Topics

E-commerce

Intranets

Mobile & Tablet

User Testing

Web Usability

See all topics...

Author

Jakob Nielsen

Don Norman

Bruce "Tog" Tognazzini

See all authors...

Recent Articles

Four Dangerous Navigation Approaches that Can Increase Cognitive Strain

User Expertise Stagnates at Low Levels

User-centric vs. Maker-centric Language: 3 Essential Guidelines

Making Usability Findings Actionable: 5 Tips for Writing Better Reports

Mobile: Native Apps, Web Apps, and Hybrid Apps

See all articles...

Popular Articles

Usability 101: Introduction to Usability

Top 10 Mistakes in Web Design

How Users Read on the Web

Subscribe to:

Jakob Nielsen's Alertbox Newsletter

Card Sorting: How Many Users to Test

by **JAKOB NIELSEN** on July 19, 2004

Topics: Information Architecture Research Methods

Summary: Testing ever-more users in card sorting has diminishing returns, but you should still use three times more participants than you would in traditional usability tests.

One of the biggest challenges in website and intranet design is creating the **information architecture** : what goes where? A classic mistake is to structure the information space based on how *you* view the content — which often results in different subsites for each of your company's departments or information providers.

Rather than simply mirroring your org chart, you can better enhance usability by creating an information architecture that reflects how *users* view the content. In each of our intranet studies , we've found that some of the biggest productivity gains occur when companies restructure their intranet to reflect employees' workflow. And in e-commerce , sales increase when products appear in the categories where users expect to find them.

All very good, but *how do you find out* the users' view of an information space and where they think each item should go? For **researching this type of mental model** , the primary method is card sorting:

1. Write the name (and perhaps a short description) of each of the main items on an index card. Yes, good old paper cards. (Taking care not to use terms that bias the users.)
2. Shuffle the cards and give the deck to a user. (The standard recommendations for recruiting test participants apply: they must be representative users, etc.)
3. Ask each user to sort the cards into piles, placing items that belong together in the same pile. Users can make as many or as few piles as they want; some piles can be big, others small.
4. Optional extra steps include asking users to arrange the resulting piles into bigger groups, and to name the different groups and piles. The latter step can give you ideas for words and synonyms to use for navigation labels, links, headlines, and search engine optimization.

Because card sorting uses no technology, this photo of a 1995 card sort looks the same as one conducted today.

Research Study

Fidelity Investments has one of the world's best usability teams, led by Dr. Thomas S. Tullis, senior VP of human interface design. Tullis and co-author Larry Wood recently reported the results of a study measuring the trade-off curve for testing various numbers of users in a card sorting exercise.

First, they tested 168 users, generating very solid results. They then simulated the outcome of running card sorting studies with smaller user groups by analyzing random subsets of the total dataset. For example, to see what a test of twenty users would generate, they selected twenty users randomly from the total set of 168 and analyzed only that subgroup's card sorting data. By selecting many such samples, it was possible to estimate the average findings from testing different numbers of users.

The main quantitative data from a card sorting study is a set of **similarity scores** that measures the similarity of user ratings for various item pairs. If all users sorted two cards into the same pile, then the two items represented by the cards would have 100% similarity. If half the users placed two cards together and half placed them in separate piles, those two items would have a 50% similarity score.

We can assess the outcome of a smaller card sorting study by asking how well its similarity scores correlate with the scores derived from testing a large user group. (A reminder: correlations run from -1 to +1. A correlation of 1 shows that the two datasets are perfectly aligned; 0 indicates no relationship; and negative correlations indicate datasets that are opposites of each other.)

How Many Users?

For most usability studies, I recommend testing five users , since that's enough data to teach you most of what you'll ever learn in a test. For card sorting, however, there's only a 0.75 correlation between the results from five users and the ultimate results. That's not good enough.

You must test fifteen users to reach a correlation of 0.90, which is a more comfortable place to stop. After fifteen users, diminishing returns set in and correlations increase very little: testing thirty people gives a correlation of 0.95 -- certainly better, but usually not worth twice the money. There are hardly any improvements from going beyond thirty users: you have to test sixty people to reach 0.98, and doing so is definitely wasteful.

Tullis and Wood recommend testing twenty to thirty users for card sorting. Based on their data, **my recommendation is to test fifteen users** .

