Herd Behavior in Assessments of Web Interface Design: Some Preliminary Evidence

Emergent Research Forum Paper

Daniel S. Soper
Department of Information Systems and Decision Sciences
Mihaylo College of Business and Economics
California State University, Fullerton
dsoper@fullerton.edu

Abstract

This paper reports on the preliminary findings of a large study aimed at investigating the role of herding in assessments of web interface design. In the context of human behavior, herding occurs when a human being unconsciously bases his or her decisions on the actions or opinions of others. Using a controlled, randomized experiment involving 678 subjects, three different web interfaces, and five different interface design characteristics, the preliminary findings indicate that herd behavior exerts a strong and highly significant influence on web interface design assessments. The effect is particularly pronounced when subjects are provided with experimentally manipulated interface design ratings from people whom the subjects believe to be very similar to themselves. Since websites commonly serve as the public face of modern organizations, these findings have obvious and important implications insofar as web design influences how an organization is perceived and its subsequent prospects for success.

Keywords

Herd behavior, web interface design, interface evaluation

Introduction and Research Questions

Herd behavior refers to the tendency of human beings to base our decisions on the actions and opinions of others (Banerjee 1992). Such behavior is typically unconscious, and has been identified in a wide variety of human endeavors including investment decisions (Avery and Zemsky 1998), management (Palley 1995), voting (Coleman 2004), restaurant selection (Banerjee 1992), fertility decisions (Watkins 1990), technology adoption (Li 2004), and so forth. In light of the increasingly interconnected nature of the modern world, some researchers have even suggested that we may now be more susceptible to the effects of herding than at any prior time in human history (Raafat et al. 2009).

Despite the large array of phenomena that have been linked to herding, the proclivity of humans to engage in herd behavior has attracted very little attention in mainstream information systems (IS) research. For this reason, I would like to report in this emergent research forum paper on the progress of a large study aimed at investigating the role of herding in a very common IS-related task in which most of us unconsciously engage on a regularly basis – judging the quality of a web interface design. In its current preliminary phase, the study seeks to provide insights into the following two research questions:

1. Does a-priori knowledge of the opinions of others influence our assessments of web interface design?
2. As the perceived degree of similarity between ourselves and the members of a reference group increases, are we more likely to rely on the group's opinion when rating a web interface?

Although this study is certainly still a work in progress, I nevertheless hope that the reader will find the brief report provided herein to be of interest.

Research Design and Methodology

Preliminary insights into the research questions above were gained by means of a controlled, randomized experiment. Inasmuch as the target population for the experiment was English-speaking adult web users, the leading global online advertising firm was engaged to craft a targeted campaign for the purpose of soliciting volunteers for the study. The firm's technology allowed subject recruitment to be explicitly limited to English-speaking web users who were at least 18 years old. IP address restrictions were also enforced to help ensure that each subject could participate in the experiment only once. Upon agreeing to participate in the experiment, subjects were asked to specify their age and gender, after which they were allocated into one of four experimental groups. Age and gender were explicitly included in the study because they have been identified by past research as the primary features by which people, in the absence of other information, unconsciously judge the degree of similarity between themselves and others (Brewer and Lui 1989). In total, data were gathered from 678 subjects, of whom 329 (48.5%) were female and 349 (51.5%) were male. Subjects ranged in age from 18 to 80 years, with the mean age being 32.91 years (std dev = 11.68). These demographic characteristics were observed to be consistent with the overall population of adult web users (Pew Research Center 2014).

The experiment itself was carried out using a custom, web-based software system. As their primary task, subjects were asked to evaluate the characteristics of three web interfaces, each of which was intentionally designed according to the general mental model of web interface design identified by Soper and Mitra (2013). The specific characteristics that were evaluated for each interface were adopted from a pre-validated, five-item subscale created to measure the attractiveness of a web interface (Aladwania and Palvia 2002). In accordance with the original instrument, subjects in the experiment were asked to respond to the evaluative statements using a seven-point, Likert-type scale anchored at 1 = *strongly disagree* and 7 = *strongly agree*. Minor modifications were made to the wording of the items in order to adapt those items to the context of the current experiment (see Table 1).

