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-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.
Copyright © 2014 ACM 978-1-4503-2473-1/14/04 ...$15.00.
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

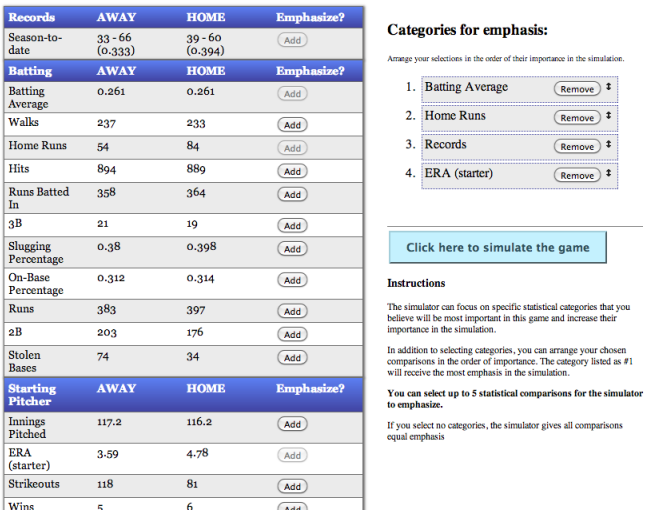

Figure 1. Customizable DSS

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

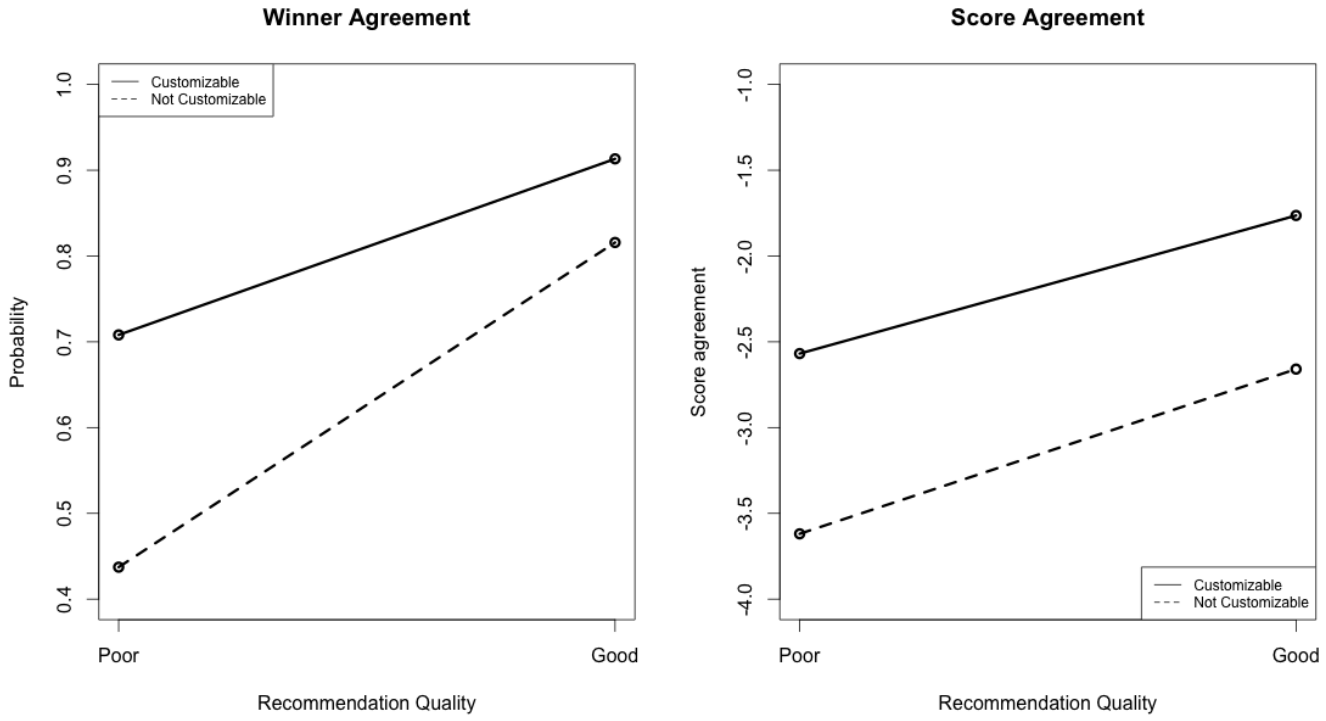