Why do I recommend testing fewer users? I think that correlations of 0.90 (for fifteen users) or maybe 0.93 (for twenty) are good enough for most practical purposes. I can certainly see testing thirty people and reaching 0.95 if you have a big, well-funded project with a lot of money at stake (say, an intranet for 100,000 employees or an e-commerce site with half a billion dollars in revenues). But most projects have very limited resources for user research; the remaining fifteen users are better "spent" on three qualitative usability tests of different design iterations.

Also, I don't recommend designing an information architecture based purely on a card sort's numeric similarity scores. When deciding specifics of what goes where, you should rely just as much on the qualitative insights you gain in the testing sessions. Much of the value from card sorting comes from **listening to the users' comments** as they sort the cards: knowing *why* people place certain cards together gives deeper insight into their mental models than the pure fact that they sorted cards into the same pile.

Why More Users for Card Sorting?

We know that five users are enough for most usability studies, so why do we need three times as many participants to reach the same level of insight with card sorting? Because the methods differ in two key ways:

- User testing is an **evaluation method** : we already have a design, and we're trying to find out whether or not it's a good match with human nature and user needs. Although people differ substantially in their capabilities (domain knowledge, intelligence, and computer skills), if a certain design element causes difficulties, we'll see so after testing a few users. A low-end user might experience more severe difficulties than a high-end user, but the magnitude of the difficulties is not at issue unless you are running a measurement study (which requires more users). All you need to know is that the design element doesn't work for humans and should be changed.
- Card sorting is a **generative method** : we don't yet have a design, and our goal is to find out how people think about certain issues. There is great variability in different people's mental models and in the vocabulary they use to describe the same concepts. We must collect data from a fair number of users before we can achieve a stable picture of the users' preferred structure and determine how to accommodate differences among users.

If you have an existing website or intranet, testing a few users will tell you whether people have trouble with the information architecture. To generate a new structure from scratch, you must sample more people.

Luckily, you can **combine the two methods** : First, use generative studies to set the direction for your design. Second, draft up a design, preferably using paper prototyping , and run evaluation studies to refine the design. Because usability evaluations are fast and cheap, you can afford multiple rounds; they also provide quality assurance for your initial generative findings. This is why you shouldn't waste resources squeezing the last 0.02 points of correlation out of your card sorts. You'll catch any small mistakes in subsequent user testing, which will be much cheaper than doubling or tripling the size of your card sorting studies.

Study Weaknesses

The Fidelity study has two obvious weaknesses:

- It's only one study. It's always better to have data from multiple companies.
- The analysis was purely quantitative, focusing on a statistical analysis of similarity scores and ignoring user comments and other qualitative data.

These two weaknesses are not fatal. I view this as a pioneering study and a great contribution to our Web usability knowledge. But, because of the study's weaknesses, it would be useful if somebody duplicated it with different information spaces, and also analyzed the qualitative data along with the numeric scores. Sounds like a good thesis project for a graduate student who's looking to research something with real-world impact (hint, hint).

Even though more data would be comforting, I have confidence in the Fidelity study's conclusions because they match my own observations from numerous card studies over many years. I've always said that it was necessary to test more users for card sorting than for traditional usability studies. And I've usually recommended about fifteen users, though we've also had good results with as few as twelve when budgets were tight or users were particularly hard to recruit.

There are myriad ways in which quantitative studies can go wrong and mislead you. Thus, if you see a single quantitative study that contradicts all that's known from qualitative studies, it's prudent to disregard the new study and assume that it's likely

to be bogus. But when a quantitative study confirms what's already known, it's likely to be correct, and you can use the new numbers as decent estimates, even if they're based on less data than you would ideally like.

Thus, the current recommendation is to **test fifteen users for card sorting** in most projects, and thirty users in big projects with lavish funding.

Learn More

Research Reports

Intranet Information Architecture Design Methods and Case Studies
Vol. 07: Navigation and Page Layout
Site Map Usability

Training Courses

Information Architecture: Day 1 and Day 2
Usability in Practice: 3-Day Camp
Measuring Usability
Advanced User Testing
University Websites

Articles

Top 10 Information Architecture (IA) Mistakes
How to Conduct a Heuristic Evaluation
Mini-IA: Structuring the Information About a Concept
Alphabetical Sorting Must (Mostly) Die
Card Sorting: Pushing Users Beyond Terminology Matches