response times did not begin until clarification of the question had been given. Response times were recorded unbeknownst to the participant. If the participant was simply unable to make a selection, that was also recorded. Two people administered the exercise. One recorded response times; the other asked the questions and recorded label selections.

Relevance rankings were calculated for each possible combination of labels within a subject list for each question. For example, if a subject list consisted of five labels, for each question there were five possible answers. Two library professionals—one with humanities expertise, the other with sciences expertise—assigned a relevance ranking to every possible combination of question and labels within a subject list. The rankings were then averaged for each question-label combination.

## Results

The analysis of the data was undertaken to determine whether the average response times of participants, adjusted by the different levels of relevance in the subject list labels that prevailed for a given question, were significantly different across the different lists. In other words, would the response times of participants using a particular list, for whom the labels in the list were highly relevant

to the question, be different from students using the other lists for whom the labels in the list were also highly relevant to the question?

A separate univariate general linear model analysis was conducted for each of the eleven questions. The analyses were conducted separately because each question represented a unique search domain. The univariate general linear model provided a technique for testing whether the average response times associated with the different lists were significantly different from each other. This technique also allowed for the inclusion of a covariate—relevance of the subject list labels to the question—to determine whether response times at an equivalent level of relevance was different across lists.

In the analysis model, the dependent variable was response time, defined as the time needed to select a subject list label. The covariate was relevance, defined as the perceived match between a label and the question. For example, a label of “Economics” would be assessed as highly relevant to the question, what is the current unemployment rate? The same label would be assessed as not relevant for the question, what are the names of four moons of Saturn? The main factor in the model was the actual list being presented to the participant. There were twelve lists used in this study. The statistical model can be summarized as follows:

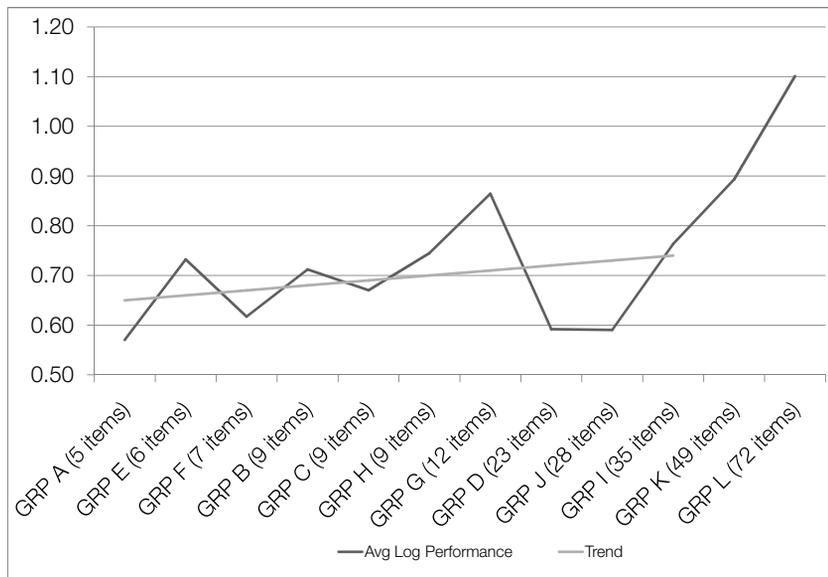
$$\text{response time} = \text{list} + \text{relevance} + (\text{list} \times \text{relevance}) + \text{error}$$

The general linear model required that the following conditions be met: First, data must come from a random sample from a normal population. Second, all variances with each of the groupings are the same (i.e., they have homoscedasticity). An examination of whether these assumptions were met revealed problems both with normality and with homoscedasticity. A common technique—logarithmic transformation—was employed to resolve these problems. Accordingly, response-time data were all converted to common logarithms. An examination of assumptions with the transformed data showed that all questions but three met the required conditions. The three

questions (5, 6, and 7) were excluded from subsequent analysis.

## Conclusions

The series of graphs in the appendix show the average response times, adjusted for relevance, for eight of the eleven questions for all twelve lists (i.e., experimental groups). Three of the eleven questions were excluded from the analysis because of heteroscedasticity. An inspection of these graphs shows no consistent pattern in response time as the number of the items in the lists increase. Essentially, this means that, for any given level of relevance, the number of items of the list does not affect response time significantly. It seems that for a single question, characteristics of the categories themselves are more important than the quantity of categories in the list. The response times using a subject list with twenty-eight labels is similar to the response times using a list of six labels. A statistical comparison of the mean response time for each



**Figure 1.** The overall average of average search times for the eight questions for all experimental groups (i.e., lists)

group with that of each of the other groups for each of the questions largely confirms this. There were very few statistically significant different comparisons. The spikes and valleys of the graphs in the appendix are generally not significantly different. However, when the average response time associated with all lists is combined into an overall average from all eight questions, a somewhat clearer picture emerges (see figure 1). Response times increase gradually as the number of the items in the list increase until the list size reaches approximately fifty items. At that point, response time increases significantly. No association was found between response time and relevance. A fast response time did not necessarily yield a relevant response, nor did a slow response time yield an irrelevant response.