Original Statement (Aladwania and Palvia 2002)	Modified Statement Used in Current Experiment
____'s website looks attractive.	This website looks attractive.
____'s website looks organized.	This website looks organized.
____'s website uses fonts properly.	This website uses fonts properly.
____'s website uses colors properly.	This website uses colors properly.
____'s website uses multimedia features properly.	This website uses multimedia features properly.

Table 1. Original and modified subscale items.

Each subject was required to evaluate all three web interfaces along just one of the dimensions listed in Table 1 so as to minimize the possibility that her ratings would be contaminated by halo error (Soper 2014b). The specific design characteristic that each subject was asked to evaluate was determined using iterative assignment, and the order in which the three web interfaces were presented to each subject was randomized in order to mitigate any ordering or self-generated validity effects (Chandon et al. 2005; Saris and Gallhofer 2007).

After specifying their age and gender subjects were allocated into one of four groups, which included a baseline group and three experimentally manipulated "treatment" groups. Subjects in the baseline group were simply shown the three interfaces and asked to rate each interface along their assigned dimension. When aggregated, the responses from baseline subjects were regarded as the true, unadulterated ratings for each interface design characteristic, and served as the basis against which subject ratings from the treatment groups would be compared. The rating tasks and experimental process for subjects in the three treatment groups were identical to those of the baseline group, excepting that subjects in the treatment groups were provided with an additional piece of information. To wit, treatment group subjects were provided with experimentally manipulated information about how other people rated the same interface and design characteristic that they themselves were currently considering. Further, the only difference among the three treatment groups was the degree of similarity between these "other people" and the subject

herself. A 34-year-old female subject assigned to Treatment Group 1, for example, might be provided with a gender-specific statement such as "The average response given by other *women* for this question is 2.59 out of 7.00" (emphasis added), while if the subject had been assigned to Treatment Group 3, she might be provided with an age- and gender-specific statement such as "The average response given by other *34-year-old women* for this question is 2.59 out of 7.00". A more complete illustration of the research design is provided in Figure 1.

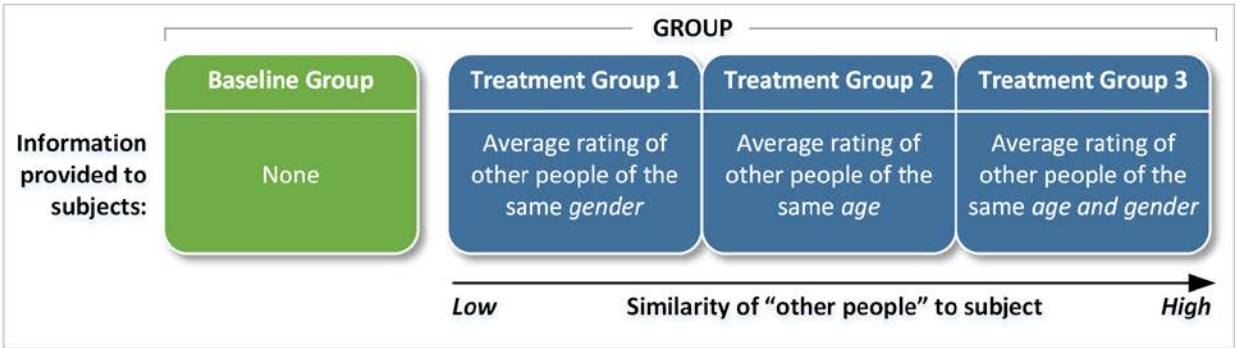

Figure 1. Research design.

