Customization Bias in Decision Support Systems

Jacob Solomon
Michigan State University
solomo93@msu.edu

ABSTRACT
Many Decision Support Systems (DSS) afford customization of inputs or algorithms before generating recommendations to a decision maker. This paper describes an experiment in which users make decisions assisted by recommendations of a DSS in a fantasy baseball game. This experiment shows that the act of customizing a DSS can lead to biased decision making. I show that users who believe they have customized a DSS's recommendation algorithm are more likely to follow the recommendations regardless of their accuracy. I also show that this customization bias is the result of using a DSS to seek confirmatory information in a recommendation.

Author Keywords
Decision Support Systems; Fantasy Sports

ACM Classification Keywords
H.1.2 User/Machine Systems: Human Factors

General Terms
Human Factors

Designers of computerized Decision Support Systems (DSS) face a difficult problem. As socio-technical systems, DSS use artificial intelligence (AI), statistical models, and related technologies to help people make decisions such as diagnosing a medical condition or choosing stocks. Frequently, DSS do this by providing specific recommendations about a decision. However, since the human users of a DSS are prone to decision-making errors or biases, it is possible that at a certain point, efforts to perfect the quality of recommendations made by a DSS will be met with diminishing returns in terms of the quality of decisions made by its users.

How do these recommendations affect the decisions of DSS users? When are recommendations likely to improve decision making, and can recommendations potentially lead DSS users to make worse decisions than they otherwise would have?

One strategy for improving DSS is to allow users to customize various aspects of the system and have influence on the recommendations that are generated. For example, in the clinical DSS *DxPlain* [1], which helps doctors perform diagnosis, the user can select information from a patient's medical profile, including current symptoms and general information such as age, and generate a list of possible diagnoses. The user can also select some information to receive special focus, which affects how the system generates the list. This strategy of customization allows users to integrate their own knowledge or expertise into the recommendations.

Customization is a design feature found not only in DSS but in many types of human-machine systems. In this paper, I make a theoretical contribution to understanding the effect that customization has on user decision making in an HCI context. I also offer a practical contribution to DSS design and policy. I show that customization can create a decision-making bias for users. This bias can lead users to erroneously follow bad recommendations generated by an algorithm, although it can also make users better recognize good recommendations.

BACKGROUND

Many DSS are a class of recommender system that use artificial intelligence or statistical models to process information relevant to a decision. Herlocker et al. [11] have argued that recommender systems suffer from some usability issues because user criteria for evaluating the quality and utility of recommendations are often different from technical criteria for evaluation. Many recommender systems offer recommendations about which decision to make among *horizontally differentiated* alternatives such as which movie to watch or which product to buy. DSS typically make recommendations about *vertically differentiated* decisions in which the result of the decision can be somewhat objectively evaluated (e.g. did the patient's condition improve or did the investment make money?). Thus, providing good recommendations is only half the battle for DSS designers. DSS need to not only provide good recommendations but also be sure that users know how to recognize and follow good recommendations, as well as help them recognize and ignore poor recommendations.

Some research has addressed this by looking at the role of transparency in presenting recommendations to users. Ehrlich et al. [9] for example studied the impact of providing explanations about how recommendations are generated to DSS users, finding that these explanations can have both positive and negative effects on decision making.

A related usability issue is automation bias or complacency. Users may make poor decisions in response to automated decision aids or recommendations because their trust and re-

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.

http://dx.doi.org/10.1145/2556288.2557211

liance on the decision aid is not well calibrated with its actual reliability or users are not vigilant in evaluating the quality of automated recommendations or alerts. This can result in DSS users erroneously following a poor recommendation, or failing to take some necessary action because it has not been recommended by the DSS. Automation bias has been observed in studies about DSS in aviation, [20], luggage screening [16], and mammography [3] among other settings. Despite widespread adoption of DSS in clinical settings, the effectiveness of DSS at improving patient outcomes and clinical decision making is still not yet clearly demonstrated [5, 4]. Coiera et al. [5] point to usability issues such as automation bias as a likely reason that DSS have had only minimal impact in medicine.

One important design issue which has received less attention regarding DSS is the role of the user in generating high-quality recommendations tailored to the specific decision. Interactive media have been shown to provide users with a sense of agency or control over the media or technology [6]. Many technologies are configurable or customizable so that users can tune them to provide the experience they desire. Sundar [21] argues that customizable media give users a sense of "self-as-source" in which they feel they part of the media they consume. This feeling of "self-as-source" leads users to prefer interactive and customizable technologies [22]. This preference can be extended to recommender systems as well. Hijikata et al. [12] found that users preferred music recommendations in a prototype recommender system when they had customized the system prior to receiving recommendations.

The existing research suggests that DSS users would prefer systems that allow for customization. However, it is not clear that this would lead to better decision making in the types of situations for which DSS are commonly used with vertically differentiated options to choose from. In fact several theories in psychology would seemingly make predictions that customization could lead users to biased decision making.

The illusion of control [14] is a phenomenon whereby people overestimate their probability of success in chance situations when there is some kind of choice or perceived element of skill, such as the ability to choose any card from a deck instead of being forced to take the top card. One study [8] examined the illusion of control in a spreadsheet-based DSS for financial decisions. DSS users who engaged in a 'what-if' strategy for using the spreadsheet, where they made adjustments to inputs and formulas in the system when using the DSS to predict outcomes, reported much more confidence in the quality of their decisions in a simulated investment task than those who used a more static and unconfigurable version. This confidence was not warranted, as performance on the task was equivalent for users of both systems.

A second theory that would predict bias induced by customization is the Forer effect [10]. Forer showed that the act of providing personally relevant information to an "expert", such as a psychologist or even a psychic, makes people more inclined to believe that expert. Customizing DSS frequently requires the user providing some information to be processed. This action could therefore conceivably cause users to overestimate the accuracy of the recommendations by eliciting the Forer effect.

A third theory that would make a similar prediction is the IKEA effect [19]. Norton et al. observed in experiments that when people put forth effort in creating a product, such as assembling a piece of furniture, those people were willing to pay more for that product than for an identical item that someone else had assembled. As customization of a DSS requires effort, it may be that this effort creates an investment in the recommendations that leads people to follow them with less criticism of quality, possibly as an example of the sunk cost fallacy.

Each of these three theories makes the general prediction that when customizing a DSS, users will be more likely to follow DSS recommendations than is warranted by the quality of recommendations. This is a critical issue that DSS designers must consider when choosing whether to implement customizable DSS.

Hypotheses

Below I describe an experiment that tests the relationship between user customization of a DSS and compliance with the recommendations provided by a DSS. In this experiment I examine whether DSS users will be more likely to follow recommendations when they believe they have customized the algorithm that produces the recommendations. I predict that based on theory, users will be more likely to follow recommendations or follow them more closely when they believe they have customized the DSS, and that they will especially rely on the DSS recommendation when the decision at hand is difficult. I also predict that customizing a DSS will harm decision-making quality by its users, even though they will be more confident in the quality of their decisions as a result of having input on the recommendations. These predictions collectively argue that customization biases decision making by DSS users. In this study, I also test some theoretical explanations for this *customization bias* such as the illusion of control, the Forer effect, and the IKEA effect.

METHODS

To explore the role of customization in decision making with DSS, I created an experiment where DSS users were given recommendations purportedly generated by a complex algorithm. Some users had the chance to customize the DSS to influence its recommendations, but in reality the customization had no effect on the recommendations. This design tests whether the act of customizing a DSS influences decisions even without affecting recommendations.

Subjects used this psuedo-DSS to inform decisions in a fantasy baseball prediction game in which they tried to predict the winners and scores of Major League Baseball games. This task has several important characteristics that make it useful for studying DSS-aided decision making. First, it is a task with a low threshold for expertise, since many people in the general population follow baseball and play similar games and can therefore be recruited for participation. Another advantage is that it is a decision that involves both a

discrete component (choosing which team will win) and a continuous component (deciding how many runs each team will score). Much existing work on DSS and automation bias has focused only on discrete decisions, even though DSS are commonly used for both discrete and continuous decisions.. Also, most existing work on automation bias has involved tasks that are difficult only because of time or multitasking pressures and not because of a lack of available pertinent information. Many decisions in medicine or finance are analytic in nature, meaning they are difficult because there is a lack of pertinent information. The fantasy baseball task offers a similar analytic decision situation.

Subjects were recruited from Amazon Mechanical Turk to play this game as part of a study to "help improve an algorithmic tool for aiding decisions in fantasy baseball." In order to complete the experiment, subjects had to first take a timed test on the basic rules and statistics of baseball. Only users who demonstrated basic knowledge of baseball rules and statistics were eligible to complete the experiment, and less than half of the Turkers who took this test were successful. This basic knowledge was equivalent to the minimum knowledge required to play fantasy baseball. Subjects were paid $3 for participation. Subjects were also promised an additional payment that would depend on their performance in the game, and were told that the average expected payment would be $2. Subjects took an average of 15.7 minutes to complete the experiment.

The final data set included 99 subjects who played a total of 1,188 rounds of the game. The subject pool was 76% male with an average age of 30 years old.

Game Play

Subjects played 12 rounds of the fantasy baseball prediction game. In this game, all subjects were shown extensive statistics about two teams and asked to make a prediction about the score of the game between the two teams. To ensure that only the available statistical information was used to inform decisions, the names of the teams were not revealed to subjects. Additionally, the games subjects were predicting were games that had already been played. Subjects were told that even though the games were past games, all statistics and algorithms in the study treated the games as if they were in the future.

I selected games for the experiment from the 2011 and 2012 Major League Baseball seasons using several criteria. I fit an existing statistical model [24] for assessing the probability of a home victory to games from these seasons. This model estimates the probability that a home team will win using the relative strength of each team in three categories: winning percentage, the Earned Run Average of the starting pitcher, and Batting Average. The model also includes an adjustment for home field advantage. This model proved useful for this purpose because it estimates the approximate difficulty of predicting a given game using only a small number of statistical categories. Since users are not able or likely to consider a large amount of data without the aid of a sophisticated tool, this model estimates probabilities in a fashion similar to how we might expect users to form predictions.

This model estimates the equivocality of the teams in a given game, which also represents a measure of difficulty of the decision. I chose games at four levels of difficulty. Level 1 difficulty gave greater than 80% chance of winning to one team. Level 4 difficulty gave less than 60%, and levels 2 and 3 were divided at 70%. The twelve games included four games each from levels 2 and 3 and two games each from levels 1 and 4. This distribution approximates the distribution of difficulty across the larger sample of baseball games.

Subjects earned points in the game by making accurate predictions about the outcome of the game. Subjects start each round with 20 points. If they choose the wrong winner, they lose 10 points. They also lose one point for the absolute difference between the predicted number of runs for each team and the actual number of runs. For example, if the final score of a game was Away 5 — Home 3 and the subject predicted Away 4 — Home 6, the subject would lose 10 points for choosing the wrong winner, lose 1 point for missing the Away run total by 1, and 3 points for missing the Home run total by 3, leaving a total of 6 points for the game. This scoring procedure offers an incentive for users to make good decisions not only in choosing the correct team to win the game (a discrete decision), but also in finding precision in predicting scores (a continuous decision). The incentive to perform well in both aspects of the decision is similar to many other DSS supported decisions. For example, a doctor must determine which medication to prescribe among discrete options, but may also need to determine dosage, frequency, or duration of treatment in more continuous decisions.

This scoring procedure serves as a measurement of bias and decision-making quality. Bias is determined by whether subjects agree more with DSS recommendations in their decisions when they have customized the DSS. Agreement is measured in two ways. A discrete form of agreement is measured as whether or not the subject chose the same team to win as the DSS, providing a binary measurement of agreement (*winner agreement*). A continuous form of agreement is measured as the absolute difference in run total between the subject's predicted score and the DSS recommended score (*score agreement*).

Subjects were also given a chance to make a wager on the quality of their prediction which measures confidence in their decisions. Subjects were given an additional 10 "confidence points" in each round. With these points, they could wager any number of them that they scored at least 15 points from their prediction, with a return of 3 to 1. Or they could keep some or all of them and add directly to their final point total for the round.

DSS and Conditions

All subjects used a DSS that provided extensive statistical information about the teams involved in each of the games. In addition to providing statistical information, the DSS also recommended its own prediction about the score of the game. Subjects were told this prediction was based on a statistical algorithm. However, the recommendations were actually pre-determined for each game. There were two types of recommendations. Good recommendations suggested the

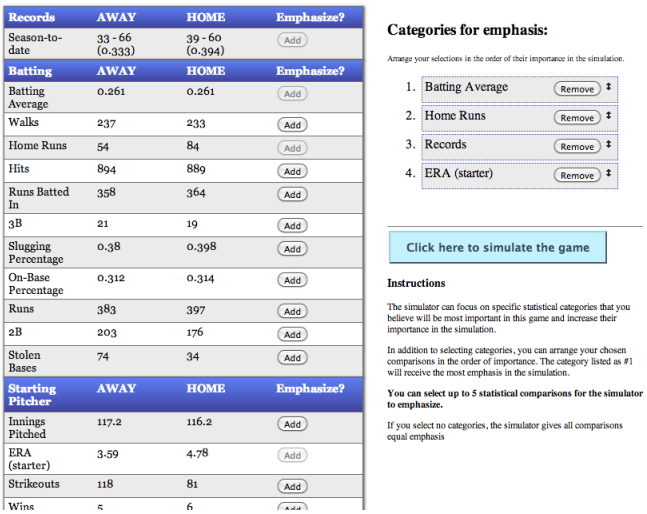

Figure 1. Customizable DSS

actual score of the game, yielding 20 points if followed exactly. Poor recommendations suggested the wrong winner, as well as a score that would yield 5 points. Subjects were given poor recommendations for four games (one randomly selected game from each of the four difficulty levels), and good recommendations for the remaining games. Over the 12 games, the average score of the DSS's recommendations was 15. Subjects were told of this average, but that there would be considerable variation in the quality of the recommendations. This degree of accuracy offers a reasonable amount of data about decisions following both recommendation qualities, and also is a reasonable and believable degree of accuracy for a sports simulation algorithm.