## Observations

We observed that there were two basic patterns exhibited when participants made selections. The first pattern was the quick selection—participants easily made a selection after performing an initial scan of the available labels. Nevertheless, a quick selection did not always mean a relevant selection. The second pattern was the delayed selection. If participants were unable to make a selection after the initial scan of items, they would hesitate as they struggled to determine how the question might be reclassified to make one of the labels fit. We did not have access to a high-tech lab, so we were unable to track eye movement, but it appeared that the participants began scanning up and down the list of available items in an attempt to make a selection. The delayed selection seemed to be a combination of two problems: First, none of the available labels seemed to fit. Second, the delay in scanning increased as the list grew larger. It's possible that once the list becomes large enough, scanning begins to slow the selection process. A delayed selection did not necessarily yield an irrelevant selection.

The label names themselves did not seem to be a significant factor affecting user performance. We did test three lists, each with nine items and each having different labels, and response times were similar for the three lists. A future study might compare a more extensive number of lists with the same number of items with different labels to see if label names have an effect on response time. This is a particular challenge to librarians

in classifying the digital library, since they must come up with a few labels to classify all possible subjects.

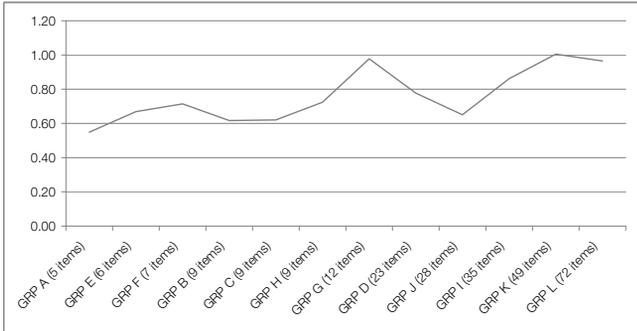
Creating eleven questions to span a broad range of subjects is also a possible weakness of the study. We had to throw out three questions that violated the assumptions of the statistical model. We tried our best to select questions that would represent the broad subject areas of science, arts, and general interest. We also attempted to vary the difficulty of the questions. A different set of questions may yield different results.

## References

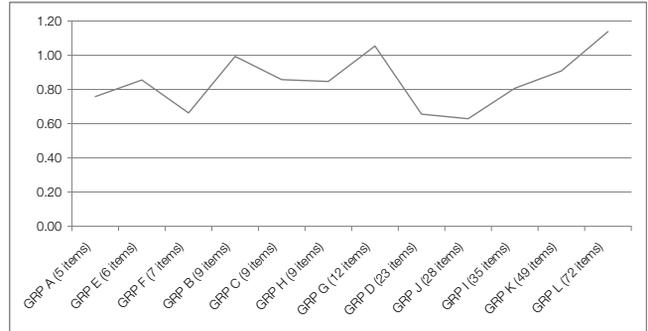
1. Steve Jones, *The Internet Goes to College*, ed. Mary Madden (Washington, D.C.: Pew Internet and American Life Project, 2002): 3, [www.pewinternet.org/pdfs/PIP\\_College\\_Report.pdf](http://www.pewinternet.org/pdfs/PIP_College_Report.pdf) (accessed Mar. 20, 2007).
2. Louise McGillis and Elaine G. Toms, "Usability of the Academic Library Web Site: Implications for Design," *College & Research Libraries* 62, no. 4 (2001): 361.
3. Judy H. Jeng, "Usability of the Digital Library: An Evaluation Model" (PhD diss., Rutgers University, New Brunswick, New Jersey): 38–42.
4. George A. Miller, "The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review* 63, no. 2 (1956): 81–97.
5. Fernand Gobet et al., "Chunking Mechanisms in Human Learning," *Trends in Cognitive Sciences* 5, no. 6 (2001): 236–43.
6. Kevin Larson and Mary Czerwinski, "Web Page Design: Implications of Memory, Structure and Scent for Information Retrieval" (Los Angeles: ACM/Addison-Wesley, 1998): 25, <http://doi.acm.org/10.1145/274644.274649> (accessed Nov. 1, 2007).
7. Ibid.
8. Kathleen Snowberry, Mary Parkinson, and Norwood Sisson, "Computer Display Menus," *Ergonomics* 26, no. 7 (1983): 705.
9. Larson and Czerwinski, "Web Page Design," 26.
10. Panayiotis G. Zaphiris, "Depth vs. Breath in the Arrangement of Web Links," [www soi.city.ac.uk/~zaphiri/Papers/hfes.pdf](http://www soi.city.ac.uk/~zaphiri/Papers/hfes.pdf) (accessed Nov. 1, 2007).
11. Panayiotis G. Zaphiris, Ben Shneiderman, and Kent L. Norman, "Expandable Indexes Versus Sequential Menus for Searching Hierarchies on the World Wide Web," <http://citeseer.ist.psu.edu/rd/0%2C443461%2C1%2C0.25%2CDownload/http://coblitiz.codeen.org:3125/citeseer.ist.psu.edu/cache/papers/cs/22119/http:SzzSzagrino.orgzSzzpaphirizSzPaperszSzexpandableindexes.pdf/zaphiris99expandable.pdf> (accessed Nov. 1, 2007).

