12-2012

Preventing Misuse and Disuse of Automated Systems: Effects of System Confidence Display on Trust and Decision Performance

Margaux Price
Clemson University, margauxmae@gmail.com

PREVENTING MISUSE AND DISUSE OF AUTOMATED SYSTEMS: EFFECTS OF SYSTEM CONFIDENCE DISPLAY ON TRUST AND DECISION PERFORMANCE

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Human Factors Psychology

---

by
Margaux M. Price
December 2012

Accepted by:
Dr. Richard Pak, Committee Chair
Dr. Leo Gugerty
Dr. Christopher Pagano
Dr. Joel Greenstein

ABSTRACT

Complex decision-making may be aided by forms of automation known as decision-support systems (DSS). However, no DSS is completely reliable and so it is imperative that users know when they should and should not trust it (calibration of trust). Previous research has shown that providing users with information about the DSS's confidence in its own advice ("system confidence") can help improve the calibration of user's trust of automation and actual system reliability on a trial by trial basis. The current study examined how the nature of the presentation of system confidence information affected user's trust calibration. The first study examined the attentional demand of each display, while the second study examined their effect on trust and performance on a decision making task. The results of this study indicate that there was no effect of system confidence display type on subjective or objective trust. The lack of differences in performance or trust between the control condition (no system confidence display) and other displays raises doubts about whether users were utilizing the system confidence information or using reliability information. The type of decision task may be crucial in determining whether to provide system confidence and these results suggest that it should be tested prior to implementation against a control group, unlike previous studies. The results of these studies have implications in the design of DSS, especially given the difficulty of providing accurate system confidence information to users. The time and resources that would be required to provide such a display may not be beneficial if it has no effect on user trust or decision performance.

DEDICATION

First, I would like to thank my committee members, Lee, Chris, Joel, and Rich for their time and feedback. To my advisor, Rich, thanks for always pushing me to do my best, for supporting my ideas, and for all of the time you and effort you invested in making me a better researcher. To my lab mates in the CATlab, I am thankful for all you, there is no way I would have made it through graduate school without you! Nicole, thank you for always giving me “hope”,keeping me smiling, and reminding me of the positive side of everything, and for being such a great friend. Stephanie, thank you for being the best partner in everything from class to HFES, to shenanigans, you’re something else and I am so glad I had the opportunity to work with you. Ashley, you were the best roommate and friend a girl could ask for, thanks for always being there for me in and out of school.

To my family, thank you for always believing in me and supporting me. Dad, thanks for always leaving big footprints to fill, for always taking an interest in everything I’m studying, and most of all for your support. I’ve always tried to make you proud, and I hope to continue to do so. Mom, thanks for always telling me that whatever it is it can be done, helping me stay focused, and making me laugh when I needed it the most.

TABLE OF CONTENTS

Page

Table of Contents (continued)

Page

Table of Contents (continued)

LIST OF TABLES

LIST OF FIGURES

List of Figures (continued)

Figure Page

INTRODUCTION

Decision support systems (DSS) are automated systems that have the potential to help users make better decisions when there are numerous options and many attributes to consider. For example, when purchasing a computer online, the consumer may input the purpose of use (e.g., emailing or gaming), what features they value most (e.g., hard drive space, processing speed). Based on this information, the web-based aid may present the best choices of machines based on the user's needs. When purchasing a Medicare prescription drug plan using parameters such as a monthly and yearly budget maximums and coverage minimums, the options vary across several attributes and what's best for one person may not be the best for everyone. These types of decisions can be difficult if the consumer has no prior domain knowledge. For example, our computer consumer might need to understand the slight differences between attributes (e.g., types of RAM, monitor backlights) while our Medicare shopper must know the jargon (e.g., gap coverage or donut hole).

Fortunately, decision aids can help consumers narrow down the options, simplifying the decision-making process. There are also other DSSs such as GPS to help users find destinations more efficiently, financial DSSs that help predict future outcomes, and medical devices that help users make healthier choices. In all of these cases, the operator is the consumer who may be making a one-time or infrequent decision (e.g., kiosk or "walk-up-and-use"). The consumer may not be able to practice extensively to develop trust in the same manner that workplace operators of control systems do based on practice over time, training, and explicitly-provided knowledge about the DSS's reliability.

Most of the research in human-machine interaction with DSSs has focused on the workplace operator using complex automation, typically in high stress, high risk scenarios (e.g., Hancock & Parasuraman, 1992; Hilburn, Jorna, Byrne, & Parasuraman, 1997; Parasuraman & Hancock, 2008). This body of research has focused on designing systems that foster appropriate trust between the user and machine so that the benefits of a DSS can be realized and catastrophic events are avoided. While research for those types of systems is undoubtedly important, consumer DSSs impact the lives and health of many people as well. It's important then, to examine whether factors that lead to appropriate trust in highly risk scenarios are the same as those in a consumer decision task and determine whether the results of those studies can be extended to the consumer domain.

Research on trust in automation has shown that providing system confidence (or an estimation of how confident the system is in making its recommendation) may help users determine when to trust and not trust DSSs. Using our prior example, the Medicare shopper may benefit from knowing that the DSS is 75% confident that the plan it is recommending matches their needs. Several studies have shown that providing the user with a display of system confidence can improve appropriate use of DSSs in identification tasks and domain-specific (e.g., aviation) strategy decisions (McGuirl & Sarter, 2004; Spain & Bliss, 2009). Providing system confidence may also be able to improve appropriate use of consumer DSS.

The main purpose of this study was to examine the use of anthropomorphic presentations of system confidence. Prior research has shown that system confidence information can improve performance by influencing appropriate trust (i.e., using the automation only when it is accurate). In this study, we are specifically examining the mode of presentation of system confidence. Furthermore, this study aims to extend system confidence and automation research to consumer based DSSs, specifically in the context of choosing supplemental prescription drug plans.

Types of Automation

Automation can be classified by the type of task it performs and how it augments human performance at different stages of information processing. Parasuraman, Sheridan, & Wickens (2000) provide a classification for types of automation that map directly onto the stages of the information processing model (see Figure 1). This classification is important because it outlines the level of interaction between the human and automation. The type of consumer decision support system that this study focuses on is one designed to support decision making by reducing the amount of cognitive resources (i.e., attention, working memory) required to compare many options consisting of numerous attributes, and instead leaves the processing up to the DSS. DSSs that help users through the decision and action selection stage can be further defined by how much autonomy is given to the user and the automation (see Figure 2). The current study focuses on a DSS that consumers are likely to encounter; systems that fall between level 3 (the computer narrows choices down to a select few) and level 4 (the computer suggests one alternative).

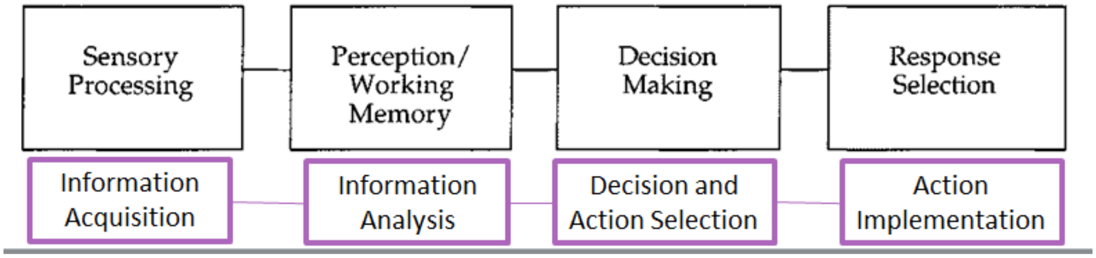
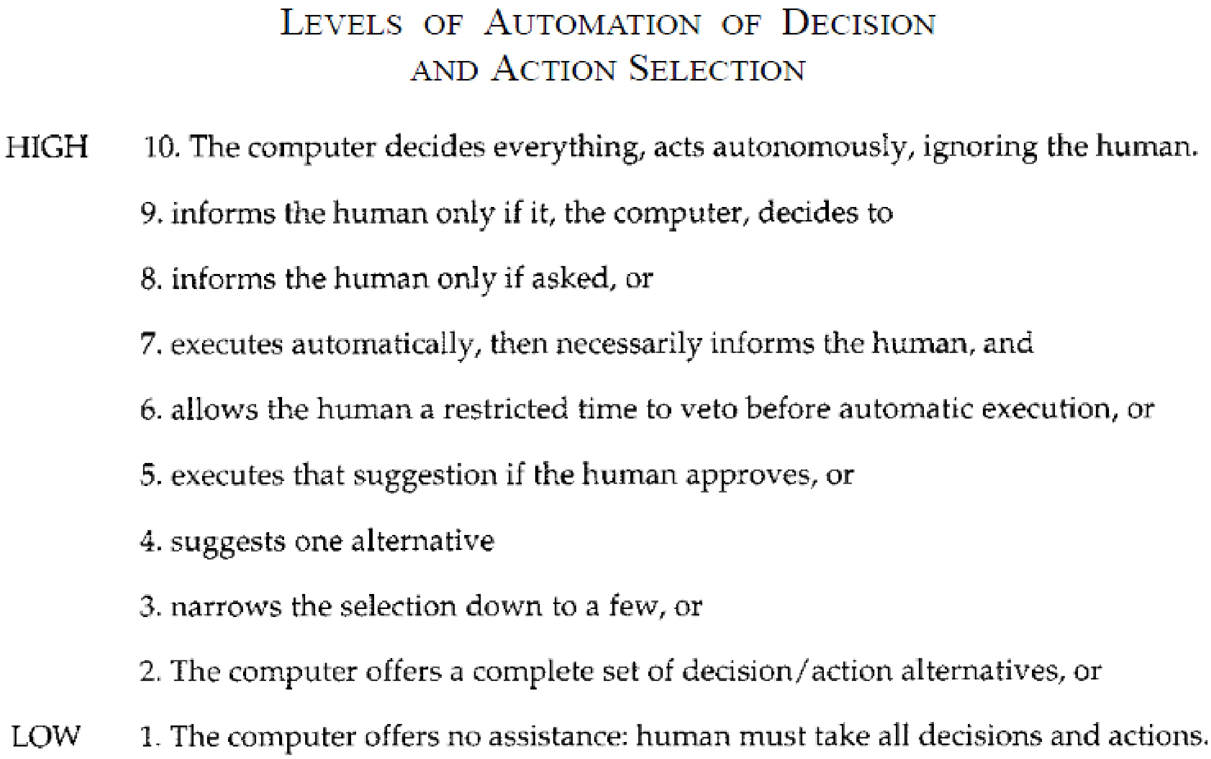


Figure 1. Stages of information processing, above, with automation classification types, below (adapted from Parasuraman, Sheridan, & Wickens, 2000).



LEVELS OF AUTOMATION OF DECISION AND ACTION SELECTION

HIGH 10. The computer decides everything, acts autonomously, ignoring the human.