There were two conditions of the experiment, and subjects were randomly assigned to one condition that they remained in over all 12 rounds. In the customizable condition, subjects had the opportunity to make adjustments to the DSS's recommendation algorithm after seeing a table of statistical comparisons between the teams (see Figure 1). The instructions stated that by default, the algorithm treated all statistical comparisons equally (i.e. the relative strength of each team in stolen bases is as influential as the winning percentages). But, they could choose up to five statistical categories to receive extra emphasis and order them according to their importance. For example, a subject could select winning percentage and place it as the most important category, followed by home runs, followed by starting pitcher ERA etc. The instructions stated that good customization improves the performance of the algorithm, but poor customization could harm performance.

In the control condition, subjects only saw the table of statistical comparisons which they could examine before clicking a button to generate a recommendation about the game. The instructions gave these non-customization users the same information regarding the general performance of the DSS.

Subjects were only shown their scores after completing all twelve rounds. This eliminated the possibility for subjects to learn from round to round, which could have confounded and complicated the results.

Survey

After completing 12 rounds of the fantasy baseball game, subjects took a short survey intended to assess theoretical explanations for any observed customization bias. The survey asked about how much control they felt they had over the quality of the recommendations, how much effort they put forth to customize in order to assess the IKEA effect, and also the degree to which they felt they were providing information to the DSS as an assessment of the Forer effect.

RESULTS

Customization Bias

The study design is a 2x2 design with a between-subjects factor (customization) and a within-subjects factor (recommendation quality). To test for the effect of customization on agreement with the DSS, I fit multilevel regression models with the experimental factors as fixed effects and a random effect for each subject to account for the repeated measurements in the design. For assessing the binary measure of winner agreement, I used multilevel logistic regression. These models are described in Table 1. The intercept of these models can be interpreted as the estimated degree of agreement when subjects do not customize the DSS and receive a good recommendation. The coefficients represent differences from this baseline group, and the table shows the standard error below each estimate in parentheses. Figure 2 shows the models' estimated degree of agreement for all four combinations of the factors. This figure converts the log odds estimated by model 1 into the probability of a subject agreeing with the DSS.

Model 1 shows a statistically significant effect of customization on the binary measurement of agreement. Customization users were overall more likely to predict the same team to win as the DSS. Subjects were less discerning of poor recommendations when they had customized the DSS, as they were more likely to agree with the DSS when receiving a poor recommendation than those who did not customize. Conversely, customization users were better at discerning good recommendations as well, being more likely to follow good recommendations than the control group.

Model 4 tests these effects in terms of score agreement. Again, customization users were biased towards agreeing with the recommendation. On average, customization users predicted scores that were 0.9 runs closer to the DSS recommendation than non-customization users. When the recommended scores were accurate, subjects predicted 0.96 runs closer on average than when the recommendation was inaccurate.

Models 2 and 3 in Table 1 add the difficulty of the game to the models to see whether the difficulty of the game would influence how subjects interpreted recommendations and whether this would be different between the two conditions. Since subjects had a higher probability of receiving a poor recommendation for difficulty levels 1 and 4 than for levels 2 and

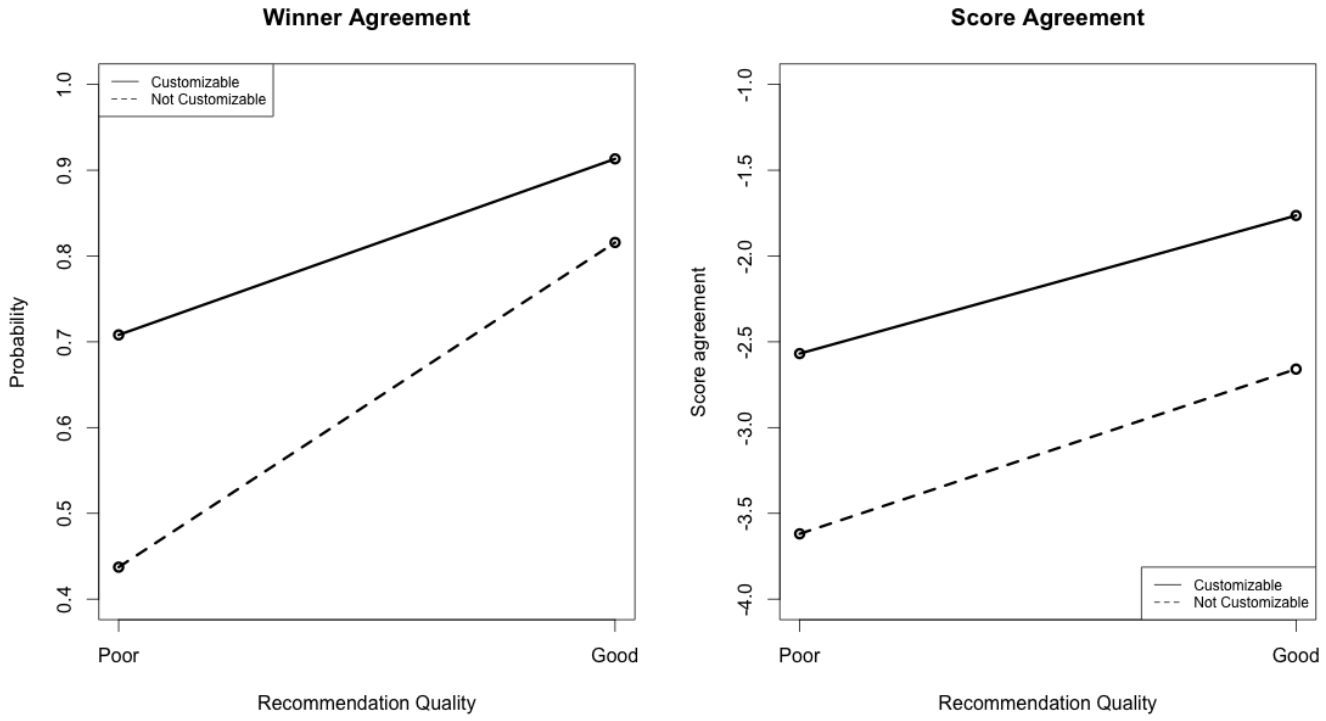

Figure 2. Agreement with DSS recommendations

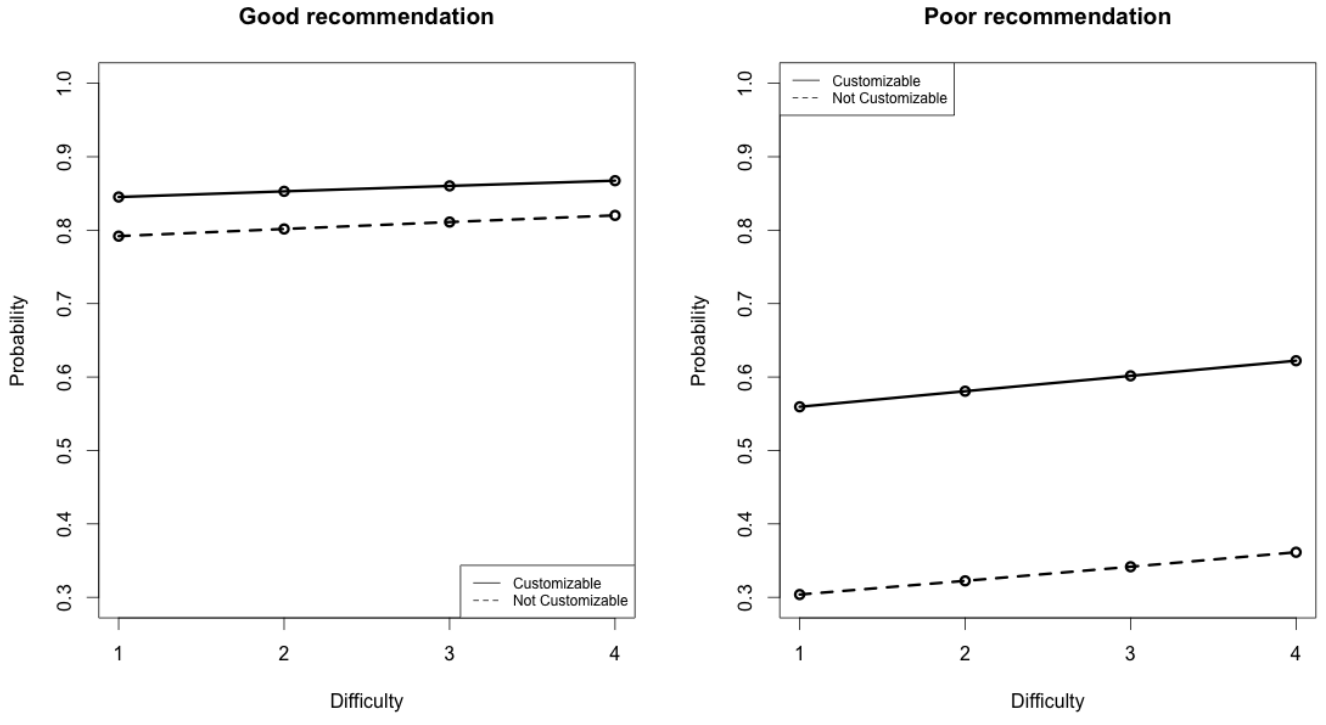

Figure 3. Effect of difficulty on customization bias

Table 1. Customization Bias Models

	Dependent variable:			
	Winner Agreement			Score Agreement
	Binary (Log Odds)			Continuous
		Good Rec	*Poor Rec*	
Model #:	(1)	(2)	(3)	(4)
Intercept	1.49^{***}	1.34^{***}	-0.83^{***}	-2.66^{***}
	(0.19)	(0.28)	(0.31)	(0.16)
Customization	0.87^{***}	0.36	1.07^{***}	0.90^{***}
	(0.27)	(0.40)	(0.41)	(0.22)
Poor Recommendation	-1.74^{***}			-0.96^{***}
	(0.21)			(0.17)
Customization x Poor Rec.	0.27			0.15
	(0.30)			(0.22)
Difficulty		0.06	0.37^{***}	
		(0.15)	(0.14)	
Customization x Difficulty		0.26	0.09	
		(0.24)	(0.20)	
Random Effect Intercept Variance	0.73	0.31	0.83	.78
Observations	1,188	792	396	1,188
Log Likelihood	−579.61	−334.06	−246.99	−2,455.30
Akaike Inf. Crit.	1,169.21	678.13	503.97	4,922.59
Bayesian Inf. Crit.	1,194.61	701.50	523.88	4,953.07
Note:				*p<0.1; **p<0.05; ***p<0.01

	Poor Rec.	Good Rec.
Customization	423.78	717.78
No Customization	470.53	690.10

Table 2. Average rank of decision quality by condition and recommendation type. Rank ranges from 1 (worst) to 1,188 (best).

3, I fit separate models for rounds with good recommendations and poor recommendations using winner agreement as the dependent variable. Figure 3 visualizes these models. When receiving good recommendations, the difficulty of the decision had almost no influence on the probability of agreeing with the DSS in either condition. When receiving a poor recommendation, subjects were slightly more likely to agree with the DSS when the decision was more difficult, and this was true in both conditions. Because the estimates for customization and non-customization are effectively parallel for both types of recommendations, it does not appear that the difficulty of the decision moderates customization bias, although it does seem that when receiving poor recommendations, users of any type of DSS will be more likely to trust it when the decision is difficult than when they have an easier decision.

Confidence and Decision Making

Subjects wagered an average of 4.22 confidence points per round (S.D. 2.5). To see whether customization influences DSS users' confidence in their decisions, I fit a similar multilevel regression model as model 4 above with customization and recommendation quality as independent variables and the number of confidence points wagered by the subject as the dependent variable. This model showed no statistically significant effect of customization on confidence in decisions.

I also tested the effect of customization on subjects' overall decision quality to see whether customization bias led to overall differences in decision making quality. I defined decision quality as the number of points earned from the prediction of a game, including points earned from confidence wagers, because the incentive of the game was to score as many points as possible. The scoring structure created a bimodal distribution because of the large number of points lost when choosing the wrong winner and the 3 to 1 return on confidence points. To correct this, I rank transformed each prediction's points earned compared to all other rounds from the experiment, with a rank of 1 being the lowest number of points. Table 2 shows the average rank in each of the four prediction categories. Subjects made the best overall decisions when they customized a good recommendation, and the worst decisions when they customized a poor recommendation, and all terms including the interaction from the model were statistically significant ($p < .01$). I supplemented this analysis by simply comparing the point totals from all twelve rounds between subjects, measuring the total performance of subjects in the customization condition agains the control group. The mean number of points earned per subject was 360.2 (S.D. 46.4). An OLS regression indicated that customization subjects earned 19.3 more points than those who didn't customize over the whole experiment.

The small difference in decision making, which actually favors customization, is likely an interaction between the ex-

perimental design and the nature of customization bias. Although subjects made worse decisions when receiving a poor recommendation, subjects also made slightly better decisions when receiving a good recommendation. Since subjects were twice as likely to receive a good recommendation in the experiment, their overall performance was slightly better as a result of believing that they customized the DSS.

