Using Closed Card-Sorting to Evaluate Information Architectures

Thomas S. Tullis
Senior VP, User Experience
Fidelity Investments
tom.tullis@fidelity.com

ABSTRACT:

A technique using closed card-sorting to evaluate candidate information architectures for a web site is described. Participants in an online card-sorting study are randomly directed to one of the architectures being evaluated. Each participant is shown the same cards but different categories to sort them into. The basic data collected is simply which cards each user put into which groups. For any one architecture being tested, the data show what percentage of the participants put each card into each group. A better architecture is one where the participants were more consistent with each other in terms of which groups they put each of the cards into. The basic "score" proposed for each card is the percentage associated with the "winning" group (i.e., the group with the highest percentage)—the higher that percentage is, the better. A consistency score for each architecture tested can then be calculated by taking an average of these percentages across all the cards. A technique for correcting this score when the different architectures have different numbers of groups is also described.

Introduction

Card-sorting has been widely used as a technique for getting input from users about how a website or other information structure should be organized. Perhaps the most common type of card-sorting study is an "open" sort, where the users are given a set of cards which they sort into groups that they create and name. This is useful when embarking on the design of a new website or a redesign of an existing website. Since the users name the groups that they create, it's also a great way to learn about the terms that they use in thinking about the cards and their relationships with each other.

While an open sort can give you a great deal of insight into organizing a website, it doesn't give you a magic solution. The data from an open sort should be used as one input to the design or information architecture of a website. After doing an open sort, it's quite common to have several different candidates for the information architecture of the website. These may represent different "slices" through the open sort data (i.e., different numbers of groups) or different terminology for the groups. That's when "closed" card-sorting can be useful as a way of comparing the effectiveness of these different architectures to each other.

In a closed card-sort, the users are presented with the cards to be sorted AND the names of groups that the cards should be sorted into. Figure 1 shows an example of an online closed card-sort.