9. informs the human only if it, the computer, decides to

8. informs the human only if asked, or

7. executes automatically, then necessarily informs the human, and

6. allows the human a restricted time to veto before automatic execution, or

5. executes that suggestion if the human approves, or

4. suggests one alternative

3. narrows the selection down to a few, or

2. The computer offers a complete set of decision/action alternatives, or

LOW 1. The computer offers no assistance: human must take all decisions and actions.

Figure 2. Levels of human interaction with automation and stages of information processing mapped onto types of automation (adapted from Parasuraman, Sheridan, & Wickens, 2000).

Decision Making and Automation

When the task requires more cognitive resources than the decision maker has available, the task can be considered a resource-limited task (Norman & Bobrow, 1975). Performance on a resource-limited task can only be improved if more resources are available to commit to the task. This is contrasted with data-limited tasks where providing more information can improve performance. For example, given a choice among 15 health insurance plans, the decision maker is faced with many comparisons along different attributes. This task is resource-demanding (as shown in a task analysis; Price & Pak, accepted) in that it requires working memory, numerical calculations, and comparisons. Particularly, non-compensatory decisions are resource demanding because users cannot make tradeoffs between attributes of options. Instead, each option must be considered attribute by attribute. In resource-limited tasks, providing more data (i.e., another option, attribute, or system confidence information) that the participant must consider will not improve performance because no resources are available to allocate to the new information.

DSSs may be most useful for decisions tasks that are resource-limited when the DSS is able to process all or part of the information, freeing resources for the decision maker. Ideally, decision makers would consider all options analytically, comparing each attribute for every decision option (i.e., expected utility approach). Decision makers faced with compensatory decisions in a resource-limited task and under time pressure cannot consider all options analytically (i.e., expected utility or tallying approach). Instead, decision makers tend to rely on other, less resource demanding strategies such as a take the best strategy (Gigerenzer & Goldstein, 1996), satisficing (Gigarenzer & Goldstein, 1996; Simon, 1955), or elimination by aspects (Tversky, 1972). These strategies reduce the amount of decision information that is attended to and the number of comparisons that are made, thus reducing resource demand. The benefit, then, of a level 4 DSS is that it does the processing required to consider all options analytically and algorithmically, provides a suggested option, but then leaves the judgment of the decision (whether or not to follow the automation's suggestion) up to the decision maker.

In level 4 automation, users must decide whether to follow the suggested option or expend resources verifying the suggestion. They have several options: 1) trust the DSS and agree with the option, 2) verify the option, then either agree or disagree, or 3) disagree, or distrust the system and find a suitable answer on their own. This stage of an information processing model of decision making, where the user must decide whether to trust the automation, is called the evaluation of outcomes stage. Users will sometimes resort to using judgment heuristics at the evaluation of outcomes stage when there is high workload or time pressure. Instead of expending time and resources double checking the automation's suggestion, the user may place value on past exemplar experiences (representative heuristic; Tversky & Kahneman, 1974), ease of retrieval (availability heuristic; Tversky & Kahneman, 1974), emotions (affect heuristic; Slovic, et al., 2005), or characteristics of the automation itself (e.g., anthropomorphic features). However, sometimes heuristic use leads to automation biases, when the user mistrusts or distrusts the system. Ideally, the user should be able to trust the DSS; however, no automated system is 100% reliable.

Basis of Trust in Automation

Trust, in the context of human-automation interaction, is the "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p 54). Trust is an affective response to features and interactions between the decision maker and the automation. A fundamental issue with automation is how to appropriately calibrate user's trust with actual system reliability, so that the user always uses automation when it is appropriate and does not use it when it is inappropriate. Calibration has been defined as the "correspondence between a person's trust in the automation and the automation's capabilities" (Lee & See, 2004; Muir, 1987). Thus, calibration depends on both system factors (e.g., how reliably the automation helps the user reach a goal) and human factors (e.g., whether the user trusts and complies with the automated).

Improper calibration can lead to performance decrements. If the user does not trust the automated system when it is in fact reliable (i.e. distrust or disuse), the consequence is less efficient performance because effort is allocated to "double checking" (i.e., a cost of verification). Misuse, on the other hand is the "overreliance on automation" (Parasuraman & Riley, 1997, p. 233) and occurs when the user inappropriately uses the automation for a task or decision it was not designed for or under conditions when the automation cannot make a valid choice (e.g., incomplete source data).

Two types of trust: Dispositional trust and history-based trust

Trust can be differentiated by its source (Merritt & Ilgen, 2008; Kramer, 1999). First, dispositional trust is the initial trust in a system *before* the user has extensive practice with the machine (i.e., the DSS). Dispositional trust is primarily affected by individual differences in personality, propensity to trust, self-confidence, as well as initial impressions of the features of the DSS (Fiske & Neuberg, 1990). For example, domain experience can also influence how users perceive errors and error attribution (Sanchez, 2006). Self confidence can affect trust because automation is more likely to be used if users have higher trust in the automation's capability to do the task than in their own ability (Lee & Moray, 1994). .

Second, history-based trust is a product of experience with a system and thus is shaped *after* the user has experience with the DSS. History-based trust is built upon how well matched the user's expectations of how the DSS will perform are to actual DSS performance (i.e., reliability and predictability). Although they are different constructs, initial perceptions (dispositional trust) can mediate human and machine characteristics and history-based trust (Merritt & Ilgen). Merritt & Ilgen stress the importance of measuring initial trust at different points of human-automation interaction, and individual differences in propensity to trust, when examining the impacts of machine characteristics on trust and outcomes of trust (i.e., performance). This distinction between dispositional and history-based trust is important in the current context because the user of common consumer-facing systems will have only dispositional trust and little or no history-based trust. Presenting system confidence information may provide this additional type of trust instantly.

DSS characteristics that influence trust

Trust may come from three general sources: performance, process, and purpose (see Lee & See, 2004 for a thorough review) (Lee & Moray, 1992; Lee & See, 2004). The performance source of trust “describes what the automation does”, and is “demonstrated by its ability to achieve the operator’s goals” (Lee & See, 2004, p 59). Dispositional trust or initial perception of the DSS can be influenced by presenting the operator with the reliability of the system. When told that reliability is high (>70%), trust in and reliance on automation increases. When reliability is low (<60%), users tend to distrust the automation, and thus will not rely on the automation and instead will likely switch to manual control (Dzindolet, et al, 2003; Lee & See 2004). However, trust and reliance are two separate constructs (Weigmann, Rich, & Zhang, 2001). Trust is an affective response, while reliance is a behavior. One can have low trust, but still rely on the automation because time constraints or workload makes it hard for the user to make the decision on their own.

Purpose describes the “degree to which the automation is being used within the realm of the designer’s intent” and “describes why the automation was developed” (Lee & See, 2004, p 59). Operators should understand the purpose of the automated aid so as not to misuse the automation in situations where it is inappropriate. Purpose may be influenced initially by the user’s initial perceptions of trustworthiness (e.g., a dispositional trust factor). How polite the system is, whether it matches their personality type, and what it looks like all may influence the decision maker's initial impression of trustworthiness. Purpose may also influence history-based trust. As the user gains experience with the system, he or she forms an opinion on how well the system matches the intended purpose.

Process is the "degree to which the automation's algorithms are appropriate for the situation and able to achieve the operator's goals" and "describes how the automation operates" (Lee & See, 2004, p 59). This dimension has less to do with actual performance and instead focuses on characteristics of the automated aid itself, such as understanding of the rules that govern the system, dispositional attributions and inferences (rather than actual reliability), and openness. Relying more on dispositional attributions and inferences makes process different than performance; performance relies on history-based trust and actual performance accuracy. This is similar to how people base trust in social interactions with humans (rather than machines) (Lee & See, 2004). When people receive advice from another human, they may judge the trustworthiness of the other person by their facial expressions, intonation, body language, personality, and etiquette (Parasuraman & Miller, 2004). History-based trust may be formed through previous interactions and outcomes of taking advice from that person. If there is not a history of interaction between the advisee and the advice giver, then initial perceptions (dispositional trust) may influence decision making. The decision maker may look for evidence of confidence from the advice giver as a source of validity of a single recommendation (Parasuraman & Miller).

As described earlier, a major problem in human-automation interaction is the calibration of a user's trust in the automation. Complacency refers to a user's over trust of automation especially when reliability is high (Parasuraman & Manzey, 2010). User's may become disconnected or lose situation awareness when this automation complacency occurs because they are "out of the loop" (i.e., do not have a good sense of how or why the automation came to its recommendation). One way to help users become more aware of the automation is by giving the user more information about the level of confidence that the automation has in its own recommendation. This may be functionally similar to getting advice from a friend where the person gives a likelihood of being correct.

System Confidence

One method of displaying a DSS's process information and thus reducing automation bias is by providing an estimate of system confidence (Parasuraman & Manzey, 2010). System confidence is the system's expression of how likely it believes it is correct and is based on the data it has available. One example of how system confidence could be used is with a GPS navigation system. The accuracy of the system is dependent on the quality of the information it is using to base its recommendation. The quality of the data could include: a) number of satellites it has locked-on to, b) the strength of the satellite signals, and c) date on which the maps were downloaded. A high confidence scenario would occur if there were a high number of locked-on satellites, high signal strength, and maps no older than 6 months. Conversely, a low confidence scenario would occur if the GPS only had a few satellites locked-on, the signal strength was weak, and the maps were older than 1 year. In this situation, the GPS may provide the user with a percentage of the signal strength (i.e., its own measure of system confidence) giving the user information about its own certainty based on the underlying data.

In a perfect system, high system confidence would always be correlated to reliability and thus trusting a highly confident system would lead to good performance. However, system confidence may not always be positively correlated with reliability and performance. If the DSS is getting degraded or is receiving incomplete information and thus has low system confidence, it still may have high reliability (or unchanged reliability), and may provide a suggestion that is accurate (and possibly better than one the user may find on their own). Accurately assessing system confidence in a real world system may be difficult to program. However, researching the effects of system confidence on trust may help determine if it is useful enough for designers of automated systems to pursue.

Providing system confidence information to users has been shown to reduce automation bias because decision makers are better able to assess the validity of the recommendation (Parasuraman & Miller). Providing system confidence may increase the user's initial perceptions of trustworthiness (i.e., dispositional trust, Lee & See, 2004). McGuirl & Sarter (2006) examined the effects of providing system confidence information on trust calibration, compliance, and performance with a DSS that helped pilots with in-flight de-icing procedures. Their definition of system confidence was, "an accurate system-generated prediction of its own accuracy" (p 660). They examined the combination of providing task level system confidence (process information) and historical information (from the past 5 trials). Participants were placed either in a "fixed" condition where only the overall reliability of the system was known (70%) or in an "updated" confidence condition, which provided the system confidence levels for the current task and the previous 4 trials

The number of errors participants made on the primary task was twice as many in the fixed (overall reliability only) condition than in the updated conditions, meaning that trust was better calibrated when participants received system confidence information on a trial by trial basis. Participants were also more likely to misuse the automated system (i.e., comply with the automation when they shouldn't) in the fixed condition, especially in the high task load trials. In addition, providing system confidence information led to less anchoring to the initial advice provided by the automation and more appropriate compliance (e.g., rejecting the advice when it's wrong after checking the information panel for additional information). Thus, automation bias (only seeking information that supports the advice provided by the automation) was reduced when system confidence information was provided. Because participants were able to use process information (i.e., system confidence), rather than performance information (i.e., overall reliability), these results suggest the possibility of a greater influence of dispositional trust rather than history-based trust.