Explanations for Customization Bias

Table 3. Behavioral measures of the theoretical explanations of customization bias

	Dependent variable:	
	Winner Agreement	
	(5)	(6)
Intercept	2.093***	2.585***
	(0.458)	(0.303)
Time (seconds)		0.002
		(0.005)
Time:Poor Recommendation		−0.004
		(0.007)
Poor Recommendation	−0.188	−1.469***
	(0.597)	(0.325)
# Categories Selected	0.160	
	(0.107)	
Poor Rec:# Cat. Selected	−0.382***	
	(0.148)	
Observations	660	660
Log Likelihood	−268.866	−272.003
Akaike Inf. Crit.	547.732	554.006
Bayesian Inf. Crit.	570.194	576.468
Note:	*p<0.1; **p<0.05; ***p<0.01	

The survey asked questions intended to provide evidence for the theoretical explanations of customization bias. Subjects in the customization condition reported feeling more in control of the recommendations than those in the no customization condition, and the difference was statistically significant ($p < .001$). However, the mean response was 4.61 for the control condition, which is just above the neutral point. Overall, subjects did not report a strong feeling of control over the recommendations. I fit a multilevel model estimating recommendation agreement with responses from the survey and controlling for experiment condition. None of explanations showed meaningful effects. Therefore, the survey offers no evidence that the illusion of control, Forer effect, or IKEA effect can explain customization bias.

The survey measures included single item self-report scales that were not validated previously. Therefore, I conducted an additional analysis of behavioral measures to further explore these theoretical explanations. There are two behavioral measures that can be mapped to the constructs from the Forer effect or the IKEA effect. When customizing the system, subjects could choose how many categories to select. Selecting categories requires effort, and each category selected also offers more information to the algorithm. I fit a multilevel logistic regression (model 5 in Table 3) using only the customization group of subjects. This model predicts winner agreement with the number of chosen categories and the quality of the recommendation as fixed effects, with a random effect of the individual subject.

This model indicates an interaction between the the quality of the recommendation and the number of categories selected. When receiving a good recommendation, there is a small positive and not statistically significant effect of choosing more categories on winner agreement. When receiving a poor recommendation, selecting more categories has a negative effect on winner agreement. Selecting more categories is an imperfect measure because it confounds effort with providing information. Adding additional categories requires effort, and it also is an act of providing information. So this model does not distinguish between the IKEA effect and Forer effect. However, for either of them to explain customization bias, there would need to be a positive effect of selecting categories, particularly for the bad recommendations. Since model 5 shows a negative effect of selecting categories when the recommendation is poor, we can conclude that there is no support for either the Forer effect or the IKEA effect in the data.

To further verify that effort had no effect, I fit a second model using the amount of time spent customizing the system as a predictor in a similar model (model 6 in Table 3). This model showed a similar pattern of effects, although no effects were statistically significant. Overall, the data from this study provide no evidence that either effort or information providing can explain customization bias.

Confirmation Bias

Because the original proposed theories of customization bias did not explain the results, I searched for post-hoc explanations. In the analysis of the difficulty of games, I noticed a slight tendency to make better decisions when the teams had clear strengths in some statistical category, even if their opponent had some other clear strength that led to equivocality and high difficulty for the game. This made me wonder whether the categories chosen by subjects when customizing the DSS would influence their interpretation of the recommendations, and subsequently their decisions. For each prediction made by a subject in the customization group, I calculated whether the recommendation provided for that game was consistent with the statistical categories chosen for customization.

For example, if a subject chose the teams' winning percentages as a statistical category for emphasis, and the system recommended the team with the better winning percentage to win the game, the recommendation was counted as consistent with the customization in that category. For each subject, the customization-recommendation consistency was measured as the percentage of selected categories that were consistent with the recommendation.

I fit a multilevel logistic regression with the quality of the recommendation and the percentage of consistency between recommendation and customization as estimators of winner

agreement. Table 4 describes this model. This model shows meaningful effects for the percentage of consistency between recommendation and customization, as well as for the quality of the recommendation. Figure 4 plots the predicted probabilities of winner agreement from this model for both types of recommendations at all levels of consistency.

It is clear that regardless of recommendation quality, subjects were more likely to agree with the DSS when its recommendation was consistent with their customization. Additionally, this effect was larger when subjects received a poor recommendation. Subjects were better able to discern poor recommendations, and then disagree with the DSS when the recommendation was completely inconsistent with their customization. However, when the recommendation was completely consistent with their customization, subjects nearly always agreed with the DSS. This finding suggests a confirmation bias by subjects because they were more likely to agree with recommendations that confirmed their customization choices.

Table 4. Confirmation Bias model

	Dependent variable:
	Log Odds of Winner Agreement
Intercept	1.23^{***}
	(0.388)
Agreement	2.73^{***}
	(0.62)
Poor Recommendation	-0.98^{**}
	(0.42)
Agreement x Poor Rec.	-0.58
	(0.87)
Observations	593
Log Likelihood	-234
Akaike Inf. Crit.	478
Bayesian Inf. Crit.	500
Note:	*p<0.1; **p<0.05; ***p<0.01

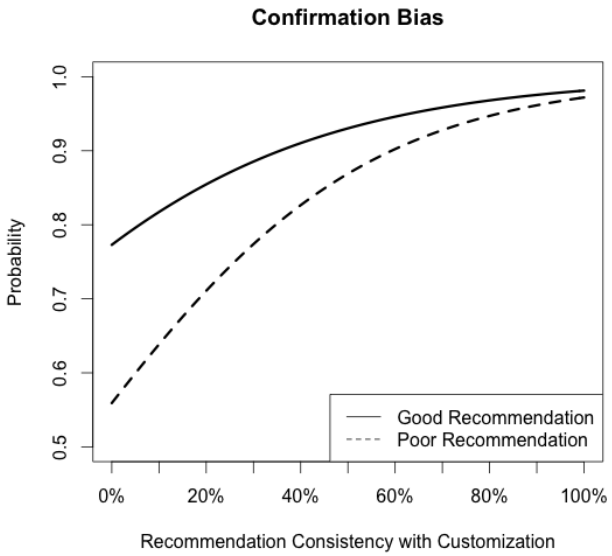

Figure 4. Customization/Recommendation Consistency on Probability of Winner Agreement

DISCUSSION

To summarize the findings:

- Subjects who customized the DSS agreed more with the recommendations of a DSS, both in a binary "yes/no" type decision and in a more continuous "how much" decision, even though their customization had no actual impact on the recommendations.
- This bias led to much worse decisions when the DSS gave a poor quality recommendation, and slightly better decisions when the DSS gave a good recommendation.
- Customizing the DSS did not lead subjects to be more confident in their decisions.
- The illusion of control, Forer effect and IKEA effect do not explain this customization bias.
- Confirmation bias is the best explanation for the bias from these data. Subjects were highly likely to agree with the DSS, regardless of recommendation quality, when the DSS (by random chance) gave them a recommendation that appeared to agree with their customization of the DSS algorithm.

As socio-technical systems, DSS cannot be evaluated purely on any technical criteria but also on their overall influence on user behavior within the context of their use. Offering useful information or recommendations to decision makers is only an intermediary goal for these systems. Ideally, DSS actually increase the quality of decisions made by users.

This study has both theoretical and practical implications. The notion that customization and interactivity create a sense of "self-as-source" [21] is supported by these data. The ability to see one's influence in the output, although spurious in reality, led users to make different decisions. Trust in automation is an increasingly important topic as artificial intelligence becomes ubiquitous. This study suggests that trust can be enhanced when the user has at least some participation in the automated process. This study makes a contribution to emerging theories of HCI, and specifically to understanding how the process of customizing a system influences users of semi-autonomous systems.

More practically, this study informs the design of DSS that incorporate both automated and human knowledge and information processing. This study offers mixed support for incorporating user customization in a DSS design. On one hand, customization can lead to very poor decisions if the customization does not help the system generate good recommendations. Even if customization improves a recommendation, there may still be better alternatives that don't receive full consideration because the user is biased towards the recommendation he or she has helped generate. On the other hand, DSS in general probably make relatively good recom-

mendations on average, otherwise they would be of little use to begin with. If the user is able to actually improve the recommendation through customization, decision making overall may be improved because customization may help them recognize good recommendations.

However, if confirmation bias is the primary cause for this customization bias, there are other important considerations as well. There may be at least two different ways that confirmation bias happened in this study. Subjects may have selected categories for inclusion that they thought were important for the game, then noticed that the recommendation agreed with those categories and felt assured that the recommendation was good. This is subtly different from another form of confirmation bias. Subjects may have selected the categories where the team they already expected to win was stronger, and then simply decided to disagree if the recommendation was not consistent with their expectations. This study cannot distinguish between these two forms of confirmation bias, but this is an important topic for future research. At issue here is the reaction to contradictory information. Does the dissonance between a recommendation and an expert's expectations or beliefs cause poor decisions?

Confirmation bias has been well studied in psychology. Nickerson [18] argues, however, that there are a large number of distinct phenomena that are frequently classified as confirmation bias. In general, confirmation bias happens whenever a decision maker selectively seeks out or processes information that is consistent with a pre-existing belief or hypothesis and ignores information that is inconsistent. Studies of DSS users have found mixed evidence for confirmation bias. Some studies [23, 17] found that users did not demonstrate much confirmation bias, even though the studies were designed to elicit this bias. Others [15, 13] have observed confirmation bias among DSS users. These studies frame a DSS and its recommendations as a source of potentially confirmatory or disconfirmatory information to be considered among other information in the decision task. For this reason, Cummings [7] argues that automation bias is a form of confirmation bias. Automation bias occurs when users over-rely on the recommendations of the automated aid, and over-relying on this aid causes users to avoid seeking out disconfirmatory information. This study introduces customization to the concept of automation bias. Previous research [8] of customization in DSS has not used a DSS that incorporated automation or artificial intelligence. And previous research on automation bias has not used customizable DSS. But it is important that these two concepts be considered together because customization allows users to *change* recommendations to match their existing beliefs.

Customization bias has important legal and ethical considerations as well. To what extent are decision makers liable when they use automated decision aids? A recent court case [2] found that the Colorado State Engineer was within his rights to make rules regarding groundwater usage based on the recommendation of a DSS, even though the DSS was shown to have scientific inaccuracies affecting its recommendations. Should such a ruling stand in a context when the decision maker has customized the DSS? For instance, should a doctor be accountable for malpractice if she made a poor decision at the recommendation of a DSS that she has customized? And from an ethical perspective, does the act of customizing a DSS create a conflict of interests for the decision maker? The results of this study certainly argue that decision makers may not always be objective if they have customized their recommendations.

Another issue that arises from these findings is the potential for feedback loops in DSS that seek to adjust and improve over time using machine learning or similar techniques. If any such systems continuously adjust their recommendation algorithms as a result of user input, it is possible that customization bias could prevent these algorithms from reaching their potential if users are simply working to confirm what they already believe rather than inform the system or themselves or seek disconfirming information. Systems that aggregate user input to make use of 'the wisdom of the crowd" may not be incorporating enough disconfirmatory information.

There are some important limitations to this experiment to consider. First, in order to maintain the credibility of the customization manipulation, subjects did not receive feedback between rounds about their scores, but instead were given their scores for all rounds after the experiment. So subjects did not have the chance to learn over repeated use about their decision making. Would subjects learn from poor decisions over time and become less biased towards the recommendations after customizing them? Also, while the subjects in this study had to demonstrate a reasonable amount of knowledge of baseball through a qualification test, there are undoubtedly differences between these subjects and experts such as doctors or experienced Wall Street traders who have spent their lives developing expertise in a limited area. Exploring customization bias directly in such settings is the next step for this research and is critical for developing better decision support systems.

Another limitation to this study is that it does not directly address the role and importance of customization as a method of decision making or as part of the process of gathering and processing information by the user. In order to customize a DSS algorithm, users need to think about how they might impact the recommendations or what information will be most useful. It is possible that the process of customizing could lead users to better understand the data and decision task, leading them to make better decisions. Since users who customized the system made overall better decisions when they received a good recommendation, it is possible that the act of customizing allowed them to better recognize these good recommendations as a result of having contemplated how to customize the algorithm. However, they may also have simply been biased to agree with the recommendation. The design of this study cannot distinguish between those two explanations for the improved decision making when customizing and receiving a good recommendation, although the results do clearly demonstrate a bias in cases of poor recommendations. Addressing these limitations is an important step both

for developing theory around customization bias and for informing design decisions about using customization in automated decision aids.

This study overall demonstrates a theoretical link between the act of customizing the recommendations of a DSS and the subsequent decision made by users. This link has important implications for HCI theory and practice as new forms of automation are developed to assist human decision making.

ACKNOWLEDGEMENTS

I would like to thank Rick Wash, Gary Hsieh, Wei Peng, and Joseph B. Walther for their helpful feedback and guidance of this research. This work was supported with funding from the College of Communication Arts & Sciences and the Graduate School of Michigan State University.