Figure 2. Agreement with DSS recommendations

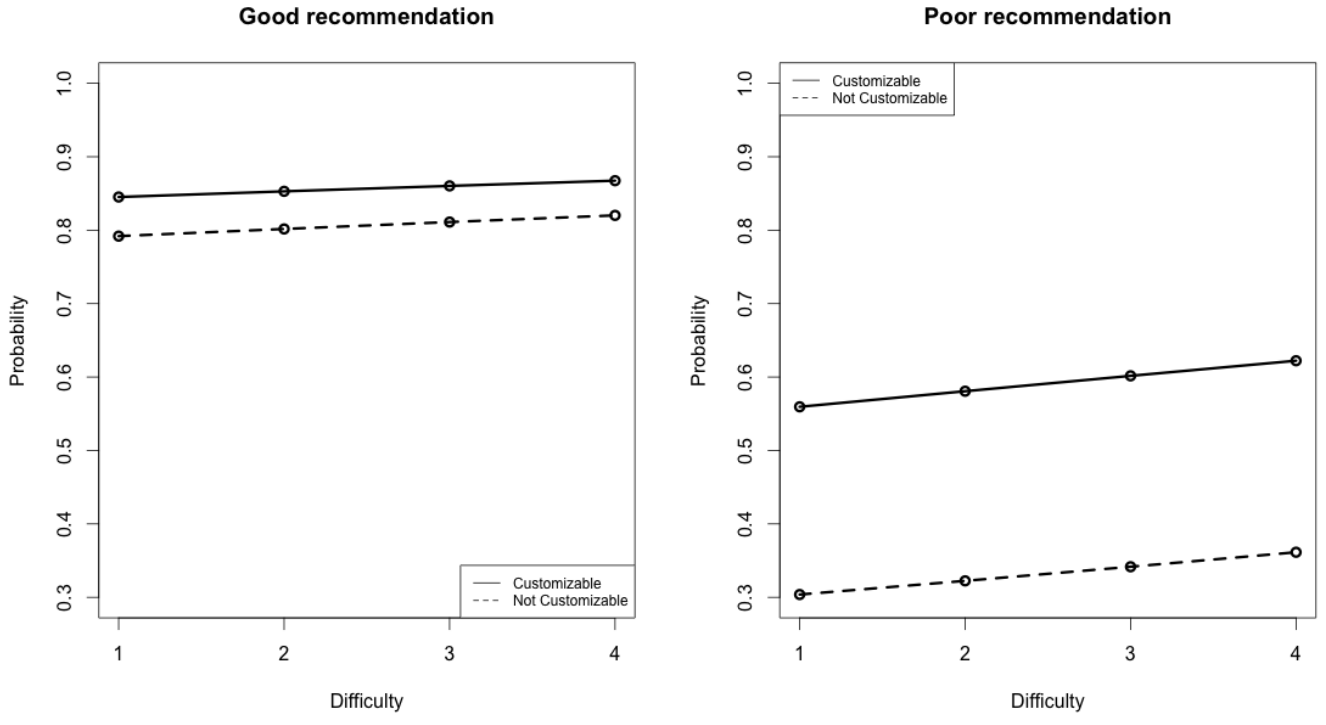

Figure 3. Effect of difficulty on customization bias

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

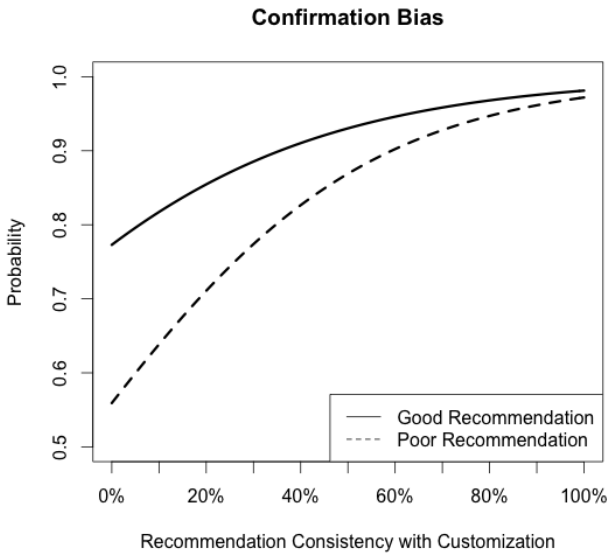

Figure 4. Customization/Recommendation Consistency on Probability of Winner Agreement

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.