Participants also monitored the panel for equipment failure, which occurred once per session. All equipment failure occurrences were missed across all conditions. The authors interpret this result positively, suggesting that providing system confidence did not pose an additional attention burden. However, another possible interpretation is that if there was a floor effect in both conditions; that is, all occurrences were missed there is no way to determine the level of attentional burden of any of the displays. Thus, it cannot be determined that the display did not require more attention to process the additional information as compared to not providing it. Facilitating more automatic processing of information by providing an additional display could allow participants to devote more remaining attention to the monitoring task. Thus, in this study, attentional demand of displays of system confidence will also be measured.

In a similar study, it was found that system confidence information allowed users to better gauge when to trust or distrust automation, and the anthropomorphic feature of pedigree (i.e., expertise) amplified the effect. The study used a target detection task, also with 4 levels of system confidence (75%, 50%, 25%, and unaided; Spain & Bliss, 2009). The study also added the anthropomorphic feature of expertise to the automation and it was hypothesized that people would trust an automated aid more when they believed the system to be an expert system over a novice system, even though this was only in the instructions, not actually manipulated within the automated system. This human-like classification of knowledge ascribed to the DSS is an example of anthropomorphism or the assignment of human like characteristics to automation. Participants in this study trusted the expert system more than the novice system. As in social interactions, the participants assumed the expert was more trustworthy than the novice system, a perception that outweighed trust built on experience with using the system.

The display in Spain & Bliss (2009) study was different from the McGuirl & Sarter study. Instead of an updating, dynamic line graph display, system confidence was displayed using a bar graph (size of the bar mapped to the confidence level, i.e., 75%

confidence had a bar ¾ filled in), the actual numerical value of the confidence level, and the bar itself was color coded (i.e., the 75% was displayed in a red bar, 50% was displayed with an orange bar, and 25% was displayed with a yellow bar). Compliance was significantly higher in the 75% (high) confidence condition than in the 50% and 25% conditions. Compliance was also greater in the high workload condition. When the image quality was degraded (an example of a data-limited, not resource-limited task), participants were more likely to appropriately comply with the automation when system confidence was high, and rely on their own judgment when system confidence was low.

Limitations of past studies on system confidence

Several problems exist with the McGuirl & Sarter (2004) and Spain & Bliss (2009) studies that may reduce the generalizability to other decision making scenarios. The type of decision tasks that were used differ from the type of decisions a consumer makes (e.g., finding the best prescription drug plan from a list of options). McGuirl & Sarter (2006) used a scenario in which there was not a precisely correct answer (e.g., different combinations of controls could produce a safe situation) and Spain & Bliss (2009) used a binary decision task. Neither studies had any trials in which the automation produced false alarms (i.e., there were never trials that said there wasn't a target when in fact there was).

Another limitation is that the motivations and strategies decision makers rely on may be different for different types of decision tasks. The risk of making an incorrect decision would have resulted in a direct loss of human life in both previous studies (i.e., crashing a plane or not detecting an enemy target), unlike the consequences involved in choosing a product or service (e.g., choosing a drug plan).

Spain & Bliss (2009) may have also confounded system reliability with system confidence. System confidence is not based on the number of trials presented over time like reliability is (i.e., percent accurate). Instead, system confidence is an estimation of how well the system believes the input information for one trial or scenario fits the algorithms that it is basing its decision on. Spain & Bliss describe their manipulation of system confidence as: “Each participant received 24 high confidence, 24 neutral confidence, 24 low confidence trials, and 24 no aid trials. In these four conditions, the base rate of a target being present was .75, .50, .25, and .50 respectively” (p. 345). System confidence in this study was coupled with reliability because the base rate of the target being present was positively correlated with actual performance of the DSS. Whenever the system confidence was high, reliability and thus performance was also high. McGuirl & Sarter’s (2004) manipulation is less clearly defined.

In a perfect system, the system confidence would be equated to how accurate the system is. It would be advantageous to test whether users base their trust on system confidence when confidence is high, but the DSS suggestion is inaccurate or unreliable or vice versa. In the cases where overall reliability is low, yet the system is 75% confident in its suggestion, will users trust the system? Can users differentiate between reliability and system confidence in a situation where there are some “good” answers and one “best” answer? This is a highly likely scenario when there are options that only vary slightly, or in situations where an algorithm isn’t able to distinguish between qualitative differences between options.

When system confidence was low, participants were less likely to comply with the system, thus having to allocate resources to the decision task (Spain & Bliss, 2009; McGuirl & Sarter, 2004). Adding information to an already resource-limited task may not help performance. The display could be designed to minimize the resources needed to process the additional system confidence information. The two studies above only looked at numerical system confidence and a trend line display. It is important to note that in face-to-face interactions with humans (and without DSS), confidence is not assessed numerically.

Reducing processing through display design

There is a body of research that provides insight into how to display numerical information to make it easier to process. For example, bar graphs are superior for tasks that require comparisons or for discriminating discrete differences in dependent variables over different levels of independent variables (Gillan et al., 1988). Line graphs have been shown to help people determine trends and patterns in data over time, whereas pie charts are recommended for showing proportions or percentages. Tables are superior when the task requires an accurate extraction of a single, absolute value (Meyer, Shinar, & Leiser, 1997). Color can be used for identifying levels of a variable (Breslow, Ratwani, & Trafton, 2009). Hybrids of these graphs (“grables”) combine numerical values as well as visualizations through perceptual features such as bar graphs, line graphs, and pie charts (Zacks & Tversky, 1999). Bar grables, line grables, and pie grables have been shown to be superior to any of the perceptual visualizations alone, across a broad range of integration and extraction tasks (Zacks & Tversky, 1999).

One explanation for the beneficial effects of visualizations is that they rely on humans' relatively automaticity at extracting perceptual features (Lohse, 1997; Breslow, Ratwani, & Trafton, 2009). Instead of trying to determine the trend over time or across variables using a table, the integration is displayed via size comparisons (bar and pie charts) or pattern direction (line charts). Tasks that require integration require working memory, a limited resource. The perception of color and size do not require effortful processing when they are used for a task that requires identification (color) or comparisons (size). However, it may be more difficult for decision makers to associate positive and negative valences because that is an artificial mapping of valence to size and color. Context is required to assign valence to these manipulations, which may in itself pose additional workload (e.g., to know if red = "bad" or "good").

Affect As Information Theory and Automatic Processing

Another way to affect one's trust in automation is to replace quantitative information with emotional information that leads to changes in trust behavior. That is, to encode quantitative information into the qualitative presentation of an emotion or affect. According to Slovic, Funicane, Peters, and MacGregor (2003), affect is the "goodness" or "badness" quality experienced as a feeling state (with or without consciousness) and determines the positive or negative qualities of a stimulus. Like perceptual features, affective stimuli are processed without attention and are likely to influence behavior whether the decision maker intends it to or not (Lee, 2006).

Emotional stimuli are thought to be processed automatically but separately from non-emotional stimuli (Ohman & Mineka, 2001), and engage different brain structures (Lee, 2006). Lee (2006; Lee & See, 2004) suggests that affect may play a major role in judgments and decision making due to the pre-attentive nature of affective processing. Evidence of this phenomenon include patients with specific brain lesions that maintain their reasoning abilities and cognitive functions (e.g., working memory), but have impaired emotions and decision making ability (Lee, 2006; Damasio, Tranel, & Damasio, 1990).

Peters, et al. (2009) found that the presence of evaluative categories, or positively and negatively valenced categories in addition to numerical information helped decision makers make better use of the numerical information because they invoked feelings of goodness and badness. Decision makers that score low in numeracy, or the ability to draw meaning from numbers, may benefit the most from affective cues such as positively and negatively valenced labels. In the context of system confidence, it may be difficult for decision makers to associate a numeric value or percentage with a social emotion such as confidence without some form of affective cue. The findings of Peter et al., suggest that the simple manipulation of associating good (positive) and bad (negative) feelings with numbers, such as system confidence level, may be enough to invoke an affective response that leads to better judgments as to whether or not to trust an automated DSS on a trial by trial basis.

One way in which emotions can guide judgments is in the example of trustworthiness. Judgments of trustworthinesss are also processed without attention.

Evidence from fMRI studies have shown different patterns of activation in the brain when judging trustworthiness of a face versus basic emotions of anger, sadness, or fear (Winston, Strange, O'Doherty, & Dolan, 2002). When participants rate the trustworthiness of faces, they reliably rate positive emotions such as happiness, as more trustworthy and negative emotions (e.g., anger, fear, sadness) as less trustworthy (Todorov, 2008). Information derived from facial emotions about trustworthiness can guide avoidance and acceptance behaviors (Todorov, 2008).

Different types of displays may affect trust differentially. In both the McGuirl & Sarter (2004) and Spain & Bliss (2009) studies, additional information regarding the system's confidence was provided in a numerical and graphic display. The additional information may have required more processing and attention. For example, in Spain & Bliss (2009) the color indications could be misconstrued. Red was used to symbolize the danger of a target being present while green was used to symbolize no danger due to the absence of a target. In a task where a target being present presents a positive outcome, the color mappings would need to be switched, such that green means less danger than red. It may require attention to understand these mappings and apply positive and negative valences to the given system confidence number. The same may be true for the size comparisons.

Facial expressions may be a better way to provide information such as system confidence information because users process emotions from facial expressions automatically. Furthermore, anthropomorphic features, or the application of human traits to computers (automation), have been shown to influence trust in situations where the only difference was the added human-like features (e.g., Gong, 2008). For example, adding the face of a smiling doctor to a diabetes decision support system increased trust when compared to a text only condition (Pak, Fink, Price, Bass, & Sturre, 2012). The match between the emotional construct of trust and the mode in which people are well adapted to read confidence and trust – faces, may provide the least attention demanding, yet most influential means of influencing trust.

Encoding System Confidence Values within Faces

Reliably conveying the system confidence through computerized facial displays (i.e., not photographed faces) has been a major hurdle in the design of interactive systems, especially when the use of an avatar is present (e.g., Oh & Stone, 2007; Takeuchi & Nagao, 1993; Walker, Sproull, & Subramani, 1994). These studies show that anthropomorphic features of a conversational agent, such as body language, intonation, and differences facial features are all needed to relay the automation's confidence. In addition to the difficulty of applying self-confidence to a DSS, implementing an interactive, conversational type of system may be impractical for companies to provide to consumers, as well as unnecessary if a more simple solution exists. Instead, relying on valence through the use of static faces with negative and positive emotional expressions may be a much less complicated and less expensive way to relay the same information.