REFERENCES

1. http://lcs.mgh.harvard.edu/projects/dxplain.html.
2. Simpson v. cotton creek circles, llc, 2008.
3. E. Alberdi, A. Povyakalo, L. Strigini, and P. Ayton. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology*, 11(8):909–918, 2004.
4. T. J. Bright, A. Wong, R. Dhurjati, E. Bristow, L. Bastian, R. R. Coeytaux, G. Samsa, V. Hasselblad, J. W. Williams, M. D. Musty, L. Wing, A. S. Kendrick, G. D. Sanders, and D. Lobach. Effect of clinical decision-support systems: A systematic review. *Annals of Internal Medicine*, 2012.
5. E. Coiera, J. Westbrook, and J. Wyatt. The safety and quality of decision support systems. *Methods of Information in Medicine*, 45(1):S20–5, 2006.
6. D. Coyle, J. Moore, P. O. Kristensson, P. Fletcher, and A. Blackwell. I did that!: measuring users' experience of agency in their own actions. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 2025–2034. ACM, 2012.
7. M. Cummings. Automation bias in intelligent time critical decision support systems. In *AIAA 1st Intelligent Systems Technical Conference*, volume 2, pages 557–562, 2004.
8. F. D. Davis and J. E. Kottemann. User perceptions of decision support effectiveness: Two production planning experiments. *Decision Sciences*, 25(1):57–76, Jan. 1994.
9. K. Ehrlich, S. E. Kirk, J. Patterson, J. C. Rasmussen, S. I. Ross, and D. M. Gruen. Taking advice from intelligent systems. In *Proceedings of the 15th international conference on Intelligent user interfaces - IUI '11*, page 125, New York, New York, USA, Feb. 2011. ACM Press.
10. B. R. Forer. The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology*, 44(1):118–123, 1949.
11. J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2004.
12. Y. Hijikata, Y. Kai, and S. Nishida. The relation between user intervention and user satisfaction for information recommendation. In *SAC '12 Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 2002–2007, 2012.
13. H.-H. Huang, J. S.-C. Hsu, and C.-Y. Ku. Understanding the role of computer-mediated counter-argument in countering confirmation bias. *Decision Support Systems*, 53(3):438–447, June 2012.
14. E. J. Langer. The illusion of control. *Journal of Personality and Social Psychology*, 32(2):311–328, 1975.
15. G. Lindgaard, C. Pyper, M. Frize, and R. Walker. Does bayes have it? decision support systems in diagnostic medicine. *International Journal of Industrial Ergonomics*, 39(3):524–532, 2009.
16. P. Madhavan and R. R. Phillips. Effects of computer self-efficacy and system reliability on user interaction with decision support systems. *Computers in Human Behavior*, 26(2):199 – 204, 2010.
17. E. K. Muthard and C. D. Wickens. Factors that mediate flight plan monitoring and errors in plan revision: Planning under automated and high workload conditions. In *Proceedings of the 12th international symposium on aviation psychology*, pages 857–62, 2003.
18. R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.
19. M. Norton, D. Mochon, and D. Ariely. The 'IKEA effect': When labor leads to love. *Harvard Business School Marketing Unit Working Paper*, (11-091), 2011.
20. L. J. Skitka, K. L. Mosier, and M. Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991 – 1006, 1999.
21. S. S. Sundar. Self-as-source. In E. Konijn, S. Utz, M. Tanis, and S. B. Barnes, editors, *Mediated Interpersonal Communication*, pages 58–74. Routledge, New York, 2008.
22. S. S. Sundar, J. Oh, S. Bellur, H. Jia, and H.-S. Kim. Interactivity as self-expression. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, page 395, New York, New York, USA, May 2012. ACM Press.
23. S. Ward. Decision support for what-if analysis and the confirmation bias. *Journal of Computer Information Systems*, 40(4):84–92, 2000.
24. T. Y. Yang and T. Swartz. A two-stage bayesian model for predicting winners in major league baseball. *Journal of Data Science*, 2(1):61–73, 2004.

Heterogeneity in Customization of Recommender Systems By Users with Homogenous Preferences

Jacob Solomon
University of Michigan
Ann Arbor, MI, USA
jacobbs@umich.edu

ABSTRACT

Recommender systems must find items that match the heterogeneous preferences of its users. Customizable recommenders allow users to directly manipulate the system's algorithm in order to help it match those preferences. However, customizing may demand a certain degree of skill and new users particularly may struggle to effectively customize the system. In user studies of two different systems, I show that there is considerable heterogeneity in the way that new users will try to customize a recommender, even within groups of users with similar underlying preferences. Furthermore, I show that this heterogeneity persists beyond the first few interactions with the recommender. System designs should consider this heterogeneity so that new users can both receive good recommendations in their early interactions as well as learn how to effectively customize the system for their preferences.

Author Keywords

Recommender Systems, Customization

ACM Classification Keywords

H.1.2. User/Machine Systems: Human Factors

INTRODUCTION

Recommender systems have the challenge of matching items in their catalog, such as movies or consumer products, to users who have heterogeneous preferences for those items. An item suitable for one user will likely be unsuitable for another. One way recommender systems can deal with this heterogeneity in user preferences is to give users a high degree of control over the recommender's algorithm, allowing them to work collaboratively with the system to find items that match their preferences. This approach to designing interactive systems is known as *customization* [1], and considerable recent HCI research has explored ways to build customizable recommenders [10, 2, 11, 14].

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05?15.00
DOI : *http* : *//dx.doi.org*/10.1145/2858036.2858513

While the goal of customization is to help the recommender accurately find items that match preferences that vary from user to user, it may simultaneously create a new form of heterogeneity among users regarding the way that they choose to configure the system. Users can arrive at a system bringing wide variability in their experience using intelligent systems, their mental models of how they work, their expectations about how the system will be helpful, and knowledge about the decision or items they are seeking recommendations about. Systems that use complex logic but lack transparency about how that logic works [9] may further complicate things for users in trying to figure out how to make the system give them what they want.

Consider two users of a movie recommender for whom the movie *Sleepless in Seattle* would in actuality be a very good recommendation (i.e. users with homogenous preferences). One user may have a mental model of the system that says the system heavily relies on the cast of a movie, and another may have a mental model or expectation that the genre is the major determinant of recommendations. One user would likely then configure the system to focus on movies with Tom Hanks, while the other would try to filter for Romantic Comedies. *Sleepless in Seattle* would likely appear in lists for both configurations, but the set of recommendations could be very different for each user and this could impact whether the user eventually chooses to watch *Sleepless in Seattle* or some other movie that would be less preferred.

This variability in users' characteristics related to interaction with a recommender creates added heterogeneity that must be accounted for by a recommender system. I present data from user studies of two different customizable recommender systems to show that even when different users have similar preferences – meaning the system should in theory provide them with similar recommendations – they are likely to configure the system in widely different ways. The accuracy of customizable recommenders may suffer due to noise in its user profiles that comes from the process of interacting with the recommender, and not just in the overall heterogeneity in users' preferences.

BACKGROUND

In order to provide good recommendations to users, recommender systems must elicit information about users' preferences. A common approach to this is collaborative filtering

in which users give explicit ratings for items (such as star ratings for movies), and then suggest items that are highly rated among other users who have given similar ratings [17]. Content-based recommenders suggest items that have similar content or attributes to items that users have highly rated [3].

Recent work in recommender systems has explored different ways to afford users control over how the system produces its recommendations beyond providing ratings for items, such as controlling the influence of some portions of a social network on the recommendations [7, 19], enabling users to sort or filter items based on attributes [8, 10], allowing users to give weight to specific attributes of items [2, 13], or providing an interface to critique recommendations [5]. A large body of research has demonstrated that giving users control over recommender systems improves the user experience [12, 10, 2, 11, 6, 7, 4], although it may also create some decision-making biases [16].

One issue with customization is that it is generally more preferred among expert users than novices or users with little domain knowledge relating to the decision being made [18, 8]. This creates a difficult paradox for the system in that 1) new users are the ones that the system needs the most input from in order to be effective (i.e. the "cold start" problem [15]) and 2) new users may never become effective at customizing the system to accurately provide recommendations if they do not get practice at using the system. However, if a system does afford customization to new users, those users may be ill-equipped to effectively express their preferences in their early interactions, which could lead the system to create inaccurate profiles of these new users that hurt the quality of recommendations it provides. If users with similar preferences (i.e. users who should be receiving similar recommendations) provide widely varying input to the system, it would add considerable noise that could inhibit its overall effectiveness.

To study this issue of heterogeneity in configuration, I sidestep the nuisance factor of heterogeneity in preferences for items by *assigning* those preferences to users, so that the sample of users in these studies form a homogenous group of users who all have identical preferences for the items being recommended. This feature of the study design allows us to see how much variability in the behavior of new users can be attributed to their inexperience with the system and not to the natural underlying heterogeneity in their preferences. In other words, this design examines how much variation there is among groups of users that would form "neighborhoods" of users if preference information could be perfectly elicited and measured. I show that even within a group of homogenous users, there is considerable variation in how they will customize a recommender, and that this variation does not quickly reduce as these users become more experienced.

TRAVEL AGENT STUDY

I created an interface to a prototype recommender system called Travel Agent for recommending travel destinations that distinguished important attributes of a destination and allowed users to specify their preference along a continuous scale for these attributes by adjusting a slider (see Figure 1).

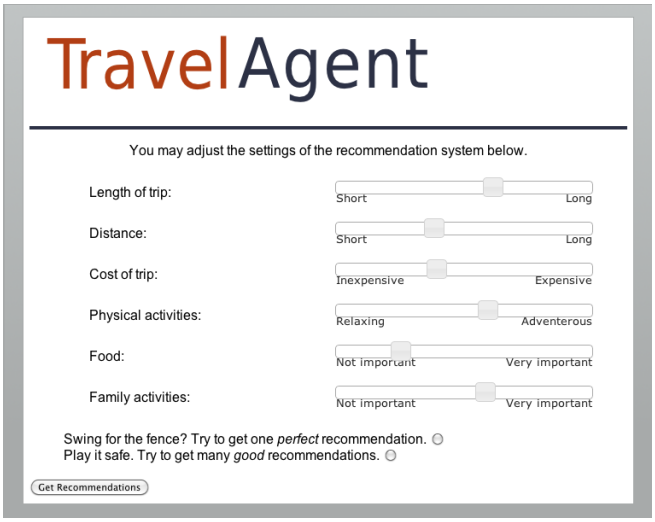

Figure 1. Travel Agent interface.

Rather than seeking a recommendation for themselves, subjects were trying to get recommendations for a fictional user persona. The persona was a 51 year old financial analyst from Chicago with 3 children looking to plan a vacation in October. The persona described details of his personality, hobbies, travel experience, and budget.

I recruited 375 subjects from Amazon Mechanical Turk to participate in this evaluation of Travel Agent. These subjects read instructions for the study and the details of the persona, and were required to pass a quiz on this information prior to using Travel Agent. After passing this quiz, they proceeded to use Travel Agent and generate recommendations, after which they answered survey questions about the recommendations.

Results

Despite the fact that all users were trying to produce recommendations for the same person, there was remarkable variability in the way users configured Travel Agent. Figure 2 shows the distribution of configuration for each option across the different users. A value of 100 corresponds with moving the slider for that option all the way to the right. For all options, the distribution spans the majority of the scale with relatively normal distributions, although the large spikes at the far-left setting are noteworthy. It should be noted that by default the sliders were pre-set to the mid-point of the scale. For each feature, users set the sliders at levels across the entire scale, although there were relatively normal distributions around a mean level for each feature.

To analyze this variance further, I conducted a k-means cluster analysis on the matrix of users by configuration options. Subjects were clustered together based on the similarity of all seven of their configuration choices. I determined that the within-cluster variance continually dropped until 20 clusters were formed, suggesting that there are about 20 distinct patterns of configuration among the 375 subjects. Figure 3 shows how each of these clusters configured the system. These clusters represent unique combinations of settings of the seven options.

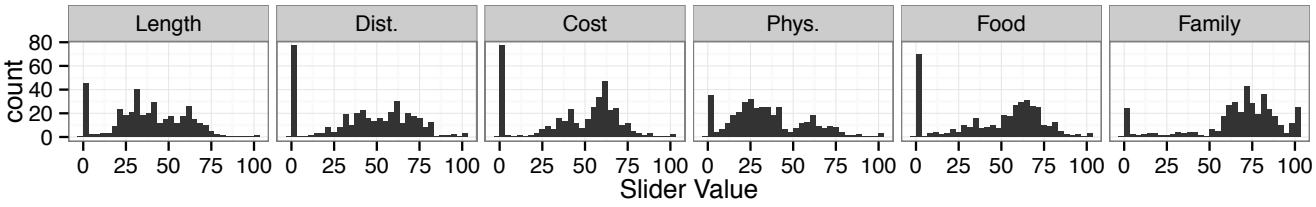

Figure 2. Distribution of configurations for each option in Travel Agent. A value of 100 corresponds to moving the slider all the way to the right.

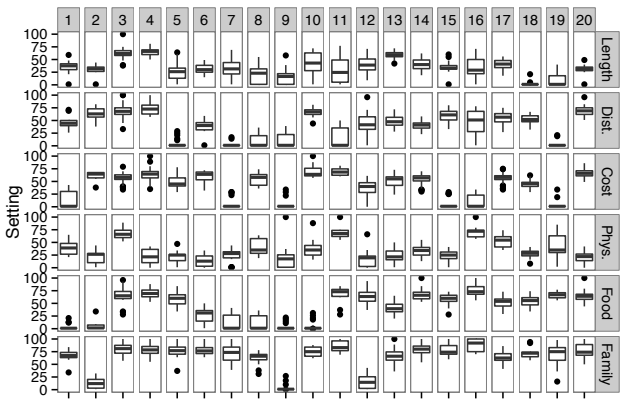

Figure 3. Variation in the way different clusters of subjects configured Travel Agent, even though all subjects were configuring for the same persona. Each column is a single cluster.

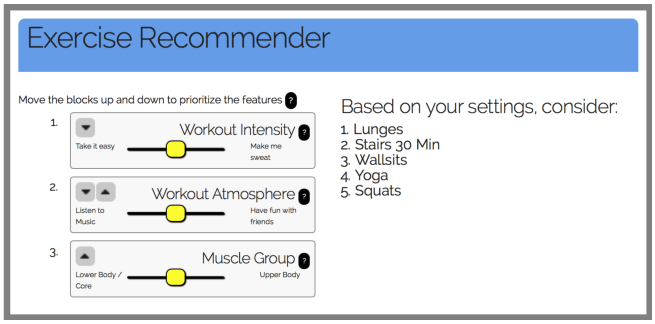

Figure 4. Exercise Recommender interface.