Past research has concluded that in the absence of specific knowledge, age and gender are the primary unconscious cues that human beings use when judging how similar they are to others (Brewer and Lui 1989). The figure above thus illustrates how a subject's perceived degree of similarity with the reference group increases as one proceeds from Treatment Group 1 to Treatment Group 3. Intuitively, approximately 50% of a large group of randomly chosen people would share the subject's gender (Treatment Group 1), while a much smaller percentage would share the subject's age (Treatment Group 2). The least likely combination of all, of course, would be to randomly choose people who share both the subject's age *and* her gender (Treatment Group 3), and in the absence of other information, it is with these people that the subject can be expected to most closely identify.

With three web interfaces and five interface characteristics, a total of 15 different configurations were possible for each group. For purposes of statistical validity, a minimum of 30 responses were required for each possible configuration (i.e., 450 responses per group). Since each subject provided three responses, the preliminary minimum sample size was determined to be 150 subjects per group. Given that the linear models (discussed below) that would be used to evaluate the study's research questions contained 11 predictors, a formal a-priori sample size analysis revealed that a minimum of 122 subjects would be required to detect a medium-sized effect (f^2) of 0.15 at a statistical power level of 0.80 (Cohen 1988; Soper 2014a). The preliminary sample size of 150 subjects per group was thus retained for the experiment. The final dataset was distributed by group according to the values in Table 2.

Group	Subjects	Responses
Baseline group	150	450
Treatment group 1 (gender)	176	528
Treatment group 2 (age)	176	528
Treatment group 3 (age and gender)	176	528
Total:	**678**	**2,034**

Table 2. Distribution of subjects and responses by research group.

As noted previously, subjects in the treatment groups were supplied with the average rating of other people for the interface design characteristic that they were currently considering. These ratings were not the true ratings given by others, however, but instead were generated with a view toward gaining insight into the study's research questions. Specifically, the artificial ratings supplied to subjects in the treatment groups were statistically derived from the distributions of baseline ratings To be more precise, the baseline mean rating and standard deviation for each combination of interface and design characteristic were used to

compute the artificial score that was supplied to subjects in the treatment groups, with that artificial score being the value associated with a cumulative probability of 0.05 on the associated baseline rating's normal distribution. For example, the true rating obtained from baseline subjects for the extent to which the third interface used fonts properly was 5.70 (on a 1 to 7 scale), with a standard deviation of 1.32. Applying the cumulative distribution function, one could readily determine that 95% of subjects would naturally rate this interface characteristic at 3.53 or above, while only 5% of subjects would supply a rating lower than 3.53. In this case, treatment group subjects would be told that the artificially low score of 3.53 was the average rating given by other people when evaluating font usage on that interface. Using this approach, it would be very unlikely for a subject in the treatment groups to naturally assign such a low rating to the interface design characteristic that she was evaluating. Any statistically significant differences in the ratings given by the baseline and treatment groups could thus be attributed to herding.

Insight into the study's research questions was gained by estimating three linear models, each of which evaluated the extent to which subject ratings in the baseline group differed from one of the three treatment groups. Each linear model was specified such that subject ratings were predicted by whether a subject belonged to the baseline group or to the model's associated treatment group, after controlling for the subject's age and gender, and the interface and design characteristic being evaluated. For this purpose, membership in the treatment group, subject gender, and the various interfaces and design characteristics were all appropriately coded using a series of binary dummy variables. The results of the linear regression analyses are presented and discussed in the following section.

Preliminary Results and Discussion

Initial estimation of the three linear regression models revealed that subject gender did not significantly affect interface design ratings. Gender was thus removed as a predictor, and the three linear models were then duly reestimated. After controlling for the effects of a subject's age, the effects of the different web interfaces, and the effects of the different interface design characteristics being evaluated, the artificially manipulated information regarding the opinions of others was found to exert a highly significant impact on subject ratings in all three treatment groups ($p < 0.001$ in all cases). These results are summarized in Figure 2. Since interface characteristics were rated on a 1 to 7 scale, the average difference between the baseline ratings and those given by subjects in Treatment Groups 1, 2, and 3 can be quantified as approximately 6.28%, 8.25%, and 10.28%, respectively.