People are well adapted to interpret basic emotions from faces; that is, people reliably and consistently are able to identify emotions such as sadness, happiness, anger, fear, and disgust from facial expressions (Ortony & Turner, 1990). Ekman (1999) theorized that because these emotions are tied to changes in physiology (e.g., differences in heart rate variability, skin conductance, etc.), these emotions help people anticipate future events in an uncertain world where fight or flight responses determine survival. Similarly, the facial expressions that convey trust or confidence may help decision makers anticipate whether they should trust the DSS automatically, using the affect-as-information heuristic when system confidence is displayed via facial expressions.

One of the first direct uses of facial affect to convey quantitative information was Chernoff faces (Chernoff, 1973). Chernoff faces map multi-variate data to specific facial features (i.e., size of the eyes, mouth, or nose, Chernoff, 1973). Further evidence showed that people were sensitive to intensities of different emotions of Chernoff faces (Hess, Blairy, & Kleck, 1997). When the intensity of the emotion is high, people rate that face higher on a scale of emotion. Thus, people should be able to distinguish differences in system confidence based on the type of emotion (e.g., sadness, happiness, or neutral) and intensity (e.g., 25%, 75%).

Will System Confidence Displayed As Facial Expressions Improve Trust Calibration?

The aim of the current studies was to examine whether an anthropomorphic display (i.e., facial expressions) of system confidence can improve trust and trust calibration. In review, our assumptions based on the literature are that 1) emotions (presented by variously valenced facial images) are processed automatically, 2) people can detect fine intensity differences between levels of facial emotion, and 3) presenting system confidence information helps calibrate trust in automation.

If system confidence information can be extracted automatically from facial expressions in the form of emotions and affect (i.e., feelings of goodness or badness), then it should be more influential on trust than a non-affective display. Evidence of less mistrust and distrust would indicate better calibration and faster reaction times would indicate more automatic processing of system confidence information.

A secondary aim of the current study was to examine whether based their trust on the DSS on system confidence or reliability. In previous studies, the relationship between system confidence and actual reliability was positively correlated. Unfortunately, this may not always be true because system confidence is a separate construct based on the information being input to the DSS, not on the actual reliability of the system. This study uncoupled system confidence from reliability and examined whether users base their trust on reliability or on the system confidence information. It was expected that participants would be able to perceive the differences between the coupled and uncoupled system confidence-reliability conditions and only rely on system confidence when it was indeed mapped (coupled) to reliability. However, adding anthropomorphic features to automation has been shown to artificially inflate trust in some studies (e.g., Spain & Bliss, 2009; Gong, 2004), thus it is imperative to test whether adding an anthropomorphic faces display can be perceived correctly or whether this type of display will artificially inflate trust.

Overview of the Current Studies

The purpose of study 1 is to determine the attentional demand of extracting a confidence value from different displays and determined which 1 of the 3 perceptual display conditions (i.e., bar chart, pie chart, or color conditions) was used in study 2. Additionally, results were used to determine if numerical system confidence information should also be included in these conditions. This study did not directly answer the question of whether an anthropomorphic display of system confidence will influence trust. Instead, its purpose was to assess the relative attentional demands of affective displays compared to the other, more conventional data display conditions.

The second study examined the effects of different displays of system confidence and the effects of coupling and uncoupling system confidence with reliability information on trust and compliance with a DSS in the context of choosing a prescription drug plan from a table of 15 options. Trust, accuracy, and reaction time was used to measure trust calibration.

STUDY 1: ATTENTIONAL DEMAND OF TYPES OF DISPLAYS

METHODS

Participants

Thirty younger adults aged 18 to 23 were recruited from the Clemson Human Participants in Research (HPR) system and received course credit for participation. Groups of 1 to 7 participants were tested simultaneously; however participants worked independently at separate workstations. The only exclusion criteria for participation were the presence of color-blindness and the inability to read a computer screen.

Demographic information (e.g., age, gender, education) was collected along with several computerized ability tests measuring perceptual speed (Digit Symbol Substitution; Weschler, 1997), working memory (Reverse Digit Span; Weschler, 1997); spatial orientation (Cube Comparison; Ekstrom, French, Harmon, & Dermen, 1976), spatial visualization (Paper Folding; Ekstrom, et al., 1976), and crystallized intelligence (Shipley Vocabulary Test; Shipley, 1986). These tests were used to identify any participants whose performance may be abnormal due to sub-average abilities, thus potentially affecting the accuracy of results of the between groups variable of display type.

Design

This study was an 11 (system confidence display type) x 2 (task load: single, dual) within subjects design. Three levels of system confidence were tested (low, 25%; neutral, 50%; and high, 75%). System confidence display type included six displays of system confidence (i.e., number, bar, pie, color, evaluative categories (eval-cats), and anthropomorphic faces (anthro-face)). For all levels except the number condition, the image was shown with and without the numerical system confidence value (e.g., the bar graph condition will be shown with the 25% value and without 25%), resulting in 5 displays without numerical information (i.e., bar, pie, color, evaluative categories, and anthro-face) and 6 displays with numerical information (i.e., number, bar, pie, color, evaluative categories, and anthro-face). The purpose of this manipulation was to determine if the perceptual features of the display (i.e., graph, size, or color) require less attention to identify system confidence level. Attention was measured by participants' response time on the graphic identification task under high workload conditions.

Procedure

Participants performed both a primary task (graphic level identification with the system confidence displays) and secondary task (playing a block game similar to the game Tetris). Experimenters gave participants a paper copy of each display condition (see Appendix A) that displayed all levels of confidence before moving on to the actual experiment. Participants first performed the block game task until they reached a score of 50. Next, they performed the graphic level identification task. Finally, they performed the tasks together. Participants were told to prioritize the block game task and perform the graphic level identification task with any reserve attention. The purpose was to make sure that participants were engaged in the block game task throughout the experiment. The full protocol can be found in Appendix B.

Tasks

Graphic identification task (primary task)

The primary task was to identify the level of the system confidence of the graphic displayed on the computer screen. Each system confidence display graphic had 3 levels: 25%, 50%, and 75%. These graphics were displayed on the right hand side of the computer screen in a designated box with a resolution of 340 x 410 pixels. Participants rated the images using the numeric keys with the following mapping: 1 =25%, 2= 50%, and 3=75%.

Block matching task (secondary task)

The block matching task was the secondary task because it was designed to place a constant attentional burden in the dual-task condition (Fisk, Derrick, & Schneider, 1986). Participants used the arrow keys (i.e., up, down, left, right) to match 3 blocks vertically or horizontally to gain points. Blocks could be switched horizontally (but not vertically) using the space bar. When 3 blocks of the same color were matched, they disappeared, similar to the game of Tetris. Blocks moved at a rate of 1 pixel every 100 ms. The goal of this task was to keep participants engaged in a secondary task that requires continuous attention and that would be difficult to automatized (Fisk, Derrick, & Schneider, 1986).

Participants completed a total of 66 trials**.** Two trials for each level of system confidence (3: 25%, 50%, 75%), in each of the 11 display type conditions (number only + 6 display types a(x2: with number/without number). Trials were displayed in random order and at random time intervals between 30 and 40 seconds to prevent subjects from anticipating the image in the dual task condition. It was not feasible to increase the number of trials due to the high number of displays being tested. Displaying the images twice at intervals of 30-40 seconds requires 45 minutes. Participants will also do this task twice, once in the single task condition and again in the dual task condition. Thus the time required to complete both of these tasks together (and not including the block matching task or the demographics) is 1hr 30min.

Independent Variables

System confidence displays and level of confidence

Six levels of system confidence displays were assessed in this first study. Each one is described in more detail and pictured below. The same 3 levels of system confidence were used as in the Spain & Bliss (2009) study: 25%, 50%, and 75%. These levels are equally spaced apart (25%) and the disparity was large enough to discriminate between levels. The 50% condition was chosen as the neutral point because it indicates that the system is equally confident as it is unconfident.

Condition 1: Numerical percentage baseline condition (Number)

In this condition, just the numerical display of system confidence was presented (see Figure 3 for the display at all 3 levels of system confidence). Tasks that require extraction of specific numerical quantities are best supported with a table to prevent misinterpretation of absolute values (Hink, Wogalter, & Eustace, 1996; Tufte, 1983). The size of the display is 134 x 70 pixels.

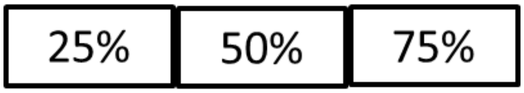


Figure 3. Number Condition.

Condition 2: Bar chart display (Bar and nBar)

The bar graph display was chosen because it provides a perceptual comparison of area that may be associated with the proportion (i.e., percentage) of system confidence. Bar graphs are typically most useful when displaying differences between some dependent variable over levels of an independent variable (Gillan, et al., 1998). There is only one level being shown at each time, rather than a discrete comparison between data points, which is what people are typically used to associating with bar graphs (Zacks, & Tversky, 1999). It is unclear whether the area comparison within one bar display will be discernible enough to obtain an absolute value of confidence, without the number. As a comparison, a second condition (see Figure 4 for both displays) that includes both the scaled gray bar with appropriate proportional area covered (i.e., 25%, 50%, & 75%) will be shown with the number.

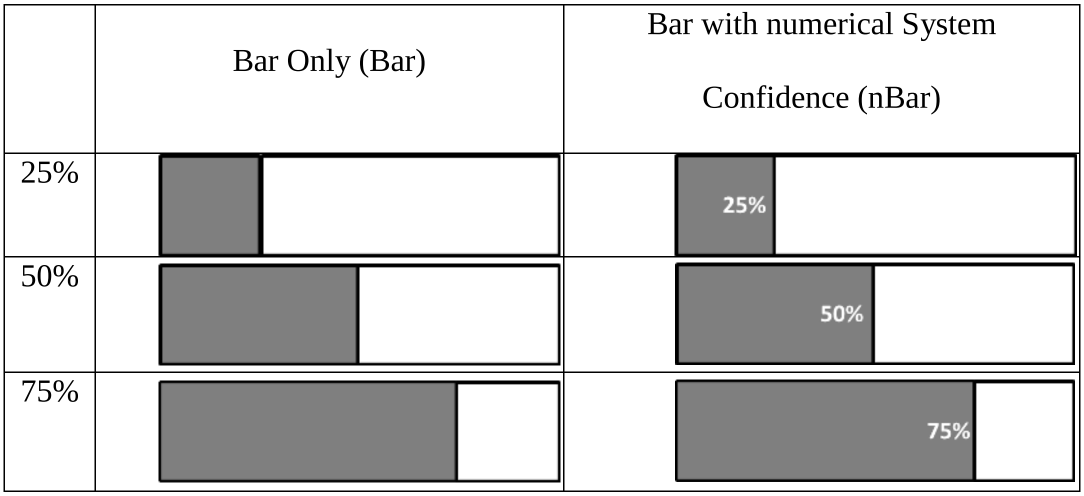


Figure 4. Bar and Numerical Bar Conditions.