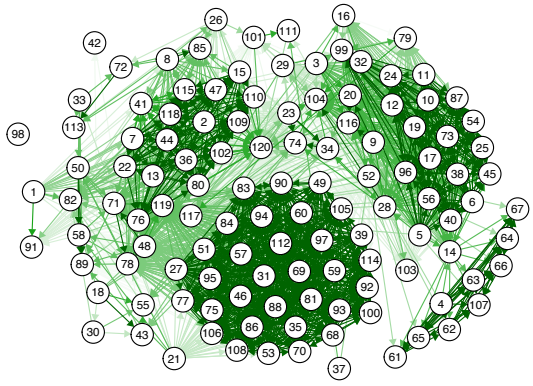

Figure 5. Similarity network of configurations for a single profile. Clusters represent groups who used similar configurations of the system.

EXERCISE RECOMMENDER STUDY

A limitation of the Travel Agent study is that the heterogeneity in the way that users customized the system may simply represent variability in the way people interpreted the persona's preferences. To address that limitation, I conducted a second study using a system called Exercise Recommender for recommending exercise activities (Figure 4) that specified preferences more specifically and gave greater incentive for users to customize the system for those preferences. I drew on experimental economics research to develop a decision task in which users are assigned preferences for attributes of an item by giving them a "payoff" for choosing an item (an exercise activity in this case) that has a given attribute. Subjects were shown five attributes of exercises they prefer (e.g. a cardio exercise, a group activity, a convenient activity etc.) and if their selected activity matched those attributes (as determined by an external panel of judges) they received an additional payment beyond their baseline compensation for participation in the study. This method has been shown to effectively *induce* preferences [20] in experiments by giving them an incentive to make decisions that fit their assigned preferences rather than their own personal preferences. By assigning concise preferences and incentivizing subjects to match those preferences, I was able to again replicate a group of homogenous users within a sample of recruited subjects.

113 subjects recruited from Amazon Mechanical Turk participated in the study. Subjects customized the system in two ways. First, they indicated through a 5-point slider their preference on three dimensions of an exercise (Workout Intensity, Social Recreation, and Muscle Group). Then, they could prioritize these dimensions by moving their input block up and down, such that the system would place greater emphasis on matching the dimension at the top of the list. The Exercise Recommender returned 5 recommended activities using a recommender formula suggested in [14]. After choosing an activity, they were shown their payoff for that choice and then redirected back to the Exercise Recommender to complete the task again using a new preference profile that gave different payoffs for different attributes. Each subject completed this task ten times so that their learning over repeated use of the system could be assessed.

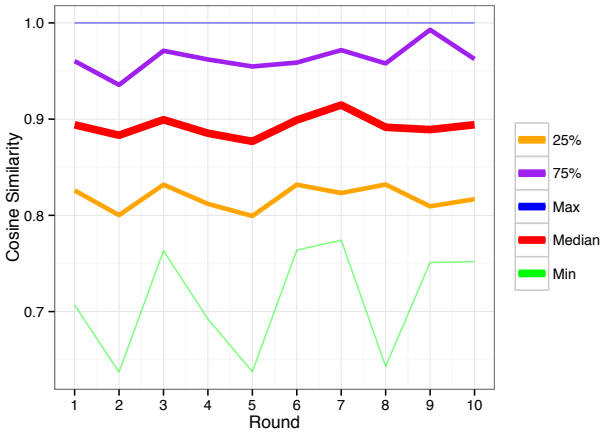

Figure 6. Similarity within cohorts using the same profile in the same round number. There is no trend for any aspect of the distributions, suggesting that subjects did not converge over time to more similar choices in customizing the Exercise Recommender.

Results

Figure 5 illustrates the variety of different ways that people tried to configure the Exercise Recommender for just one of the profiles. I calculated the similarity between each subject's configuration when using the system for a particular profile and every other subject when using that same profile as the cosine similarity between the six configuration values (Muscle Group setting, Muscle Group priority, Social Recreation setting, etc. ...) of each subject.

As in the Travel Agent study, there was again significant variability in the ways that people customized the Exercise Recommender even when trying to get recommendations to match an identical set of preferences. In the profile represnted in Figure 5, there is a single dominant cluster and three smaller clusters, as well as a non-trivial number of unclustered configurations. The dominant cluster for this profile in fact accounts for only 28% of the pool, leaving 72% of customization choices spread out among many different approaches. Across the 10 profiles, there were between two and four clear profiles for each cluster, with the largest cluster never accounting for more than 50% of users. This variability is further evidence that a customizable algorithm presents a difficult usability challenge to users who must figure out how to express their preferences and control the recommender, but may have widely varying mental models or intuitions about how to do that successfully.

I wanted to see whether this variability was stronger in the earlier rounds of the study than in later rounds to see whether users would begin to homogenize in their configurations after gaining some experience with the system and feedback about decisions. I divided the dataset into cohorts of subjects who used the same preference profile in the same round number. Since the preference profiles were randomly ordered for each subject, this resulted in 100 cohorts of 9 to 13 subjects. I calculated the full cosine similarity matrix within each cohort, and extracted the quartiles of each matrix. I found that over time, there was no trend of configurations becoming more or less similar to each other among people with the same preference profile. Figure 6 illustrates the pattern of change over time in the quartiles of cohorts' distributions, and regression analyses suggested that there was no meaningful trend towards either greater or lesser similarity over time.

DISCUSSION

Customizable recommender systems provide an alternative to more traditional ratings-based approaches such as collaborative filtering or content-based recommenders in eliciting preference information from users. Customizable recommenders provide an interface for users to interact more directly with the system's algorithm or recommender logic. While this can have an advantage of providing a positive user experience [4], it may not be an effective way to elicit reliable information about preferences, particularly from novice users of a system. Customizing an algorithm directly may be somewhat of a skill that requires training, experience, and knowledge to perfect. This heterogeneity may also be a significant source of noise for customizable systems, since users who want the same thing won't necessarily do the same thing within the system.

These findings suggest that customizable algorithms actually require less flexibility than what is apparent in user interfaces to those algorithms. Users should be given multiple paths to reach the same destination, meaning that there should be several different ways to configure a system within the UI that effectively result in the same recommendations. In the Exercise Recommender, an ideal design would have given about three different paths that would have led to similar recommendations, since there were typically about 3 distinct clusters of strategies for configuring the system. Additionally, these results suggest that customizable recommender systems need to explore feedback mechanisms to help users perceive what effect they have on a recommender algorithm. This will help users adapt their mental models and their configuration choices to better fit the algorithm.

A limitation of this study is that it merely quantifies the degree of heterogeneity that designers might expect, but does not provide specific information about the different mental models users have of an algorithm. As these mental models may be highly specific to particular systems or decision contexts, a critical part of a good user-centered design process will involve user research to determine all the specific mental models or customization strategies that users will take, and building affordances into the customization process that fit the varying mental models.

ACKNOWLEDGMENTS

Rick Wash, Emilee Rader, Joshua Introne, and Wei Peng, and members of the BITLab at Michigan State University provided valuable feedback on various stages of this work. This work was funded by the College of Communication Arts and Sciences at MSU.

REFERENCES

1. Jan Blom. 2000. Personalization: a taxonomy. In *CHI'00 extended abstracts on Human factors in computing systems*. ACM, 313–314. DOI: http://dx.doi.org/10.1145/633292.633483
2. Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. Tasteweights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 35–42. DOI: http://dx.doi.org/10.1145/2365952.2365964
3. Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12, 4 (2002), 331–370. DOI: http://dx.doi.org/10.1023/A:1021240730564
4. Dina Burkolter, Benjamin Weyers, Annette Kluge, and Wolfram Luther. 2014. Customization of user interfaces to reduce errors and enhance user acceptance. *Applied ergonomics* 45, 2 (2014), 346–353. DOI: http://dx.doi.org/10.1016/j.apergo.2013.04.017
5. Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 125–150. DOI: http://dx.doi.org/10.1007/s11257-011-9108-6
6. Yoshinori Hijikata, Yuki Kai, and Shogo Nishida. 2012. The relation between user intervention and user satisfaction for information recommendation. In *SAC '12 Proceedings of the 27th Annual ACM Symposium on Applied Computing*. 2002–2007. DOI: http://dx.doi.org/10.1145/2231936.2232109
7. Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 43–50. DOI: http://dx.doi.org/10.1145/2365952.2365966
8. Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. 2011. Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 141–148. DOI: http://dx.doi.org/10.1145/2043932.2043960
9. Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1–10. DOI: http://dx.doi.org/10.1145/2207676.2207678
10. Benedikt Loepp, Katja Herrmanny, and Jürgen Ziegler. 2015. Blended Recommending: Integrating Interactive Information Filtering and Algorithmic Recommender Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 975–984. DOI: http://dx.doi.org/10.1145/2702123.2702496
11. Denis Parra. 2013. *User controllability in a hybrid recommender system*. Ph.D. Dissertation. University of Pittsburgh.
12. Denis Parra and Peter Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies* 78 (2015), 43–67. DOI: http://dx.doi.org/10.1016/j.ijhcs.2015.01.007
13. J Ben Schafer, Joseph A Konstan, and John Riedl. 2002. Meta-recommendation systems: user-controlled integration of diverse recommendations. In *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, 43–51. DOI: http://dx.doi.org/10.1145/584792.584803
14. J Ben Schafer, Joseph A Konstan, and John Riedl. 2004. View through MetaLens: usage patterns for a meta-recommendation system. *IEE Proceedings-Software* 151, 6 (2004), 267–279. DOI: http://dx.doi.org/10.1049/ip-sen:20041166
15. Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 253–260. DOI: http://dx.doi.org/10.1145/564376.564421
16. Jacob Solomon. 2014. Customization bias in decision support systems. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 3065–3074. DOI: http://dx.doi.org/10.1145/2556288.2557211
17. Xiaoyuan Su and Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009 (2009), 4. DOI: http://dx.doi.org/10.1155/2009/421425
18. S Shyam Sundar and Sampada S Marathe. 2010. Personalization versus customization: The importance of agency, privacy, and power usage. *Human Communication Research* 36, 3 (2010), 298–322. DOI: http://dx.doi.org/10.1111/j.1468-2958.2010.01377.x
19. Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 351–362. DOI: http://dx.doi.org/10.1145/2449396.2449442
20. Vernon L. Smith. 1976. Experimental Economics: Induced Value Theory. *The American Economic Review* 66, 2 (May 1976), 274–279. http://www.jstor.org/stable/1817233

Don't Wait! How Timing Affects Coordination of Crowdfunding Donations

Jacob Solomon
Michigan State University
solomo93@msu.edu

Wenjuan Ma
Michigan State University
mawenjua@msu.edu

Rick Wash
Michigan State University
wash@msu.edu

ABSTRACT

Crowdfunding sites often impose deadlines for projects to receive their requested funds. This deadline structure creates a difficult decision for potential donors. Donors can donate early to a project to help it reach its goal and to signal to other donors that the project is worthwhile. But donors may also want to wait for a similar signal from others.

We conduct an experimental simulation of a crowdfunding website to explore how potential donors to projects make this decision. We find evidence for both strategies in our experiment; some donate early while others wait till the last second. However, we also find that making an early donation is usually a better strategy for donors because the amount of donations made early in a project's campaign is often the only difference between that project being funded or not. This finding suggests that crowdfunding sites need to develop designs, policies and incentives that encourage people to make immediate donations so that the site can most efficiently fund projects.

Author Keywords

Crowdfunding

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

General Terms

Human Factors; Design

INTRODUCTION

Crowdfunding sites are a type of online community where large groups of people come together in order to realize a new idea or project that requires financing. Crowdfunding is a form of collective action that requires the participation of people in varying user roles, such as project creators, donors, or site administrators. Success in crowdfunding is achieved when the individual actions of users are well coordinated so that everyone's effort is put to use, not duplicated by other users, or not withheld from the group. Crowdfunding sites have many features that enable this coordination, either by affording direct communication between users, by providing information about users and their collective behavior, or by providing rules and structure that assist users in their decision making.

One feature in many sites that assists in coordination is the setting of a deadline for projects to collect the required funds. This deadline typically accompanies an "All-or-nothing" style of crowdfunding where refunds are given to donors if a project does not meet a specified goal before the deadline [20]. Often in this style of crowdfunding, the site is designed to provide real-time status information about the progress of a project towards its goal. Kickstarter, for example, displays for each project the up-to-date total for how much the project has received, how many people have donated, and how long until the deadline (see Figure 1).

This combination of design features presents an interesting choice for anyone who might consider contributing to a crowdfunding project. *When* should one make a donation? Should donors immediately donate, or wait some period of time before donating? There are good reasons for either choice. It may make sense to wait to see whether other people donate, because the donations of others may be sufficient and a donor could free-ride and reap the benefits of the project without contributing. It might also make sense to donate immediately, because a donation may be used as a signal to others that the project has quality and encourage them to donate.

Both choices are relevant to coordination. Signals sent to other users of a site through donation (or lack of) will influence the way groups coordinate to fund projects, and likely the degree to which they are successful at this collective effort.

We explore this decision from the perspective of crowdfunding donors by conducting an experimental simulation of a crowdfunding site. This study seeks to understand how the decision of when to donate affects coordination on crowdfunding sites. We primarily explore how the degree of interest or preference that donors have for projects (both as individuals and collectively as a cohort of potential contributors) affects their decisions, and how this subsequently affects coordination and crowdfunding outcomes. We show that making an immediate donation is generally a better strategy for potential donors, particularly if they are strongly motivated to see a project completed. We do find however, that waiting till the end can be a successful strategy in some circumstances.

CSCW 2015, March 14–18, 2015, Vancouver, BC, Canada.

ACM 978-1-4503-2922-4/15/03 ...$15.00.
http://dx.doi.org/10.1145/2675133.2675296

Figure 1. An example Kickstarter project

Later, we will discuss the implications of this finding for the design of crowdfunding sites.

CROWDFUNDING, COORDINATION AND TIMING

Recent years have seen the rise of a wide variety of crowdfunding websites [8], including Kickstarter (which funds creative projects), IndieGoGo (which funds a wide variety of ideas and new businesses), Spot.Us (which funds investigative journalism), and Donors Choose (which funds K–12 classroom projects). These websites have enabled people with ideas to raise large amounts of money to support their projects. Kickstarter has raised over $1 billion in funding for a variety of projects from over 6 million people[1]. Donors Choose has raised over $240 million for over 450,000 projects in K-12 classrooms[2].