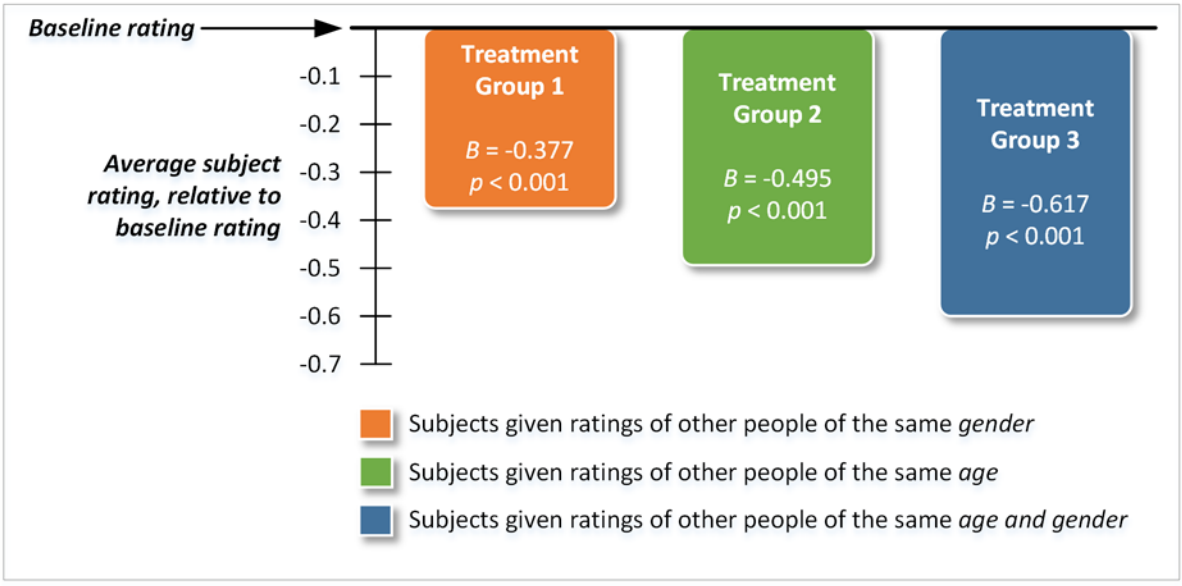

Figure 2. Effects of the opinions of others on subject assessments of web interface design.

The study's first research question inquired into whether a-priori knowledge of the opinions of others would influence interface design assessments. As shown in the figure above, when subjects were provided with artificially low ratings which they were told reflected the aggregate opinions of other people, they would, on

average, assign a statistically lower rating to the interface design characteristic than baseline subjects. Put another way, by simply providing a subject with (false) information about the group opinion, the subject will provide interface design ratings that are much closer to the (false) ratings ostensibly assigned by the group. Further, given that subjects were perfectly free to choose whatever rating they liked for the various interface design characteristics, the tendency of subjects to align their ratings with those of the group indicates that they were not consciously aware of the fact that their decision-making had been artificially manipulated. From these observations it is possible to conclude not only that people's opinions regarding the design of a web interface can be easily manipulated, but also that herd behavior can indeed contaminate assessments of web interface design.

The study's second research question inquired into whether a subject would be more likely to rely on a group's opinion when rating a web interface as the degree of similarity between herself and the members of the group increases. Recalling that the similarity between a subject and the reference group increases as one moves from Treatment Group 1 to Treatment Group 3, it is observationally evident from the figure above that subjects increasingly align their interface ratings with the artificially low ratings of the group when told that the group members are more and more like themselves. A one-tailed statistical comparison of the parameter estimates for Treatment Group 1 and Treatment Group 3 confirmed this observational evidence ($t_{1952} = 1.656$, $p < 0.05$), indicating that strong perceptions of similarity between a subject and the members of a reference group lead the subject to unconsciously shift her ratings in the direction of the group opinion. It is thus possible to conclude that even when there are no social consequences for disagreeing with the group, people will nevertheless unconsciously seek to align their interface design assessments with those of the group, particularly when they believe the members of the group to be very similar to themselves.