Condition 3: Pie Chart Display (Pie and nPie)

A pie chart condition was chosen because performance has been shown to improve when a pie chart is used to represent proportionate values (e.g., percentages) (Gillan et al., 1998). Similar to the condition above, including the numerical quantity allows a comparison between conditions to examine if there is a significant difference when the number is included with the perceptual comparison. Previous research has also suggested that hybrid displays of pie charts that include the table values (i.e., a pie "grable") produces significantly more accurate results than just the pie chart itself when the goal is to extract an absolute value (Hink, Wogalter, & Eustace, 1996). The pie charts shown are smaller in Figure 5 than the actual size displayed (~274x268 pixels).

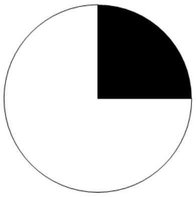
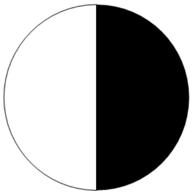
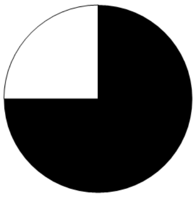
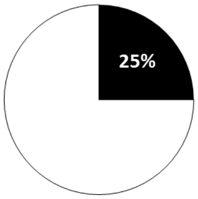
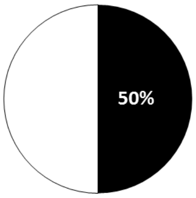
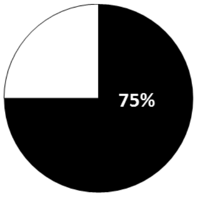
	25%	50%	75%
Pie Chart			
Pie Chart with numerical system confidence (nPie)			

Figure 5. Pie and Numerical Pie conditions.

Condition 4: Color Display (Color and nColor)

The color display condition uses a multi-colored heat map scale to indicate the level of confidence (Figure 6). Breslow, Ratwani, & Trafton (2009) found that when the task requires the identification or extraction of an absolute value, using a multi-colored scale can facilitate identification. The task used in that study involved finding the color in a legend, then finding a region on a map. The task in this study requires the identification of a proportionate confidence level. Thus, this condition was included to examine whether the benefits of a multi-colored heat map can be extended to an extraction task where the subject knows that they only have 3 options to rate the level of confidence.

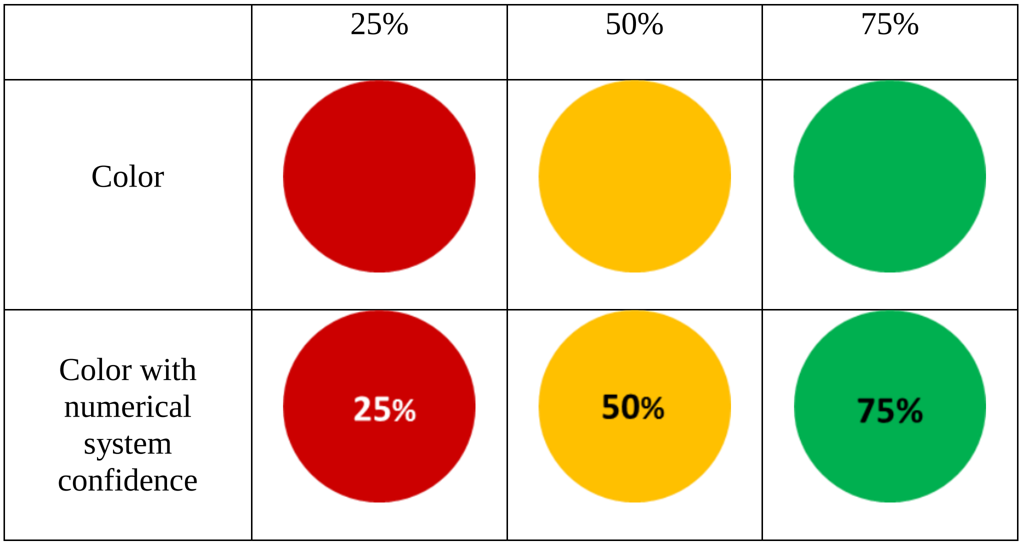


Figure 6. Color and Numerical Color Conditions.

Condition 5: Evaluative Categories Display (EvalCat and nEvalCat)

Evaluative categories can add meaning to numerical data by invoking an affective response that can improve decision making, especially for those low in numeracy (Peters, et al., 2009). Three labels were created with boundary lines similar to the Peters, et al. (2009) study to help participants decide whether to trust the system (see Figure 7).

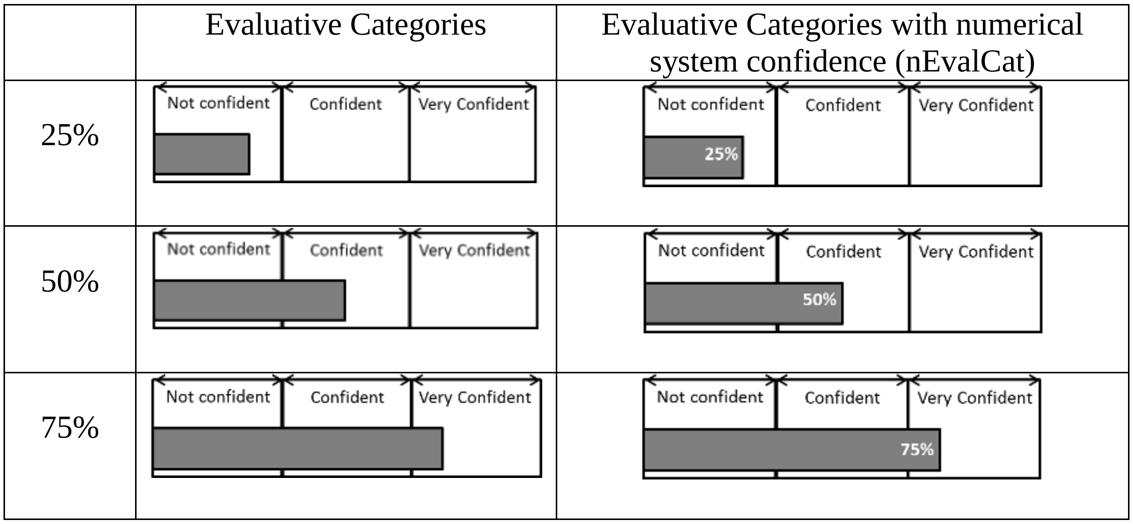


Figure 7. Evaluative Categories and Numerical Evaluative Categories Conditions.

Condition 6: Anthropomorphic Display (Anthro and nAnthro)

Chernoff faces were created using the statistical package R (Figure 8). A wide range of emotion was chosen to maximize the discriminability between confidence conditions. For the 25% condition, the emotion of sad was chosen because it is one of the six emotions that people are able to best recognize and because it should evoke a negatively valenced affect. The 50% or neutral condition is the same for both happy and sad, so this was chosen as the midpoint. The 75% happy and 100% happy faces were chosen to represent high system confidence.

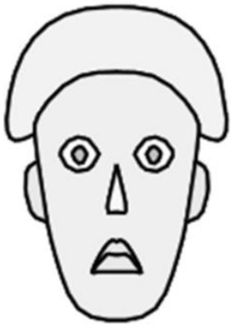
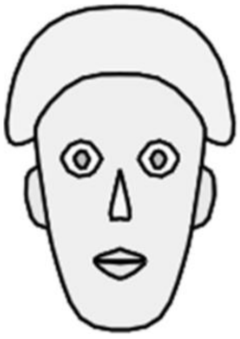
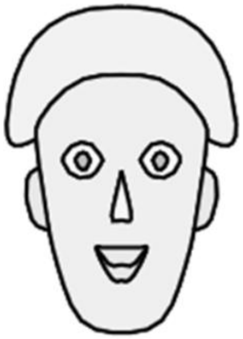
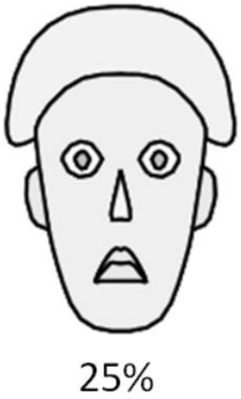
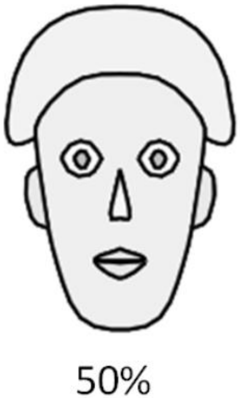
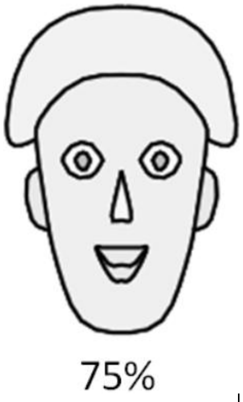
	25%	50%	75%
Anthropomorphic			
Anthropomorphic with numerical system confidence display	 	 	

Figure 8. Anthropomorphic display and Numerical Anthropomorphic Conditions.

Dependent variables

Graphic identification accuracy

Accuracy on the graph identification task was a binary measure, either correct or incorrect. Correct scores indicate that participants pressed the correct number key associated with the image's confidence level (where1=25%, 2=50%, & 3=75%).

Graphic level identification time

Graphic identification time was recorded in milliseconds from the time the graphic appeared to the time the participant pressed the number key to make a rating. Faster reaction times indicate lesser attentional demands.

Block game score

A block game score was assessed by the program automatically by recording 1 point when 3 blocks are successfully matched and cleared. A total score was then calculated. The score was used to screen out participants who were not engaged in both tasks, defined as scores below 3 standard deviations from the mean.

RESULTS

Thirty participants (15 female, 15 male) between the ages of 19-27 (*M*=23.43, *SD*=2.74) participated in study 1. Participant characteristics are included in Table 1. Unfortunately, scores for the spatial visualization (paper folding) and spatial orientation (cube comparison) were not recorded for 3 subjects, thus only 27 subjects are included in the reported means and standard deviations for those abilities tests.

One subject was dropped from the performance analyses because of computer malfunction. The remaining 29 subjects were included in the remainder of the analyses.

Table 1.
Study 1 Participant Descriptives

Category	*M*	*SD*
Age[1]	23.43	2.74
Spatial Visualization (Paper Folding)[2]	23.04	2.59
Spatial Orientation (Cube Comparison)[2]	24.56	9.29
Working Memory (Reverse Digit Span)[1]	7.73	2.8
Perceptual Speed (Digit Symbol Substitution)[1]	96%	8.9%
Vocabulary (Shipley)[1]	29.43	4.28

Note: [1]N=29; [2]N=27.

Block Game Task Score: A Manipulation Check of Attentional Load

The predetermined criterion for exclusion from analyses was a score below 3 standard deviations from the mean on the block game task. It could be assumed that these participants were not following instruction or were not devoting attention to the task. The mean block game score in the dual task condition was a score of 88.7 (blocks cleared) and the standard deviation was 22.06, and thus the criterion for exclusion was a score of 22.52. The lowest score on this task was 49, well above the exclusion criterion so all remaining 29 subjects were included in the analyses.