Projects on crowdfunding websites offer two forms of value to donors. Many projects offer a *public good* – a valuable good that everyone can benefit from, even if they don't back the project. For example, recently Kickstarter was used to raise money to create a new Veronica Mars movie that we all can now watch. Additionally, many crowdfunded projects offer specific value to individual donors, usually in the form of project-related rewards, product pre-orders, or equity in what is produced. Crowdfunding is rarely used solely as a sales mechanism; almost all crowdfunded projects include some public goods component. Belleflamme et al.[6] argue that in the absence of any public good aspect, crowdfunding theoretically "yields exactly the same outcome as seeking money from a bank or equity investor."

Crowdfunding is an effective method of raising funds for projects. Many crowdfunding websites have funding rates higher than 40%: 43%–49% of projects on Kickstarter are fully funded [15, 12]; 43.5% of projects on Spotus are fully funded [11], and almost 70% of projects on Donors Choose are fully funded [19]. Mollick [15] observes that projects that ask for less money have higher funding rates than larger projects.

Crowdfunding Needs Coordination

Crowdfunding enables a large number of people to collaborate through the creation of and donations to projects to produce a public good. Crowdfunding works because people donate to crowdfunding projects irrespective of geography [1]. However, this creates a coordination problem for the people involved: with scarce resources, how can donors and creators decide which projects to put their effort and money behind?

Crowdfunding requires several forms of coordination. Much of the CSCW research on crowdfunding has looked at coordination between project creators and donors or potential donors. Gerber and Hui [7] found that establishing long-term connections with backers and building awareness are important motivations for project creators to use crowdfunding. In other work [10] they show that building a community around one's project and engaging with that community is important for the success of crowdfunded projects. Likewise, Xu et al. [21] showed that effective project updates during the course of a Kickstarter campaign keep backers engaged and positively influence the chance of success for a project. Mitra and Gilbert [14] describe how the language used in project descriptions signals the quality of projects to potential donors.

Another form of coordination on crowdfunding sites is coordination and collaboration between project creators. Hui et al. [10] found that a number of communities have formed around crowdfunding, where creators discuss and critique ideas. Project creators may learn not only from other creators, but also from their own repeated experiences as project creators [9, 16].

Wash and Solomon [20] demonstrated that the discrete nature of many crowdfunding projects creates complementarities in donors' preferences, which means that coordination is required in order for all donors to get what they desire out of crowdfunding. They showed that in some cases, such as when a crowdfunding site uses an "all-or-nothing" mechanism and refunds donations to incomplete projects, getting more in donations may not lead to more projects being funded if the donations are not well coordinated.

Crowdfunding donors face a coordination challenge in that, while donors typically know how much they value a project, they don't necessarily know how others value it. This may be inferred from the amount of backing a project has received previously, but if people are free-riding, this inference could be incorrect and a project may not receive some donations simply because it does not appear to donors that others will also contribute.

Theory of Donation Timing in Charitable Giving

Crowdfunding projects collect donations over an extended period of time. The choices donors make about when to donate may have significant effects on the outcomes of crowdfunding projects. Agrawal et al. [2] argue that crowdfunding prompts "rational herding" where people are more likely to donate to projects when they have already received some donations from others, and that as a project nears its deadline the rate of donations accelerates. This finding has similarly been shown in a micro lending sites [22], which resemble crowdfuding. Kuppaswamy and Bayus [13] found that projects on Kickstarter tend to experience a "bathtub" pattern of donations over time: projects typically get many donations immediately after being posted, go through a period where few donations are made, and then as the deadline approached, re-

[1] `https://www.kickstarter.com/help/stats`, retrieved on June 4, 2014

[2] `http://www.donorschoose.org/about/impact.html`, retrieved on June 4, 2014

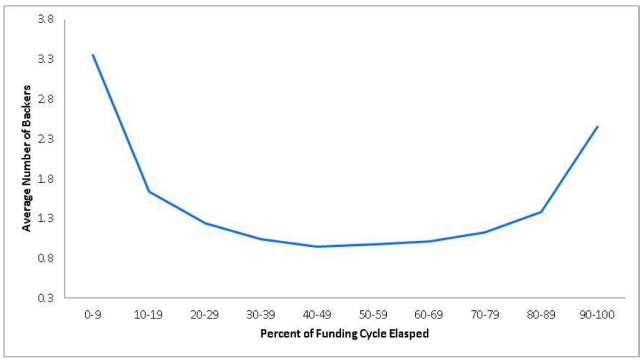

Figure 2. The timing of donations made to Kickstarter.com, as reported by [13]

ceive a last-minute surge of donations. Figure 2 illustrates their finding.

Public goods are often funded through philanthropy: charitable giving by wealthy (and sometimes not-so-wealthy) donors. Economists have identified a number of rational strategies for choosing when to give to a public good. First, many people choose to free-ride. That is, they choose to not donate to the public good, allow others to fund the good, and then reap the benefits once it is funded. This is a very common and rational strategy. However, if everyone chooses this strategy, then no public good would ever be funded.

A second strategy is to wait until asked to donate, and then contributing to whichever charity asked. Dubbed "the Power of the Ask", this strategy is a minimal-effort strategy that allows people to make charitable donations without the effort of deciding where to donate [4]. Together, these first two strategies represent a *lack of coordination*. Anyone using these donation strategies is explicitly forgoing coordinating with other donors about which projects or charities to fund.

Many donors often intentionally choose a wait-and-see approach to charitable donation. Because there is risk associated with donating, it is rational for a potential donor to wait and see how many other people are donating. This allows the potential donor to assess the likelihood that the project will be completely funded, or to assess the quality of a project based on the assessment of the crowd [3]. By waiting, donors may also benefit if the project is able to be completed without their donation at all.

Because many potential donors take a wait-and-see approach, another strategy is to make a large donation early during a fundraising campaign. Charities often solicit a "leadership donation" – a large donation from a well-known donor for 25%-50% of the total funds needed. Andreoni argues that leadership donations provide a credible signal that a charity or project is high quality and also that the project is likely to receive the funds it needs [5]. It also reduces the remaining funds needed, making the fundraising goal easier to achieve. Thus, by giving early, leadership donors can influence which projects are funded both directly through their donation and indirectly by inducing others to donate.

Timing Donations on Crowdfunding Websites

These donation timing strategies present an interesting coordination dilemma for donors on a crowdfunding site. Typically, no single donor can fund a crowdfunding project, so it is important that donors coordinate their donations to ensure that the project of their choice receives the funds it needs. However, each person might prefer a different project be funded, and therefore must coordinate with others to decide which projects will be funded with the scarce resources available in the crowd.

Wash [19] found evidence of a "completion bias" in crowdfunding projects on Donors Choose: the last person to make a donation to a completed project typically donates far more than an average amount. This could be a selection-and-timing effect, where people who want to make large donations intentionally wait until they are certain that their donation will go to a successful project – the wait-and-see approach identified for charitable giving. Or it could be that donors who hear about a project later experience very little risk, and thus are willing to make an increased donation.

Shin and Jian [17] found evidence of leadership giving in crowdfunding. They observed that most early donations to projects come from friends or family of the project creator (who may be most motivated for the project to succeed). However, leadership giving is risky; if the project never reaches its goal and receives the funds it needs, then the money that was donated is tied up for a period of time, unable to be used more productively.

It is difficult to determine from existing research and data on crowdfunding which of these strategies are being used. Perhaps those who wait till the end of a campaign are waiting because they are trying to free-ride off the donations of others. Or they may be herding and only see value in a project because enough other people have expressed value through their donations. We do not know with enough precision the motivation or value that a donor has for making the donation. We also can't know whether people who donate early are the ones who really like the project most and want to encourage others, or if they are simply expressing support for the creator because of the personal connection and not as much for the project itself. Similarly, those who wait may be the ones who like the project most, and the completion bias found by Wash [19] offers some evidence for this. Without having more specific information about individual donors' motivations and valuations of projects, our understanding of the timing dynamics is incomplete. In this paper, we report on a lab-based study where we control people's preferences and thus are able to better understand the strategies being used.

METHODS

In order to understand timing choices in crowdfunding, we created an experiment that provided people an opportunity to make decisions about when to donate to crowdfunding projects. This experimental approach allows us to completely control the environment: we controlled exactly which projects could receive donations, each person's budget for donating to crowdfunding projects, and the number of other potential donors.

Also, critically, we were able to assign *preferences* to subjects. For each available project, every subject was assigned a payoff – a number of credits that they would receive if that project is fully funded (regardless of whether that subject donated to the project). At the end of the experiment, subjects exchanged the credits that they earned for real money (100 credits = $1 USD). Vernon Smith showed that by assigning preferences and paying based on credits earned, this structure effectively induces the subjects to value the projects as they are assigned to [18]. Once subjects value the projects, they are likely to make decisions about these projects in similar ways as they do about real-world projects that they value. This Induced Value Theory [18] has been the basis of much of experimental economics in the last 30 years.

Assigning preferences to subjects also allows us to control the difficulty of the coordination problem. If we assign high preferences to many people for a given project, then that project is easy to fund. Likewise, if few people value a project, then that project will be more difficult to fund. Thus, we can effectively create a variety of different types of projects simply by varying the distribution of preferences for a given project across subjects.

Simulated Crowdfunding Website

Our experiment followed a similar setup to the crowdfunding game used by Wash and Solomon [20]. A group of six subjects formed a crowd of visitors to a simulated crowdfunding site. In this setup, subjects were allotted credits that they could donate to the three projects on our simulated crowdfunding site. Projects were available for donations for 60 seconds, and subjects could make donations at any point during the period. The three projects had no descriptions and were labeled only as "Red", "Yellow", or "Blue" projects. Each project had a goal of 100 credits. Figure 3 shows the site as subjects saw it.

Subjects were instructed that if the project was funded by the end of 60 seconds, they would receive their pre-specified payout as a bonus payment in credits. Each subject was given a budget of 30 credits per project that could be donated, and this budget could not be transferred to other projects. This feature of the design ensures that projects on the site are not actually in direct competition with each other for donations. Donating to the Red project, for example, does not in any way diminish one's ability to donate to the Blue project. Although they appear on the site simultaneously, there was no economic incentive to withhold donations or make a strategic choice about timing donations from one project because of the status of any other project. This reflects the common view that people make independent donation decisions about crowdfunding projects, rather than comparing projects and deciding which to donate to. This also allows us to treat the projects as independent in our analysis.

Subjects were free to donate any amount within their budget at any time during the 60 seconds. Subjects were also free to donate as many times to a project as they wished. For this reason, the number of donations that could be made to a project was limited only by the time available. The total amount in credits of donations was only limited by the budget.

Project:	Difficult		Easy		Medium	
Condition:	1	2	1	2	1	2
Subj. 1	0	15	30	30	45	25
Subj. 2	0	15	30	30	45	25
Subj. 3	0	15	30	30	45	25
Subj. 4	15	15	30	30	15	25
Subj. 5	15	15	30	30	15	25
Subj. 6	15	15	30	30	15	25
Total	45	90	180	180	180	150

Table 1. Payoff a subject receives, in credits, when a given project is funded. In condition 1, the medium project has the same total payout as the easy project but unevenly distributed payouts. In condition 2, the medium project has a lower total payout but everyone has identical payouts.

A strategy that we expected to see was for subjects to wait until the last possible moment to submit a final donation. Because we wanted to be sure that we captured all such attempts to make a "last-second donation" and not have the results depend on subjects' ability to time their clicks of the mouse button precisely, we configured the interface to treat the final ten seconds of the round as being effectively the final second. For this reason, in the final ten seconds of each round, project totals stopped being updated on subjects' screens and subjects could only submit one tim in this period. Subjects were instructed to treat this period as the "final second."

Groups played three practice rounds followed by fifteen live rounds where earnings were recorded. After each round, groups were re-formed to avoid problems that can arise from repeated games [3].

We recruited 120 undergraduate students (54% female, average age of 20 years) by email from our university to play this crowdfunding simulation. Only 33% of subjects indicated that they had ever visited a crowdfunding website previously. Subject earned an average of about $20 for participating in this hour-long study.

Creating Projects and Preferences

Projects varied only in the preferences assigned to the potential donors, as induced by the payouts offered for completion. Table 1 lists the payout structure of the experiment. This payout structure created three classes of projects that relate to how much total interest there was in a project (i.e. the sum of all donors' payouts) and the distribution of those payouts (i.e. evenly spread out so all donors receive the same payout, or uneven payouts where some donors receive larger payouts than others).