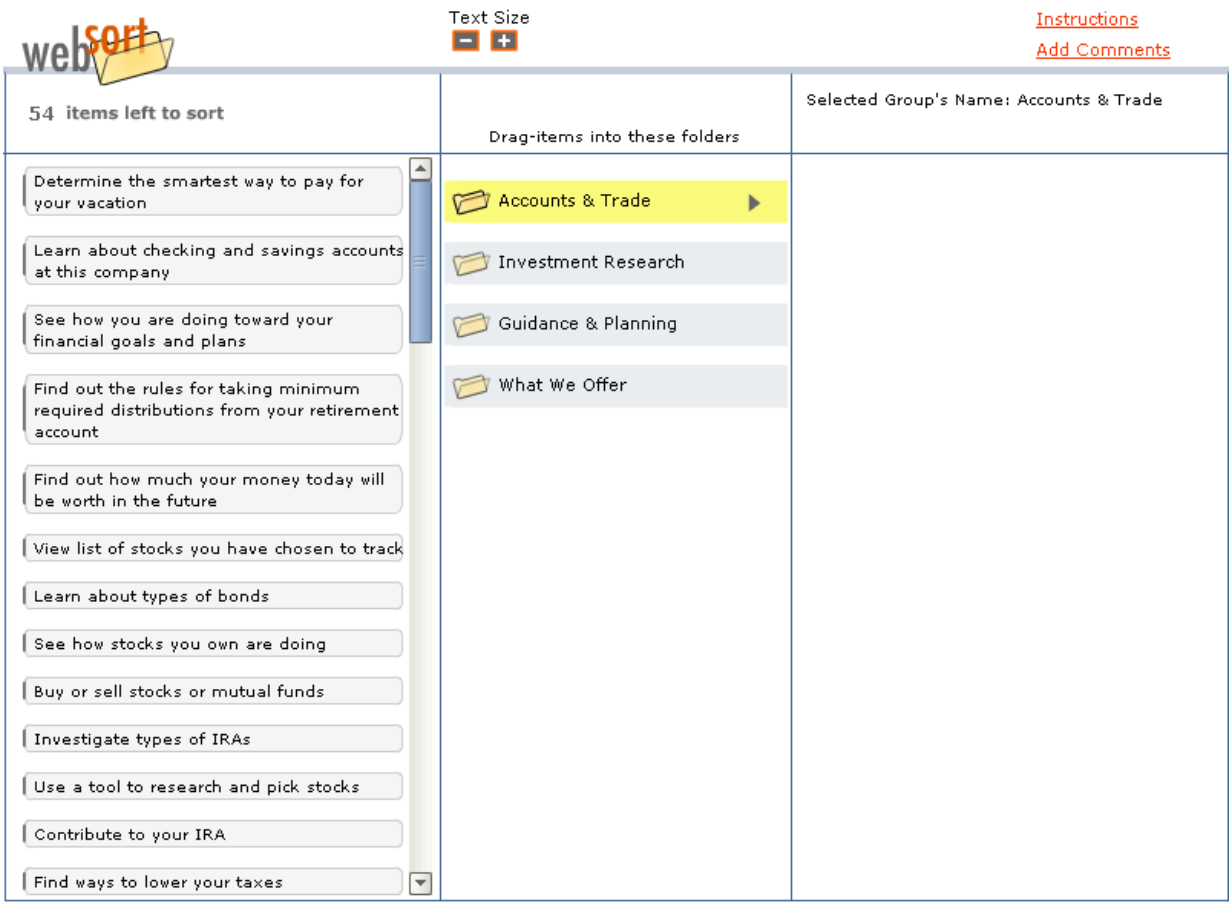

Figure 1. An example of an online closed card-sort. In this example, 54 functions of a hypothetical financial-services website are shown on the left. The user must sort them into the four predefined groups in the middle. This example uses the Websort online tool (www.Websort.net).

The New Technique

There's nothing new about open or closed card-sorting. What is new is the method we have developed for conducting simultaneous closed card-sorts of different candidate architectures and then evaluating the results to determine which ones worked the best for the users.

Let's assume that you have three different information architectures being considered for a website. Each of these information architectures, or frameworks, would be represented in a different closed card-sort. Since the contents of the website are the same in each case, the same set of cards would be used in each card-sort. In order to get large numbers of participants quickly, we have done these as online card-sorting studies. A "launch" page that describes the study is created and all participants are directed to it. Background information is collected by that page if needed. From that page, participants are then randomly directed to one of the three closed card-sorting studies.

The data that you get from each participant is simply which cards that person put into each of the groups. The primary goal of the analysis is to see how consistent the participants were with each other in placing the cards into groups. A framework where the participants were far more consistent in placing the cards into groups is considered to be a better framework than one where they were much less consistent.

Consider the following data, which is from a real closed card-sort (although the actual cards and categories are not identified):

Card	Category A	Category B	Category C
Card #1	17%	78%	5%
Card #2	15%	77%	8%
Card #3	20%	79%	1%
Card #4	48%	40%	12%
Card #5	11%	8%	81%
Card #6	1%	3%	96%
Card #7	46%	16%	37%
Card #8	57%	38%	5%
Card #9	20%	75%	5%
Card #10	4%	5%	92%

These percentages represent the percentage of the participants in the study who put each card into each of the three categories. Each row sums to 100% (although there may be rounding error). In this example, there was far more agreement among the participants about what category Card #6 belonged in (namely Category C, which 96% of the users put it in) than, say, Card #7, where they were more distributed across all three categories.

One metric we have used to measure the degree of agreement among the participants in the study who used a given framework is simply the average of the maximum percentages across all the cards, as follows for the example just shown:

Card	Category A	Category B	Category C	Max
Card #1	17%	78%	5%	**78%**
Card #2	15%	77%	8%	**77%**
Card #3	20%	79%	1%	**79%**
Card #4	48%	40%	12%	**48%**
Card #5	11%	8%	81%	**81%**
Card #6	1%	3%	96%	**96%**
Card #7	46%	16%	37%	**46%**
Card #8	57%	38%	5%	**57%**
Card #9	20%	75%	5%	**75%**
Card #10	4%	5%	92%	**92%**
			Average:	**73%**

So the “Max” entry for each card is simply the percentage associated with the “winning” category for that card. In this example, those percentages average to 73%. The higher that average is, the more agreement there was among the participants in sorting the cards into the categories.