## APPENDIX. Response times by question by group

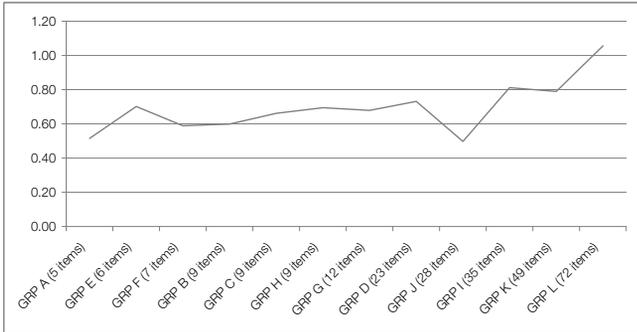
Question 1



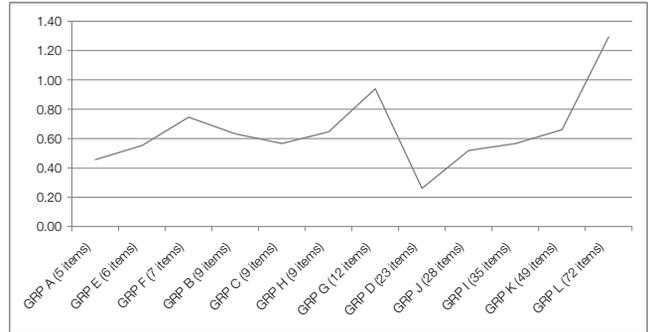
Question 8



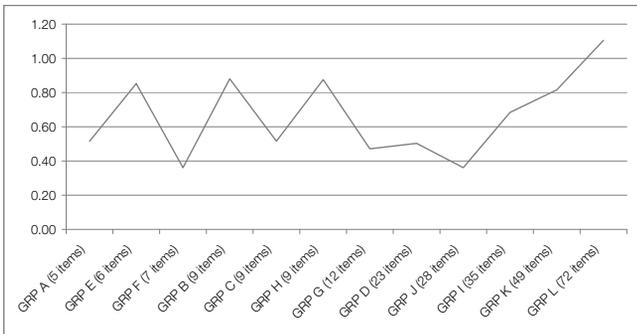
Question 2



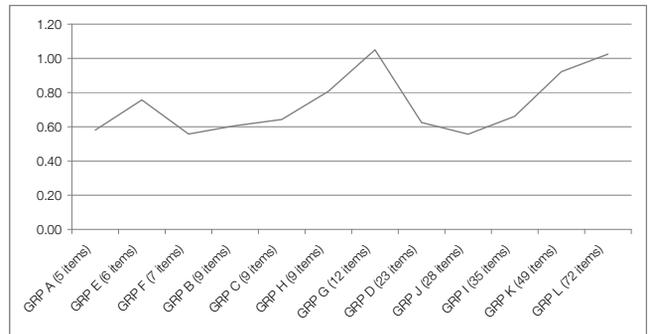
Question 9



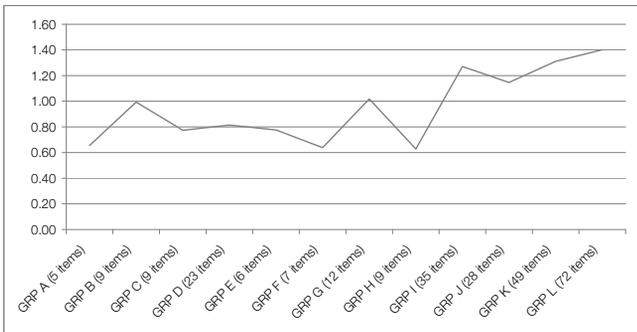
Question 3



Question 10



Question 4



Question 11