Attentional Demand of System Confidence Displays

An 11 (system confidence display type) x 2 (single task vs. dual task) within subjects ANOVA was used to analyze task time and accuracy on the graphic identification task. The purpose of this analysis was to identify the graphic display (e.g., Bar, Pie, and Color) that required the least amount of time (as a measure of attention) to identify the system confidence level. The condition that required the least amount of attention was then used in study 2. The purpose of including the other displays was to provide relative attentional differences between conditions as a possible explanation for performance differences in study 2. For example, if differences are found in attentional load between displays, these results may help explain performance differences in study 2.

Task time

For task time, there were significant main effects of task type ($F(10, 24)=50.72$, p<.000, η_p^2 =.68) and display type ($F(10, 240)=6.77$, p<.000, η_p^2=.22). Task type (single vs. dual task) also significantly interacted with the display type on task time ($F(10, 240)=3.04$, p=.001, η_p^2=.11). Task time was measured in seconds, thus all means and standard deviations are presented in seconds.

A Post-hoc Bonferroni analysis revealed 13 significant differences between the 11 display conditions in the single task condition and these are summarized in Table 2 and presented in Figure 9. The purpose of study 1 was to determine which of the graph conditions should be used in study 2, thus these are the only differences that will be discussed in-text because of the high number of differences found in study 1.

Table 2.
Significant Post-hoc Comparisons for Task Time (seconds)

Display Type	*M*	*SD*	*Direction*	Display Type	*M*	*SD*	*p*
			Single Task (n=29)				
Bar	1.04	0.34	Faster	Color	1.52	0.42	.007
nBar	1.03	0.32	Faster	Anthro	1.40	0.43	.001
nPie	1.09	0.49	Faster	Color	1.52	0.42	.005
nPie	1.09	0.49	Faster	Anthro	1.40	0.43	.045
nPie	1.09	0.49	Faster	NAnthro	1.23	0.50	.013
Color	1.52	0.42	Faster	Number	1.12	0.48	.001
EvalCat	1.12	0.35	Faster	Color	1.52	0.42	.000
EvalCat	1.12	0.35	Faster	Pie	1.32	0.43	.004
EvalCat	1.12	0.35	Faster	Anthro	1.40	0.43	.000
nEvalCat	1.06	0.34	Faster	Color	1.52	0.42	.000
nEvalCat	1.06	0.34	Faster	Pie	1.32	0.43	.006
nEvalCat	1.06	0.34	Faster	Anthro	1.40	0.43	.000
nEvalCat	1.06	0.34	Faster	NAnthro	1.23	0.50	.050
			Dual Task (n=29)				
nPie	1.08	0.10	Faster	Pie	1.98	0.08	.003
nPie	1.60	0.08	Faster	nEvalCat	2.01	0.11	.018
nPie	1.60	0.08	Faster	Anthro	1.90	0.09	.013

For the single task condition, the Color condition (*M*=1.52, *SD*=.42) was significantly slower than the Bar condition (*M*=1.12, *SD*=.45, *p*=.007), nBar condition (*M*=1.03, *SD*=.32, p<.000), and nPie condition (*M*=1.09, *SD*=.49, *p*=.005). In the dual task condition, nPie was faster (*M*=1.6, *SD*=419.8933) than Pie (*M*=1.98, *SD*=.08), p=.003. The nPie condition was also the only graph display that was significantly faster than affective conditions, nEvalCat (*M*=2.01, *SD*=.11, *p*=.018) and Anthro (*M*=1.9, *SD*=.09, *p*=.013).

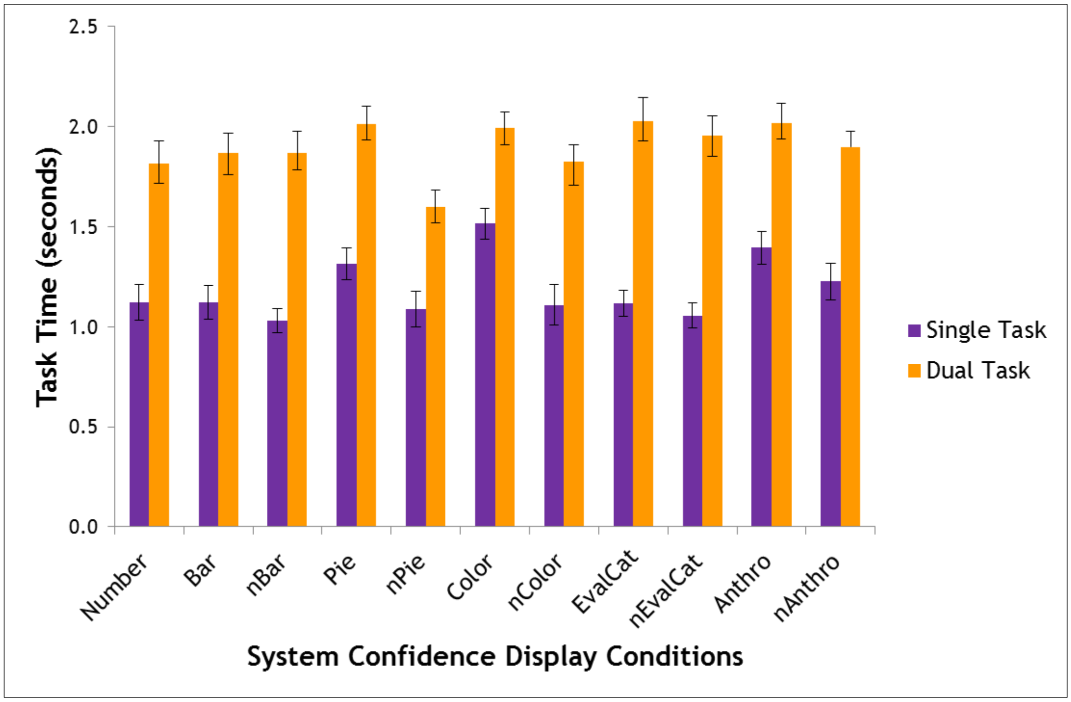


Figure 9. Mean task time by system confidence display condition (in seconds). Error bars represent standard error of the mean, N=29.

Accuracy

Mauchly's test indicated that the assumption of sphericity had been violated for display type (*χ2* (54) = 91.76, *p* < .001) and the interaction between task type and display type (*χ2* (54) = 78.86, *p* < .019). Degrees of freedom were corrected using Huynh-Feldt estimates of sphericity for these two analyses (display type, *ε* = .822; Task Type by Display Type, *ε* = .894) (Huynh & Feldt, 1976). Task type did not violate the assumption of sphericity, thus the sphericity assumed values are reported.

There were significant main effects of task type (*F*(1, 24)=28.41, p<.000, η_p^2 =.54) and display type (*F*(8.22, 197.26)=4.82, p<.000, η_p^2 =.17). Additionally, there was a significant interaction between task type and display type (*F*(8.94, 214.56)=2.12, p=.03, η_p^2 =.08). A Post hoc Bonferroni analysis indicated that there were no differences between single and dual task accuracy in the Pie condition (single – *M*=5.64, *SD*=.17; dual –*M*=4.56, *SD*=.27), Color condition (single – *M*=5.08, *SD*=.29; dual –*M*=4.52, *SD*=.25), and nColor condition (single – *M*=5.64, *SD*=.18; dual –*M*=5.36, *SD*=.18). The remaining 8 display conditions showed a significant reduction in accuracy with the addition of the block task compared to the single task condition (p<.05). These results are graphed in Figure 10.

In the single task conditions, the Number display (*M*=5.86, *SD*=.44) and EvalCat display (*M*=5.83, *SD*=.76) were significantly more accurate than the Pie display condition (*M*=5.21, *SD*=1.01). However, in the dual task conditions the nColor display (*M*=4.79, *SD*=1.68) was significantly more accurate than the Bar display (*M*=4.07, *SD*=1.88), Anthro display (*M*=4.26, *SD*=1.39), and Color display (*M*=4.14, *SD*=1.58).

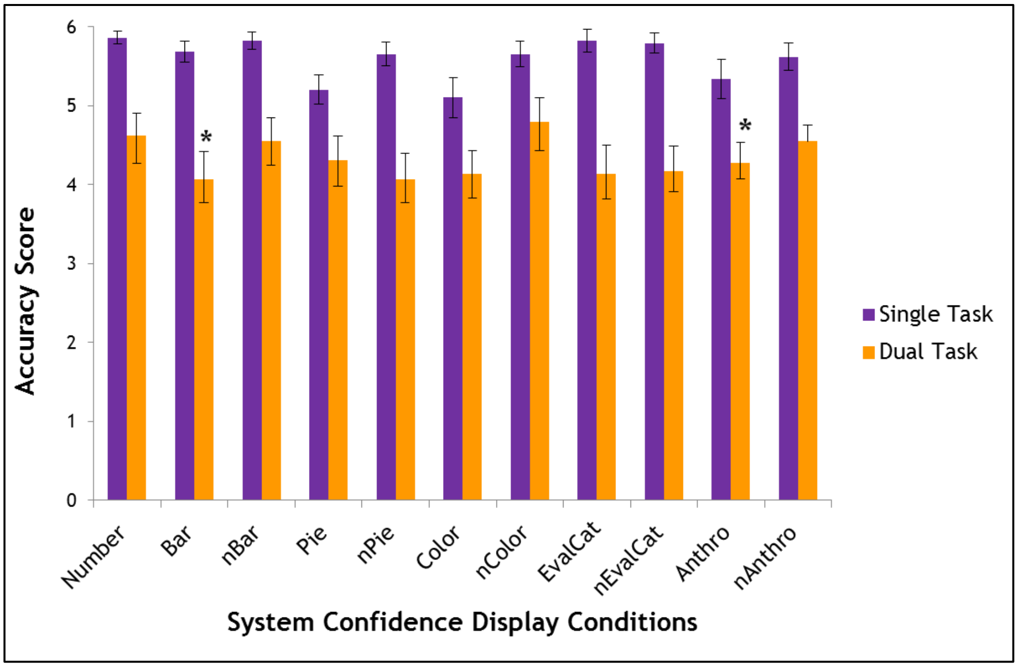


Figure 10. Mean accuracy score by system confidence display condition. Error bars represent standard error of the mean, N=29.

DISCUSSION

The purpose of study 1 was to identify which of the graph conditions (i.e., Pie, nPie, Color, nColor, Bar, or nBar) commanded the least amount of attention under high workload and should be brought forward in study 2. The remaining conditions were included to determine if there were major differences in including or excluding the numerical value of system confidence (e.g., 25%) in the conditions that this could be excluded. Additional information may require more attention and thus reduce performance-based measures such as task time. Attentional load was determined using task response time, however accuracy was also measured.