We have labeled these projects in this description according to their relative difficulty in funding, as determined by the results of the study. Easy projects were funded most frequently, medium projects were funded somewhat rarely, and difficult projects were almost never funded (because funding this project required irrational donating). We anticipated how difficult each type of project would be to fund through pilot testing, and structured the experiment into two conditions that represent two different versions of a crowdfunding site. Each

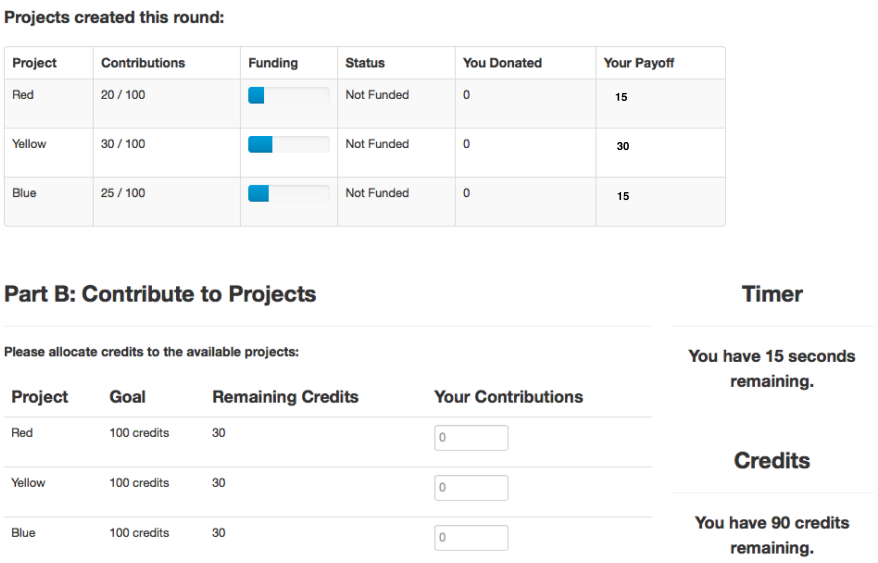

Figure 3. Crowfunding interface used in the experiment

site had three projects, one that was easy, one medium, and one difficult. The primary difference between the two sites was the nature of the medium project. In condition 1, the medium project had a high degree of overall interest, based on the sum of all payouts, but the payouts were unevenly distributed so that some people valued it more than they had budget to donate, and others had only a small preference. This created difficulty because if any one person with a high value decided to free-ride or take a wait and see approach, it became difficult for the rest of the group to rationally fund the project. Likewise in condition 2, the medium project had evenly spread out preferences but they were smaller overall. This similarly reduces the margin of error that users have for coordinating their donations successfully and still earning a payout from the project.

Subjects knew their own payouts for each project, but were not explicitly given any information of others' payouts. Any information they gained about other donors had to be inferred by observing donations to projects over the course of the 60 seconds.

RESULTS

Descriptive analysis

How much did people donate?

Of the 30 credit allotment given for each project, subjects donated an average of 10.39 (SD = 10.23) credits to each project on the site. Figure 4 shows the distribution of donations made to a project broken down according to the subject's payout if the project was completed. Subjects on average made 1.3 donations to a given project over the 60 seconds in the round. Figure 5 describes this distribution. Repeated donations from the same donor occurred in less than half of all observations. It was more common to either free-ride or to only make a single donation.

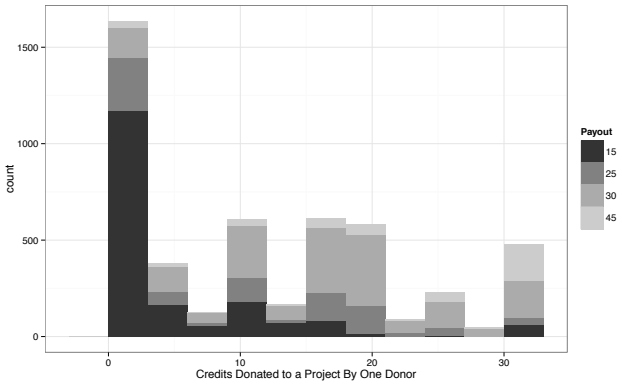

Figure 4. Distribution of donation amounts

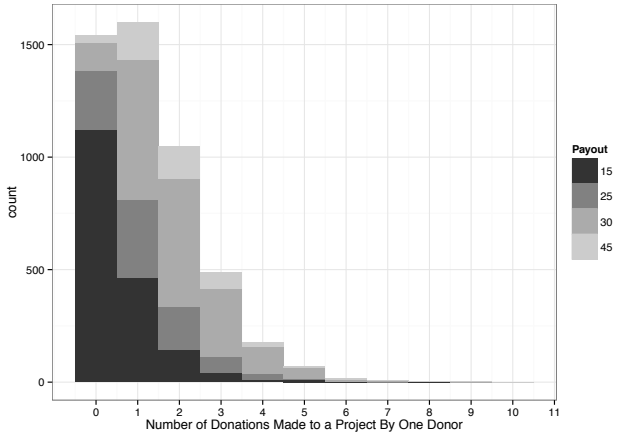

Figure 5. How many times each person donated

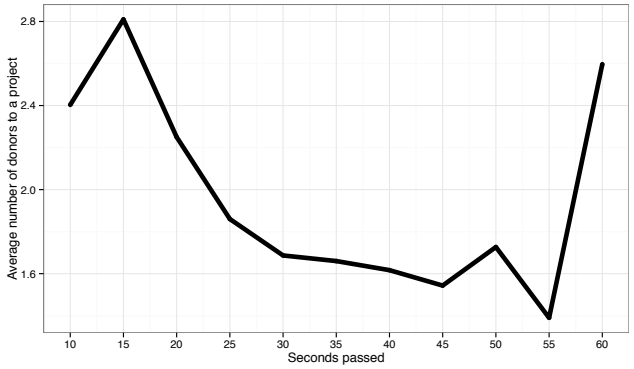

Figure 6. Average number of donations to a project over time

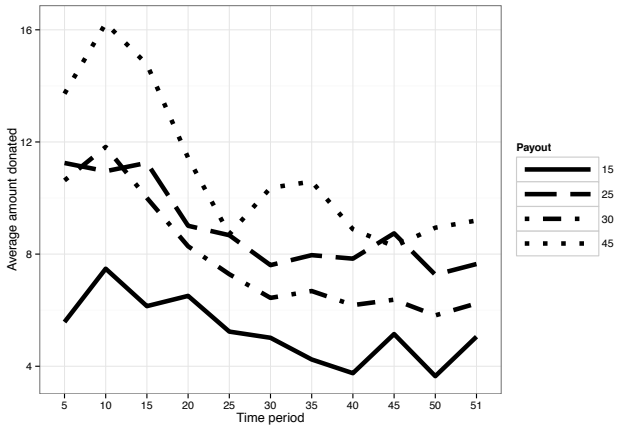

Figure 7. Average donation amount at each time period. These averages exclude non-donations in the period

Together, these results illustrate that free-riding – donating zero or a very small amount and letting others contribute most of the needed funds – was a popular strategy among donors. Free-riding was particularly common when subjects had a payoff of 15 credits, which was a low payout amount.

When did people donate?

Figure 6 describes how many donations were made to projects over the course of the 60 second campaign. Donations quickly hit a peak between about 10 and 15 seconds into the 60-second fundraising campaign and then slowly decreased in frequency until another peak right before the end. This pattern is highly similar to what has been observed by other research on live crowdfunding sites (compare Figure 6 to the Kickstarter data shown in Figure 2 above).

When subjects did make donations, there was a general tendency to make larger donations towards the beginning of the round, and smaller donations later on. Figure 7 shows the average size of donations made at each 5-second block within the round (not including donations of 0). This is in contrast to the finding of Wash [19], who found that the last donation of a project was usually much larger than other donations.

How successful were the crowds at funding projects?

	Project:	Difficult	Easy	Medium
Cond. 1	Success Rate	0%	54%	35%
	Credits Received	13.7	96.0	85.9
	# of Donations	1.3	11.8	9.5
	Payouts	Uneven	Even	Uneven
Cond. 2	Success Rate	1%	52%	18%
	Credits Received	19.2	94.0	65.3
	# of Donations	1.8	11.5	7.2
	Payouts	Even	Even	Even

Table 2. Donation statistics for each project type, averaged across all projects of that type

The projects varied widely in their likelihood of being funded. The easy project was, unsurprisingly, the most frequently funded project, and the difficult project was only funded once. The medium projects, however, showed an interesting pattern. The medium project with a high total payout but preferences unevenly spread through the crowd was funded 35% of the time. The other medium project, with a lower total payout but an even distribution of preferences, was funded 18% of the time. This difference is statistically significant according to a Chi-square test ($p < .001$). This suggests that it is better to have more total interest in the community than to have everyone like it somewhat. Table 2 describes the outcomes for all types of projects.

Identifying Strategies

One of our goals was to identify different strategies that subjects use when making donations. To do this, we used k-mean clustering analysis. The unit of analysis is the amount of donation from an individual at a time period. We cut the entire 60 second session into 3 periods: The first 15 seconds, 16 to 50 second and 51 to 60 seconds. This was based on our initial analysis (Figure 6) where we noted that these three periods had different donation patterns overall. In this analysis, we only examine first-time donations to a project. The first-time donation is the best representation of a donor's strategy, and first time donations are more fairly comparable to each other than subsequent donations. This is because donor's have the choice to make a first-time donation at any point in the round, and they always have an equal amount of budget when making the donation. Subsequent donations are biased towards the end of the round, which makes it difficult to assess the degree to which they act as leadership donations. Additionally, the vast majority of users make only 0 or 1 donation to a project (see Figure 5).

The goal of the cluster analysis was to find donations that were similar in terms of the timing and amount of the donation, the donor's payout, and the amount the project had already received. This is done by finding groups that maximize the distance between clusters and minimize the distance within. Through exploration, we found that seeking four clusters provided only marginal improvement in overall fit compared to three, so we settled on three clusters.

Figure 8 illustrates the clusters, broken down by the timing and amount of donation made by the observations. Each

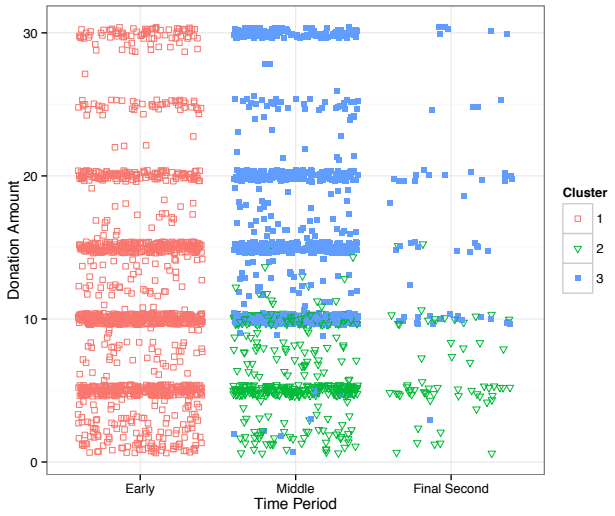

Figure 8. Strategy clusters

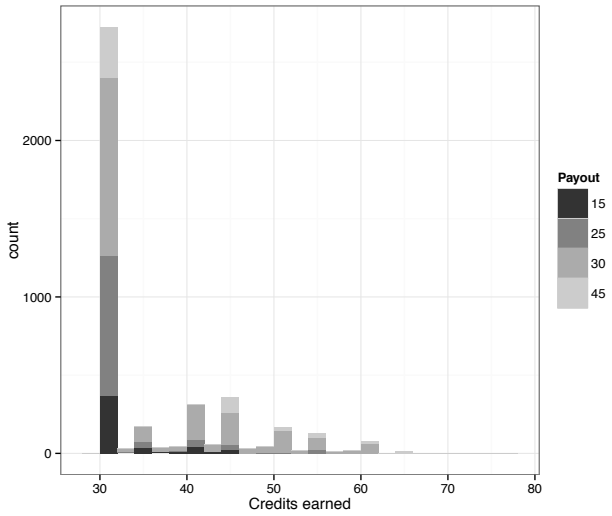

Figure 9. Distribution of the number of credits earned from a project

of the three clusters represents a separate strategy taken by donors. One strategy involved making a donation of any size in the initial 15 second period of the round (cluster 1). Another strategy was to make a larger donation (approximately 10 credits or greater) in either the middle period or the "Last chance" period (cluster 2). The third strategy was to make only a small donation in the middle or final stages of the round (cluster 3).

These clusters only include instances where a subject made any donation in the round. There is a fourth strategy which can clearly be seen in Figure 4 which was to completely free-ride and never make a donation.

We speculated that the size of one's payout would be related to which strategy a donor chose to take. However, removing payouts as one of the factors led to almost no changes in the clusters. We also noted in a visual analysis of donations that the proportion of donations made by each payout level remained constant over the course of the round on average. This means that the size of one's payout did not influence donors' strategies.

These clusters roughly correspond to the strategies identified in the literature review above. Donations in cluster 1 are leadership donations – early donations that signal to others which projects are likely to succeed. Cluster 4 (non-donations) are free-riding. Clusters 2 and 3 appear to be variations on the wait-and-see strategy, though it isn't clear how that strategy is playing out.

Having an Impact

To examine which strategy worked best for donors, we plotted the growth patterns for the three types of fundable projects in Figure 10. Projects that were ultimately funded differ from the unsuccessful projects. A separation in growth appears within the first 15 to 20 seconds for each type of project. For the High Interest projects – the easy project and the medium project with a large payout and uneven spread – this separation does not grow much, if at all, over the rest of the round. For these types of projects, the difference between being funded or not is a direct effect of the donations made early on. In the middle period of the round, the growth rates are equivalent. But the extra donations made early make it more likely that when the final second comes, that someone will donate to complete the project.

For the Medium difficulty project with lower, evenly spread payouts, the early donations are even more important. Unlike the High Interest projects, donations slow down considerably for this kind of project when it does not receive many early donations. When this project was funded, it received a slightly higher rate of initial donations which then was sustained somewhat constantly over the round until it was funded. When these early donations did not happen, donors ignored this project and it received few further donations.

This suggests that as a strategy for personal gain in our crowdfunding simulation, it is generally better for a donor to make an early donation than to wait. To examine this idea in more detail, we ran a set of regression models that estimate the profitability of different donation strategies.

When a project is not funded, the all-or-nothing structure of the site meant that the subject would still earn 30 credits. Since less than 40% of projects were funded, the distribution of earnings was somewhat unusual and heavily inflated with earnings of 30 (see Figure 9). Therefore, we built two models to analyze whether donating early led to better outcomes for donors. The first model estimates the probability that a subject earned any "profit" at all (payout greater than 30) using logistic regression.