This average can then be used to make comparisons between the frameworks tested. For example, consider the following data from the same study (i.e., the same cards) but where the participants were presented with three other categories to sort the cards into:

Card	Category X	Category Y	Category Z	Max
Card #1	23%	62%	14%	**62%**
Card #2	27%	54%	19%	**54%**
Card #3	47%	40%	13%	**47%**
Card #4	53%	30%	16%	**53%**
Card #5	6%	12%	82%	**82%**
Card #6	5%	6%	89%	**89%**
Card #7	47%	15%	37%	**47%**
Card #8	85%	14%	2%	**85%**
Card #9	35%	60%	4%	**60%**
Card #10	6%	20%	74%	**74%**
			Average:	**65%**

This 65% average reflects that there was less agreement among the participants in using this framework than the previous one which had a 73% average. By treating the category with the highest percentage for each card as the "winning" category, it is also possible to calculate a score for each individual participant which reflects the percentage of cards that person put into the "winning" categories. In essence, this can be thought of as an "accuracy" score for each person: ranging from 0% (if that person put none of the cards into the "winning" categories) to 100% (if that person put all of the cards into the "winning" categories). Having these individual participant scores then allows you to calculate a confidence interval for the averages outlined above. This is illustrated here with data from a real study just completed in which six different frameworks for organizing the same functions were tested:

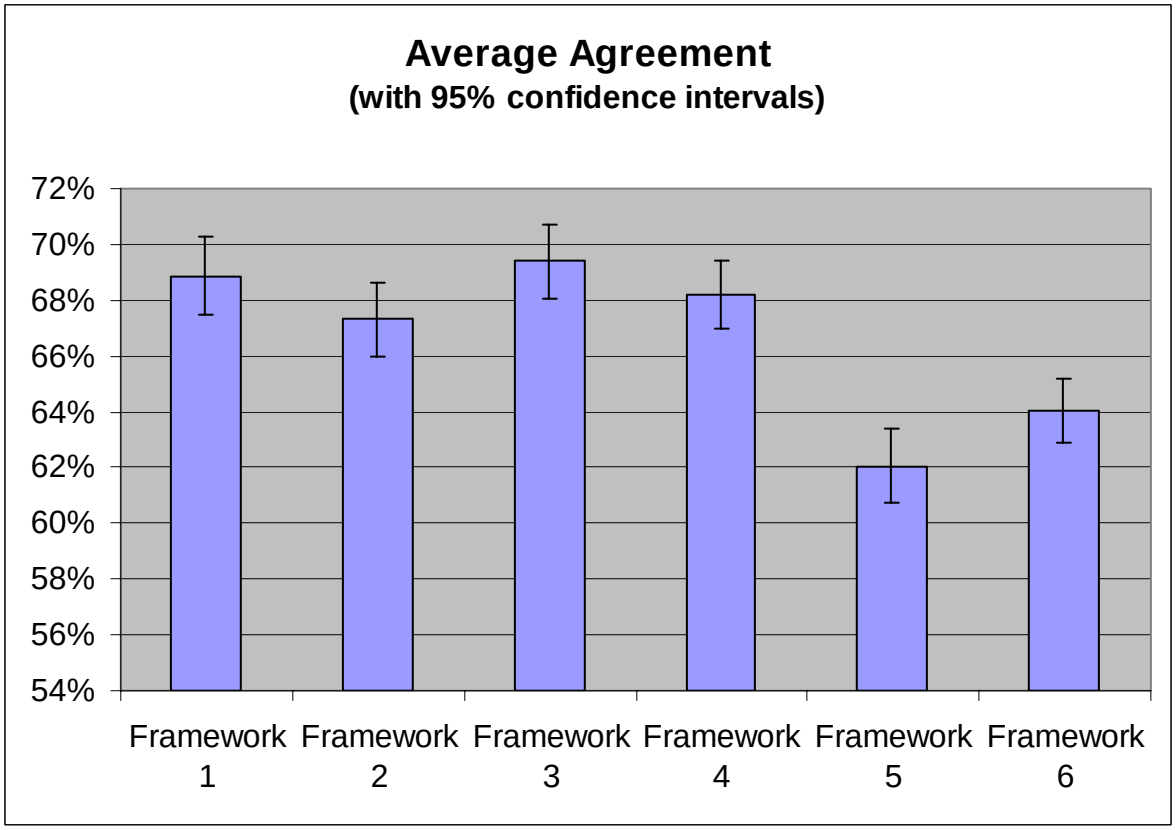

Figure 2. Data from a closed card-sort in which six different frameworks were evaluated.

As can be seen from these data, Frameworks 5 and 6 performed significantly worse than Frameworks 1 through 4.

But what if the different frameworks you want to compare have different numbers of categories? If users were performing randomly in the sorting exercise, you would expect a 3-category framework to yield an average of 33% using this method while a 10-category framework would yield an average of 10%. We've experimented with various ways of correcting for this problem. The solution that seems to work best is to simply look at the difference, for each card, between the percentage associated with the "winning" category and the percentage associated with the "2nd place" category, as in this example:

Card	Category A	Category B	Category C	Max	2nd Place	Difference
Card #1	17%	78%	5%	**78%**	**17%**	**61%**
Card #2	15%	77%	8%	**77%**	**15%**	**62%**
Card #3	20%	79%	1%	**79%**	**20%**	**60%**
Card #4	48%	40%	12%	**48%**	**40%**	**8%**
Card #5	11%	8%	81%	**81%**	**11%**	**70%**
Card #6	1%	3%	96%	**96%**	**3%**	**93%**
Card #7	46%	16%	37%	**46%**	**37%**	**8%**
Card #8	57%	38%	5%	**57%**	**38%**	**18%**
Card #9	20%	75%	5%	**75%**	**20%**	**55%**
Card #10	4%	5%	92%	**92%**	**5%**	**87%**
			Average:	**73%**		**52%**

So in this example, the cards that did "well" are the ones that yielded a large difference between the max percentage and the 2nd-place percentage (such as Card #6, with a difference of 93%). The ones that did poorly are the ones that yielded a small difference (such as Card #4, with a difference of only 8%). The average of these differences can then be used to make comparisons between frameworks that have different numbers of categories. In essence, it's a measure of how strongly the winning category drew cards to it.

Conclusion

We've now used this technique in a variety of studies. We've evaluated frameworks ranging from 3 to 10 categories and numbers of cards ranging from 24 to 54. Our studies have had as many as 1,786 participants each. We've successfully used it to compare a wide variety of information architectures.

BIBLIOGRAPHY

Ahlstrom, V., and Allendoerfer, K. (2004) Information Organization for a Portal Using a Card-Sorting Technique. DOT/FAA/CT-TN04/31 Technical Report. Available at http://acb220.tc.faa.gov/technotes/dot-faa-ct-tn04-31.pdf#search=%22card-sorting%20technique%22.

Dong, J. (2002) A Methodology to Verify and Improve an Existing Large-scale Information Architecture. UPA 2002 presentation, Orlando, Florida.

Faiks, A., and Hyland, N. (2000) Gaining user insight: a case study illustrating the card sort technique. College and Research Libraries, Volume 61, 2000, 349-357. Available at http://www.ala.org/ala/acrl/acrlpubs/crljournal/backissues2000b/july00/faiks.pdf.

Tullis, T. (2003) Using Card-sorting Techniques to Organize Your Intranet. *Intranet Journal of Strategy and Management*, March 2003.

Tullis, T., and Wood, L. (2004) How Many Users Are Enough for a Card-Sorting Study? Poster presentation at UPA 2004 Conference, Minneapolis, MN.

Tullis, T., and Wood, L. (2005) How Can You Do a Card-sorting Study with LOTS of Cards? Poster presentation at UPA 2005 Conference, Montreal, Canada.