In the dual task condition, there were no significant differences in response time between the graph conditions. The task itself was relatively simple and did not require the user to comprehend meaning of the system confidence value, just rate it between 1, 2, or 3. However, in the high workload condition, nPie display was faster than the affective conditions nEvalCat and Anthro. Interestingly, when the number was not included with the Pie display accuracy was unaffected by the addition of the block matching task. The combination of quicker response times and lack of diminishing accuracy as workload increased determined that the nPie condition would be brought forward.

Ironically performance did not improve when numerical confidence rating was included in the graphic. This finding indicates that the additional information may not be necessary in determining the value of system confidence; the graphic alone was enough to convey numerical meaning. Peters, et al. (2009) found that evaluative categories improved performance when used in conjunction with a number and a graph and the results of this study gave no indication that the display should be minimized by not displaying the numerical system confidence value. Thus, the nEvalCat display was used in study 2, as opposed to the EvalCat display (without the number). The Anthro display was used in study 2 without the numerical information because the results of study 1 indicate that participants are able to distinguish differences between the faces without the numerical value. The numerical values were not included because it would be more difficult to conclude from the results that the processing of affect from the face display is influencing trust or if participants are also using the numerical value. The idea was that completely separating the two – numerical information from affective information – would better test the hypothesis that affective information would have a greater influence on trust behavior.

STUDY 2: EFFECTS OF TYPE OF SYSTEM CONFIDENCE DISPLAY ON TRUST AND PERFORMANCE

The purpose of study 2 was to examine the effect of an anthropomorphic face display of system confidence on trust behavior compared to more traditional displays. In study 2, the effects of 5 different types of system confidence displays on trust and performance with a DSS was examined in a prescription drug plan decision context. Three levels of system confidence were examined, low (25%), neutral (50%), and high (75%), similar to previous studies examining system confidence (i.e., McGuirl & Sarter, 2006; Spain & Bliss, 2009). An additional focus of this study was to answer the question of whether participants will use system confidence information in the same manner when it is not positively correlated with reliability. If no differences in trust are found when system confidence is coupled versus uncoupled with reliability, it may indicate that the additional system confidence information is not beneficial over providing reliability information.

METHODS

Participants

One hundred younger adults ages 18 through 27 were recruited through flyers, advertisements, and through the Clemson University Sona Systems Participant Pool website. Participants were able to choose from one of the following incentives for participation: 1) enter a drawing for an 8GB iPod Nano, 2) earn $10, or 3) extra credit in a psychology course. The only exclusion criterion was the presence of color blindness (self-reported).

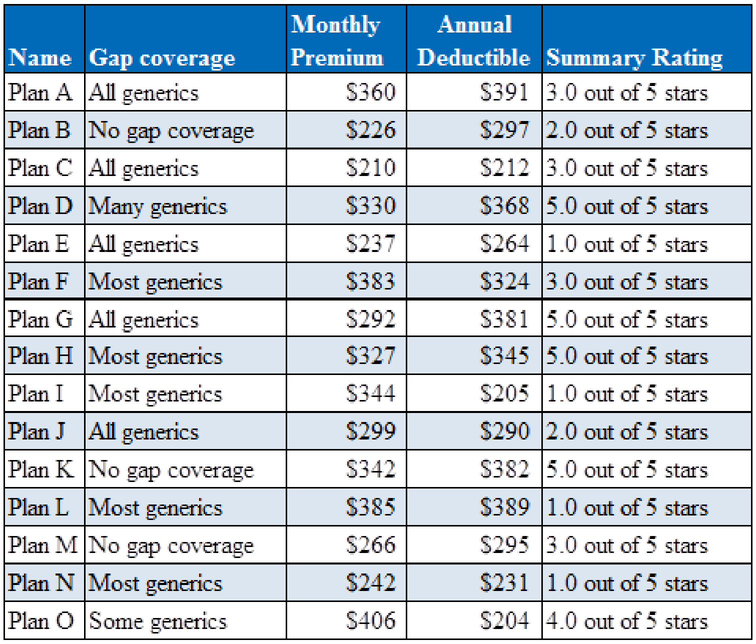
Design

The study was a 4 +1 (system confidence display: none (control), number, nPie (from study 1), evaluative categories, anthropomorphic) x 3 (system confidence level: 25%, 50%, & 75%) x 2 (reliability-confidence relationship: coupled, uncoupled) mixed factorial design. System confidence display type was a between subjects variable, system confidence level and reliability-confidence relationship were within subjects variables.

Each participant completed 60 trials of a decision-making task over two blocks that were counterbalanced between participants. One block contained 30 trials with uncoupled reliability-system confidence, with 10 trials at each of the 3 levels of system confidence. The other block contained 30 trials with coupled reliability-system confidence, also with 10 trials at all 3 levels of system confidence. In a previous study (Pak, et al., 2012), the automated aid did not fail for the first 8 trials, which was sufficient for participants to build trust in the system (i.e., not immediately discount its advice). Aid failures were placed randomly on the remaining 22 trials. Similar to previous findings (Sanchez, 2006; Pak, et al., 2012), trust did not diminish after failure and instead recovered with the first aid success.

Decision Task

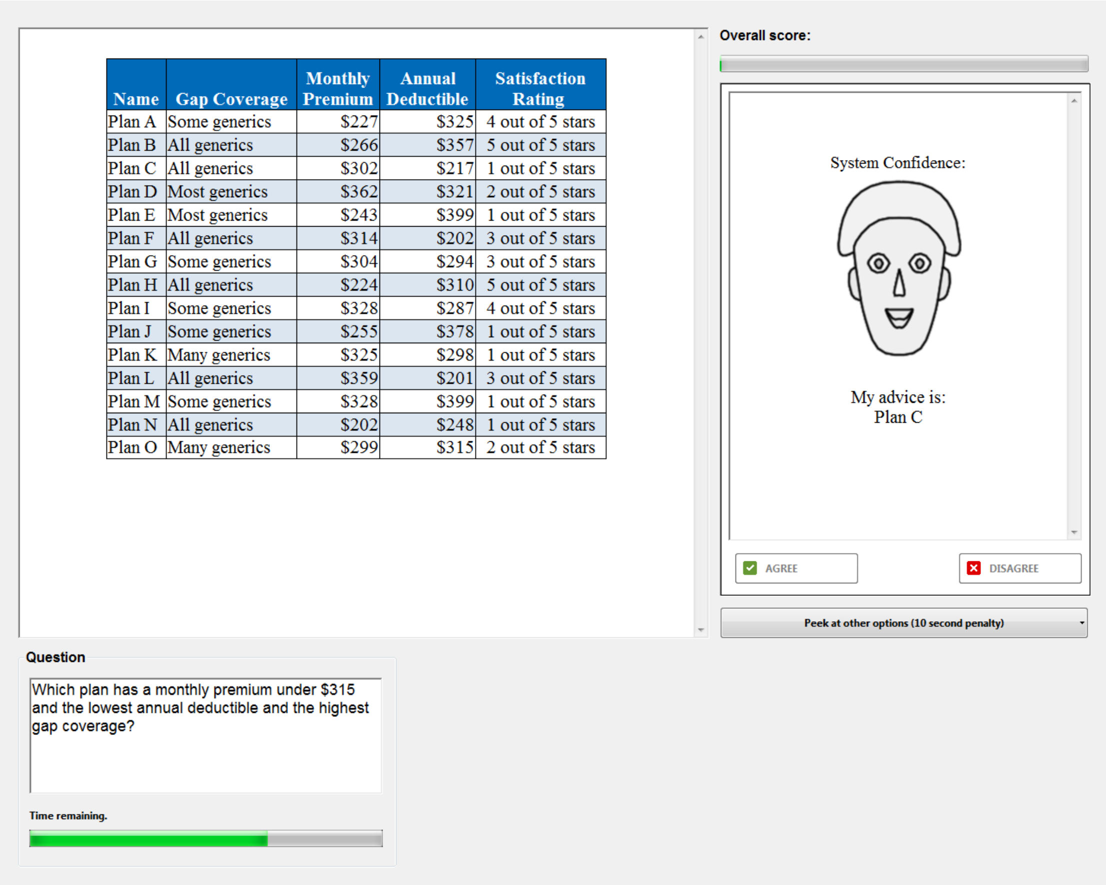
Participants had to choose a prescription drug plan from a table with 15 plan options and 4 attributes (see Figure 11). The table with the plan data had the same 4 attributes (i.e., monthy premium, annual deductible, gap coverage, and satisfaction rating) used in Price & Pak (accepted). Additionally, the question contained criteria that the plan had to meet. For example, one question read, "Which plan has the most gap coverage, a monthly premium under $325, and the lowest annual deductible?" One plan option met all of the criteria, 5 plans will met 2 of the 3 criteria, 5 plans met 1 out of the 3 criteria and 4 met none of the criteria. The table and attributes were explained to participants at the beginning of the experiment, and they will be given a sheet of paper which the definitions of each of the attributes (e.g., gap coverage, summary rating, see Appendix C) to refer back to during the experiment. These were also read aloud by the experimenter and participants were given the opportunity to ask questions specifically about these terms before moving on. This method of explaining the task was identical to that of the table condition in Price & Pak (accepted).



Name	Gap coverage	Monthly Premium	Annual Deductible	Summary Rating
Plan A	All generics	$360	$391	3.0 out of 5 stars
Plan B	No gap coverage	$226	$297	2.0 out of 5 stars
Plan C	All generics	$210	$212	3.0 out of 5 stars
Plan D	Many generics	$330	$368	5.0 out of 5 stars
Plan E	All generics	$237	$264	1.0 out of 5 stars
Plan F	Most generics	$383	$324	3.0 out of 5 stars
Plan G	All generics	$292	$381	5.0 out of 5 stars
Plan H	Most generics	$327	$345	5.0 out of 5 stars
Plan I	Most generics	$344	$205	1.0 out of 5 stars
Plan J	All generics	$299	$290	2.0 out of 5 stars
Plan K	No gap coverage	$342	$382	5.0 out of 5 stars
Plan L	Most generics	$385	$389	1.0 out of 5 stars
Plan M	No gap coverage	$266	$295	3.0 out of 5 stars
Plan N	Most generics	$242	$231	1.0 out of 5 stars
Plan O	Some generics	$406	$204	4.0 out of 5 stars

Figure 11. Example of plan data.

Figure 12 shows a screen shot of the program that was used for the decision task**.** In all conditions, participants were shown a screen with the plan table on the left, the question below the plan table, and a timer bar that counted down 45 seconds. The DSS suggested plan was be displayed on the right of the screen, with the options to agree, disagree, or peek below it. The plan suggestion was presented as “Plan X”, in an effort to avoid confounding features of language that may influence trust. The system confidence display was on the right above the DSS’ suggested plan in all conditions except for the no system confidence condition. An overall score bar was placed above the system confidence display (on the top right) which increased as participants answered questions both quickly and correctly.