Concluding Remarks and Future Research

Despite being a work in progress, the findings reported above have obvious and important implications. A website now commonly serves as an organization's public face, and website design has critical consequences for how the organization is perceived, thus influencing its prospects for success. Eliminating bias from web interface design assessments should therefore be of particular interest to managers seeking to align their organization's website with the needs and expectations of their users.

There is clearly much more to be learned about the role of herd behavior in assessments of interface design, and several additional facets of this phenomenon will be explored as this project matures. The preliminary results reported here, for example, consider only the effects of a-priori knowledge of the group's opinion on interface design assessments. Are subjects willing to revise their ratings if provided with the group's opinion on an ex post facto basis? Further, the current study only attempted to discern if herding would cause subjects to unconsciously lower their web interface ratings. Can herding also cause ratings to increase? Although questions such as these remain to be answered, herding clearly exerts a powerful influence on assessments of user interface design, and it seems likely that herding also plays an important role in many other phenomena which lie at the intersection of technology and human behavior. It is hoped that the preliminary work reported here will serve as a point of embarkation for a long and fascinating stream of research in this area.

References

Aladwania, A.M., and Palvia, P.C. 2002. "Developing and Validating an Instrument for Measuring User-Perceived Web Quality," *Information & Management* (39:6), pp. 467-476.

Avery, C., and Zemsky, P. 1998. "Multidimensional Uncertainty and Herd Behavior in Financial Markets," *American economic review* (88:4), pp. 724-748.

Banerjee, A.V. 1992. "A Simple Model of Herd Behavior," *The Quarterly Journal of Economics* (107:3), pp. 797-817.

Brewer, M.B., and Lui, L.N. 1989. "The Primacy of Age and Sex in the Structure of Person Categories," *Social Cognition* (7:3), pp. 262-274.

Chandon, P., Morwitz, V.G., and Reinartz, W.J. 2005. "Do Intentions Really Predict Behavior? Self-Generated Validity Effects in Survey Research," *Journal of Marketing* (69:2), pp. 1-14.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Coleman, S. 2004. "The Effect of Social Conformity on Collective Voting Behavior," *Political analysis* (12:1), pp. 76-96.

Li, X. 2004. "Informational Cascades in It Adoption," *Communications of the ACM* (47:4), pp. 93-97.

Palley, T.I. 1995. "Safety in Numbers: A Model of Managerial Herd Behavior," *Journal of Economic Behavior & Organization* (28:3), pp. 443-450.

Pew Research Center. 2014. "Internet User Demographics," in *Pew Research Internet Project Survey*. Washington, DC: Pew Research.

Raafat, R.M., Chater, N., and Frith, C. 2009. "Herding in Humans," *Trends in cognitive sciences* (13:10), pp. 420-428.

Saris, W.E., and Gallhofer, I.N. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: John Wiley & Sons.

Soper, D.S. 2014a. "A-Priori Sample Size Calculator for Multiple Regression [Software]." Retrieved 01 Feb 2014, from http://www.danielsoper.com/statcalc

Soper, D.S. 2014b. "User Interface Design and the Halo Effect: Some Preliminary Evidence," in: *Proceedings of the 20th Americas Conference on Information Systems (AMCIS)*. Savannah, GA.

Soper, D.S., and Mitra, S. 2013. "An Inquiry into Mental Models of Web Interface Design," in: *Proceedings of the 19th Americas Conference on Information Systems (AMCIS)*. Chicago, IL.

Watkins, S.C. 1990. "From Local to National Communities: The Transformation of Demographic Regimes in Western Europe, 1870-1960," *Population and Development Review* (16:2), pp. 241-272.