When removing projects that were not funded, the remaining distribution is approximately a poisson distribution. Therefore in our second model we estimated the number of credits earned from a funded project based on the timing of the first donation grouped into 15 second intervals. For this model we used a poisson regression – a generalized linear model where

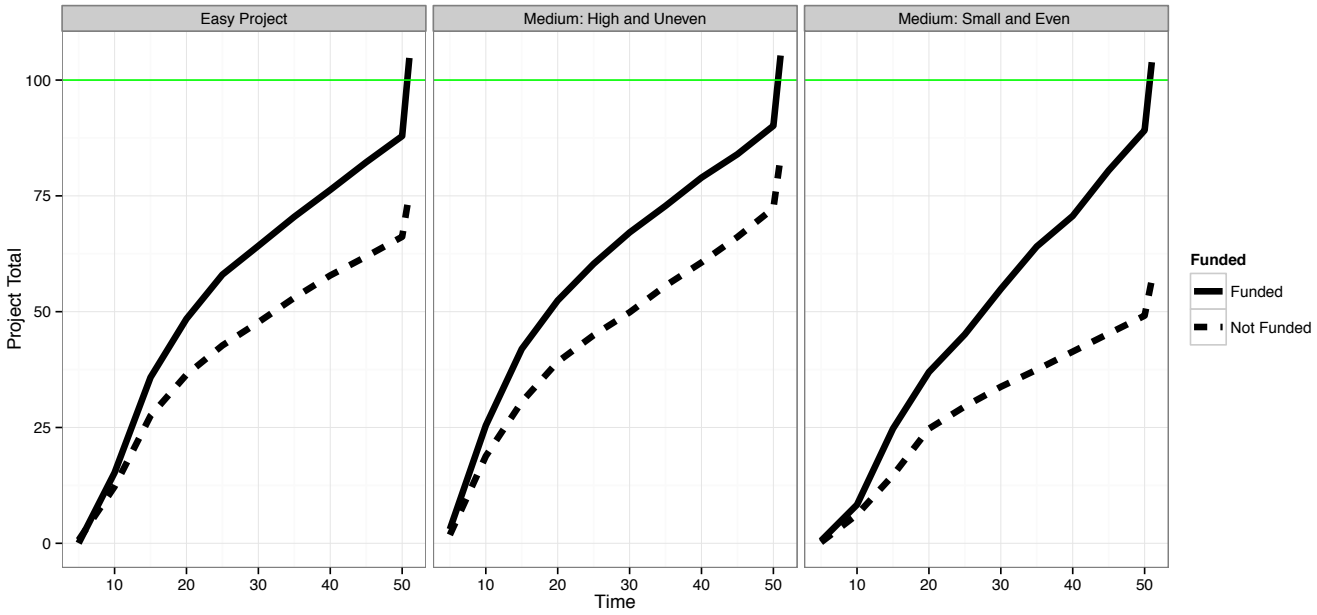

Figure 10. Comparison of growth of funded and un-funded projects

the errors are distributed according to a poisson distribution and the natural logarithm as a link function. These models are represented in Table 3. Both models include the subject's payout for the given project and a random effect of the subject (due to the repeated nature of the game). The Intercept in these models represents those who donated in the first 15 seconds. The estimates represent the changes in log odds of earning a profit (Model 1) and the natural logarithm of credits earned above 30 (Model 2) that can be expected from a one unit change in the independent variable.

Model 1 suggests that it is a poor strategy to wait longer than the initial 15 second period to make a donation, though waiting until the last moment is almost as good. Subjects who donated between 15 and 45 seconds were less likely to earn additional credits from the project. Additionally, subjects who never made a donation were very unlikely to earn additional credits. This model describes the probability of earning a profit, but in our study, this is nearly completely synonymous with the probability of a project being funded (since subjects almost always donated less than their payout). Therefore, model 1 can also be interpreted as the effect of the timing of one's donation on the probability of that project being funded.

However, model 2 suggests that if a project was funded, the subjects who waited till the end or did not donate did in fact earn more credits. To get a better overall picture for the value of being an early contributor, we conducted a Wilcoxon Rank Sum test on the total credits earned from a project that compared those who donated early (within the first 15 seconds) and those who did not. The test indicated that those who donate early did earn more credits ($p < .001$), although the difference in means between the two groups was less than 1 credit.

	Dependent variable:	
	Profited?	Earnings
	(1)	(2)
Intercept	-0.102	1.462^{***}
	(0.073)	(0.047)
15-30 Seconds	-0.239^{***}	0.033^{*}
	(0.083)	(0.019)
30-45	-0.282^{**}	0.186^{***}
	(0.113)	(0.025)
After 45	-0.091	0.295^{***}
	(0.130)	(0.027)
Never	-1.626^{***}	0.569^{***}
	(0.134)	(0.030)
Amount Donated	-0.019^{***}	-0.022^{***}
	(0.005)	(0.001)
Payout		0.041^{***}
		(0.001)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3. Effect of donation timing on earnings from projects

It is worthwhile to note that free-riding was a very poor strategy in this study. Free-riding on average led to much lower earnings than when even small donations were made. This likely speaks to the value of a donation both as an act of funding a creator's idea but also as a coordination signal sent to other potential donors.

DISCUSSION

Donors in the experiment, like real users of crowdfunding sites, face a difficult decision around when to make a dona-

tion. Donating early sends a behavioral signal to others that encourages them to donate as well, and improve the chances of the project being successful. However, the project may be successful without one's donation, and by waiting, donors may be able to reap the benefits of a completed project without donating or may only need to make a smaller donation towards the end. This is a potential cost of sending coordination signals to other users.

In our study however, donating early was overall a better strategy for donors. This was particularly true for donors with the most to gain from a project being funded; donating early leads to a higher likelihood of profit. Though, once you know the project will be funded, it is better to donate late because that allows you to partially free-ride using a smaller donation. Conversely, those with only small payouts, equating to those with relatively low preference, may do better by waiting till the end and making a small donation if the project is close to being funded.

This result has some important implications for crowdfunding sites. First of all, it suggests that the rigid all-or-nothing deadline structure of many crowdfunding sites may create some inefficiencies with the way people coordinate to fund projects. Deadlines may create an incentive for people to wait, and if too many people wait then projects that otherwise have enough interest to be completed may not be funded. In our simulation, the collective payouts of most projects was over 100, meaning that it was profitable for the crowd of donors as a whole to fund the project. Since this happened less than 40% of the time for these projects, we argue that there was inefficient coordination among the donors on a site.

A noteworthy result is that the medium project with uneven preferences was funded less frequently than the easy project. Although this was expected, it is noteworthy because the sum total of all payouts to donors was actually the same for both projects. The only difference between them was the difference in how those payouts were distributed. In some subsequent exploration of this result, we noted that the medium project almost always failed if one of the three donors with high preference (45 credit payout) decided to wait or to free-ride altogether. When projects have an uneven distribution of preference across the population of potential donors, it is critical that those with with high preference donate early because if they do not, there is not likely to be anyone else who will. When preferences are evenly distributed, there is a greater chance of early donation because if one person free-rides or waits, there is a larger pool of potential replacements for that donation.

Deadlines do have an important role in crowdfunding. Deadlines are necessary for the all-or-nothing style or crowdfunding to be functional. All-or-nothing crowdfunding minimizes the risk for a donor associated with donating [20]. Therefore, it seems logical that this approach would encourage more early donations. Our data do show many early donations, as do real crowdfunding data collected from Kickstarter [13]. But in our study, one or two "missing" donations in the early period was frequently the only difference between funded and un-funded projects. Therefore, it is critical for projects to absolutely maximize their early donations. Even though all-or-nothing may minimize the incentive to wait, it does not eliminate it. People can still estimate for themselves that they may be able to free-ride and still reap the benefits of the project, or at least minimize the size of contribution they need to make. It is also very important to consider that waiting to donate towards the end was a good strategy if the project was funded. That is to say, if a project only needed a small donation at the deadline, the person who waits till the end then makes that small donation ends up with a large profit.

Design Implications

What can crowdfunding sites do get people to donate at the start of a project's campaign? One existing structure likely has a positive influence. Many projects offer potential donors some form of personalized reward or perk in exchange for a donation. Often, projects set limits on how many of these rewards are given out to donors, which gives an incentive to donate immediately. It is a limitation of our study design that we have treated crowdfunding projects as pure public goods, when in fact the rewards offered by projects add some additional complexity.

One potential design for crowdfunding sites that could maintain the all-or-nothing structure, but possibly lead to more early donations, would be to set a mandated pace for donations. Projects might have multiple check-in points during the time period of the campaign, and failure to maintain a pre-specified funding pace at any of these points would result in the project being closed and donations being immediately refunded. This design may relieve the coordination dilemma that can occur because it encourages people to send signals of interest in a project (by way of donating) immediately, which then gives other potential donors a more accurate estimate of the true interest the crowd has in a project.

Another design might keep the current status or the total funding goal hidden from donors when the project gets close to reaching its goal. In this design, when a project meets its goal, it remains open for some period of time and can collect additional donations, since the status is not communicated. As a project grows beyond its goal, donations would not go to the project but rather to early donors. In our experiment, project status was not updated to donors in the final period and as a result, most projects that were funded received some excess donations. Returning these donations to early donors would offer a new incentive to donate early and express one's preference for a project rather than waiting or free-riding.

Limitations

This study has several limitations that should be considered. In our experiment, we created a simplified replication of a crowdfunding site that is has some clear differences from most real crowdfunding sites, such as the extremely limited amount of time of a campaign and the limited amount of information available about projects. In particular, our experiment was designed such that the only information donors had with which to coordinate was a project's funding total at a given moment in the campaign, whereas much richer information is available to donors making real crowdfunding

decisions. Future work should examine how other type of information or communication afforded by crowdfunding platforms might influence timing decisions, or how our results may be moderated in more authentic crowdfunding scenario.

Our study also simulates the decisions of a large number of distinct people by having a smaller number of people make multiple decisions, such as how many times to donate and how many projects to donate to. This is demanded by the complexity of recruiting and coordinating participation in an experiment. Although the study successfully replicated the results of real crowdfunding sites by condensing the "crowd" in this way, it should be noted as a limitation because our simulation may have achieved similar results but for different reasons.

CONCLUSION

Overall, we can conclude that all-or-nothing crowdfunding with a deadline creates an inefficiency because it encourages people to withhold their donation. By withholding donations, people not only withhold funds from a project but also a signal to others about the crowd's interest in a project. Without these signals, donors do not efficiently coordinate and fund projects even when there is sufficient interest. Crowdfunding sites should explore ways to increase early donations so that crowd interest in projects is effectively communicated and donations are coordinated.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation, award number CCF-1101266.

REFERENCES

1. Agrawal, A., Catalini, C., and Goldfarb, A. The geography of crowdfunding. NBER Working Paper #16820, February 2011.
2. Agrawal, A. K., Catalini, C., and Goldfarb, A. Some simple economics of crowdfunding. Tech. rep., National Bureau of Economic Research, 2013.
3. Andreoni, J. Why free ride?: Strategies and learning in public goods experiments. *Journal of Public Economics 37*, 3 (1988), 291 – 304.
4. Andreoni, J. Philanthropy. In *Handbook of Giving, Reciprocity, and Altruism*. Elsivier / North Holland, 2005.
5. Andreoni, J. Leadership giving in charitable fundraising. *Journal of Public Economics 8*, 1 (2006), 1–22.
6. Belleflamme, P., Lambert, T., and Schwienbacher, A. Crowdfunding: Tapping the right crowd. Tech. Rep. 1578175, Social Science Research Network (SSRN), April 2012.
7. Gerber, E. M., and Hui, J. Crowdfunding: Motivations and deterrents for participation. *ACM Transactions on Computer-Human Interaction (TOCHI) 20*, 6 (2013), 34.
8. Greenberg, M., and Gerber, E. Crowdfunding: A survey and taxonomy. Tech. Rep. 12-03, Northwestern University Segal Design Institute, 2012.
9. Greenberg, M. D., and Gerber, E. M. Learning to fail: experiencing public failure online through crowdfunding. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, ACM (2014), 581–590.
10. Hui, J. S., Greenberg, M. D., and Gerber, E. M. Understanding the role of community in crowdfunding work. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '14, ACM (New York, NY, USA, 2014), 62–74.
11. Jian, L., and Usher, N. Crowd-funded journalism. *Journal of Computer Mediated Communication 19*, 2 (January 2014).
12. Kickstarter. Statistics. `http://www.kickstarter.com/help/stats`, June 2014.
13. Kuppuswamy, V., and Bayus, B. L. Crowdfunding creative ideas: the dynamics of projects backers in kickstarter. *SSRN Electronic Journal* (2013).
14. Mitra, T., and Gilbert, E. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '14, ACM (New York, NY, USA, 2014), 49–61.
15. Mollick, E. The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing 29*, 1 (2014), 1–16.
16. Pak, C., and Wash, R. Importance of recent experience and initial overconfidence:factors of crowdfundiners' limited learning. Working Paper, June 2014.
17. Shin, J., and Jian, L. Driving forces behind readers' donation to crowd-funded journalism: The case of spot.us. Working paper, University of Southern California, 2012.
18. Smith, V. L. Experimental economics: Induced value theory. *American Economic Review 66*, 2 (May 1976), 274–279.
19. Wash, R. The value of completing crowdfunding projects. In *International Conference on Weblogs and Social Media (ICWSM)*, AAAI Press (July 2013).
20. Wash, R., and Solomon, J. Coordinating donors on crowdfunding websites. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '14, ACM (New York, NY, USA, 2014), 38–48.
21. Xu, A., Yang, X., Rao, H., Fu, W.-T., Huang, S.-W., and Bailey, B. P. Show me the money!: An analysis of project updates during crowdfunding campaigns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, ACM (New York, NY, USA, 2014), 591–600.
22. Zhang, J., and Liu, P. Rational herding in microloan markets. *Management science 58*, 5 (2012), 892–912.