Name	Gap Coverage	Monthly Premium	Annual Deductible	Satisfaction Rating
Plan A	Some generics	$227	$325	4 out of 5 stars
Plan B	All generics	$266	$357	5 out of 5 stars
Plan C	All generics	$302	$217	1 out of 5 stars
Plan D	Most generics	$362	$321	2 out of 5 stars
Plan E	Most generics	$243	$399	1 out of 5 stars
Plan F	All generics	$314	$202	3 out of 5 stars
Plan G	Some generics	$304	$294	3 out of 5 stars
Plan H	All generics	$224	$310	5 out of 5 stars
Plan I	Some generics	$328	$287	4 out of 5 stars
Plan J	Some generics	$255	$378	1 out of 5 stars
Plan K	Many generics	$325	$298	1 out of 5 stars
Plan L	All generics	$359	$201	3 out of 5 stars
Plan M	Some generics	$328	$399	1 out of 5 stars
Plan N	All generics	$202	$248	1 out of 5 stars
Plan O	Many generics	$299	$315	2 out of 5 stars

Question

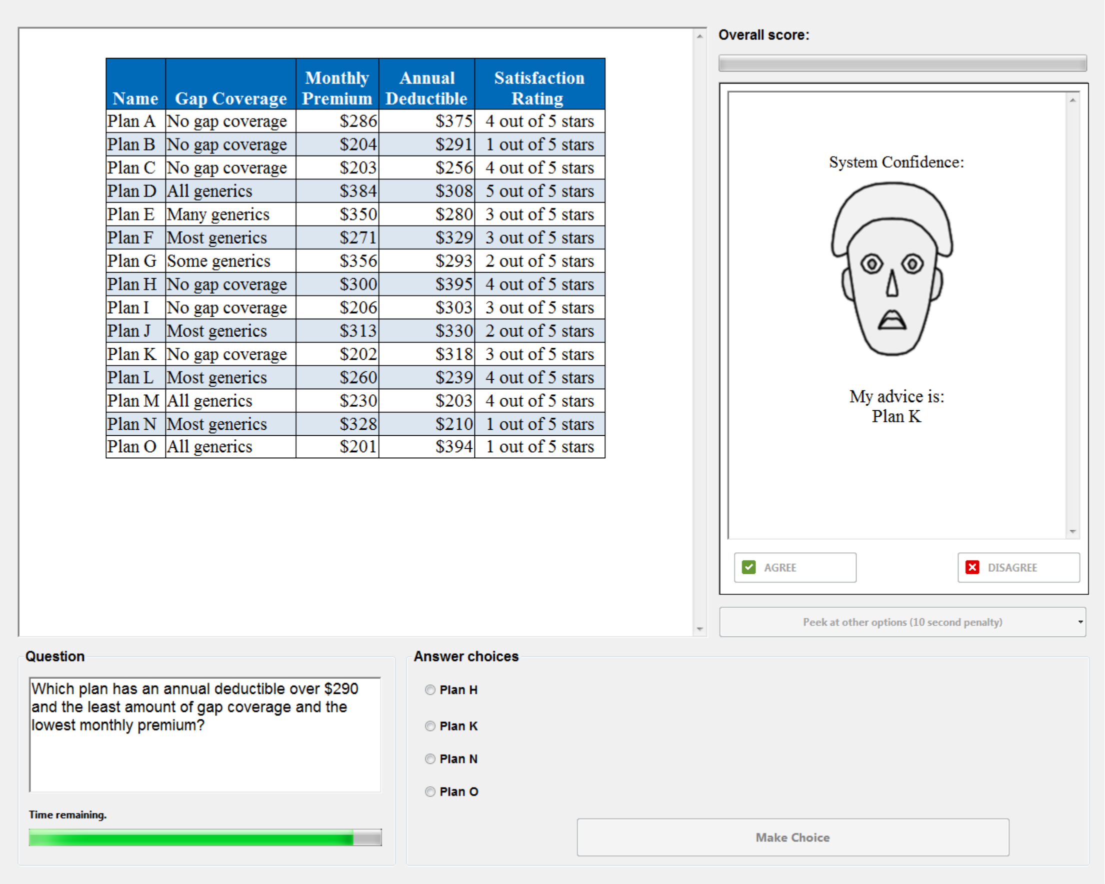
Which plan has a monthly premium under $315 and the lowest annual deductible and the highest gap coverage?

System Confidence:

My advice is:
Plan C

Figure 12. Experiment screen in the anthropomorphic faces condition. The plan table is in the upper left, with the question below it, and the timer bar at the bottom left. The DSS system confidence display is on the upper right, with the DSS' suggested plan below it. Options to agree, disagree, or peek are on the lower right. The overall score is in the upper right.

The participant could choose to agree, disagree, or peek at other options. If the participant disagreed, the participant had to choose 1 of 4 possible answers (see Figure 13). Participants were told that the 4 answers are other possible answers that could be correct. In pilot testing and a previous study (Price & Pak, accepted), participants did not have sufficient time to do the task on their own, so providing all answers would make the task impossible. Furthermore, the main focus of this study is on trust of automation, so once the user decided to disagree or peek and disagree, whether they solved the problem accurately on their own was not of particular interest in this study.



Name	Gap Coverage	Monthly Premium	Annual Deductible	Satisfaction Rating
Plan A	No gap coverage	$286	$375	4 out of 5 stars
Plan B	No gap coverage	$204	$291	1 out of 5 stars
Plan C	No gap coverage	$203	$256	4 out of 5 stars
Plan D	All generics	$384	$308	5 out of 5 stars
Plan E	Many generics	$350	$280	3 out of 5 stars
Plan F	Most generics	$271	$329	3 out of 5 stars
Plan G	Some generics	$356	$293	2 out of 5 stars
Plan H	No gap coverage	$300	$395	4 out of 5 stars
Plan I	No gap coverage	$206	$303	3 out of 5 stars
Plan J	Most generics	$313	$330	2 out of 5 stars
Plan K	No gap coverage	$202	$318	3 out of 5 stars
Plan L	Most generics	$260	$239	4 out of 5 stars
Plan M	All generics	$230	$203	4 out of 5 stars
Plan N	Most generics	$328	$210	1 out of 5 stars
Plan O	All generics	$201	$394	1 out of 5 stars

Overall score:

System Confidence:

My advice is:
Plan K

Question

Which plan has an annual deductible over $290 and the least amount of gap coverage and the lowest monthly premium?

Time remaining.

Answer choices

○ Plan H

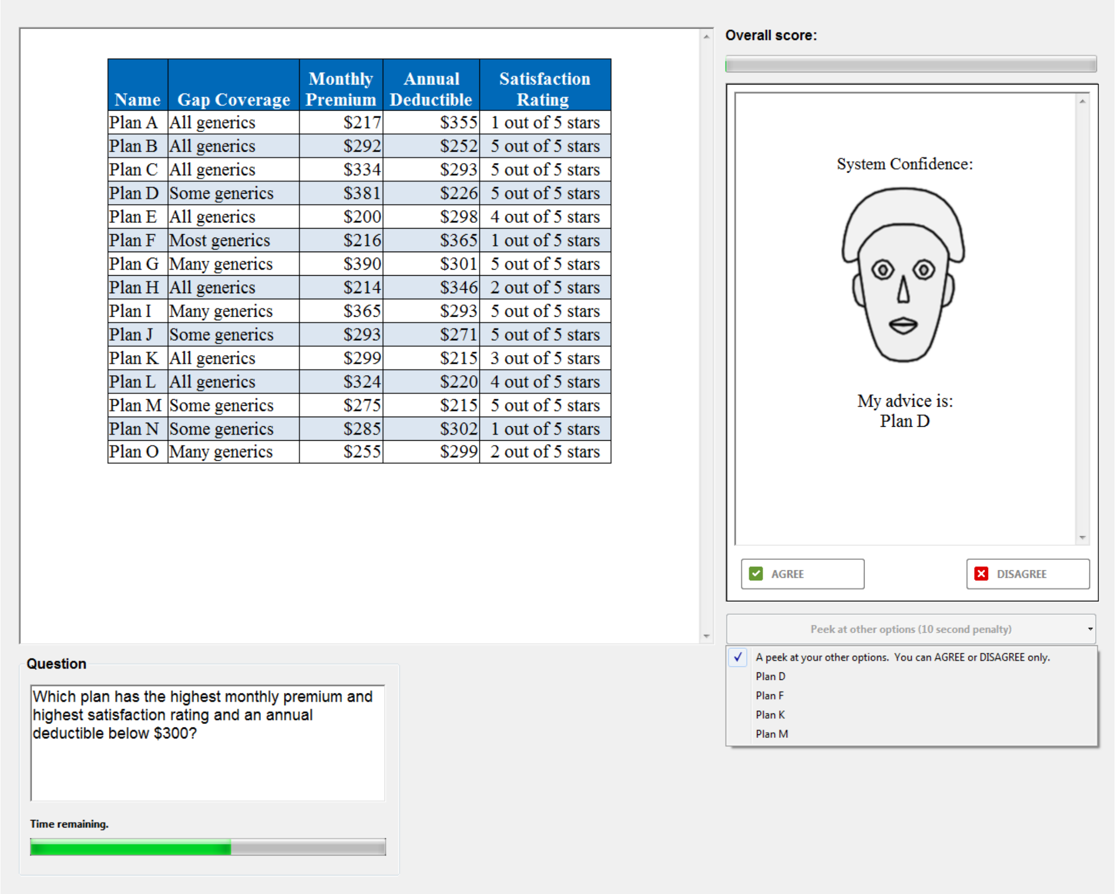
○ Plan K

○ Plan N

○ Plan O

Figure 13. Experiment screen when the user chooses “Disagree”.

The same 4 answers were displayed if the user clicked the peek button but were not selectable (see Figure 14). The participant still had to choose agree or disagree after peeking. After the participant selected an answer, they rated their trust in the DSS and their confidence in their answer (Figure 15), and then received feedback on whether their answer was correct or incorrect (see Figure 16).



Name	Gap Coverage	Monthly Premium	Annual Deductible	Satisfaction Rating
Plan A	All generics	$217	$355	1 out of 5 stars
Plan B	All generics	$292	$252	5 out of 5 stars
Plan C	All generics	$334	$293	5 out of 5 stars
Plan D	Some generics	$381	$226	5 out of 5 stars
Plan E	All generics	$200	$298	4 out of 5 stars
Plan F	Most generics	$216	$365	1 out of 5 stars
Plan G	Many generics	$390	$301	5 out of 5 stars
Plan H	All generics	$214	$346	2 out of 5 stars
Plan I	Many generics	$365	$293	5 out of 5 stars
Plan J	Some generics	$293	$271	5 out of 5 stars
Plan K	All generics	$299	$215	3 out of 5 stars
Plan L	All generics	$324	$220	4 out of 5 stars
Plan M	Some generics	$275	$215	5 out of 5 stars
Plan N	Some generics	$285	$302	1 out of 5 stars
Plan O	Many generics	$255	$299	2 out of 5 stars

Question

Which plan has the highest monthly premium and highest satisfaction rating and an annual deductible below $300?

Time remaining.

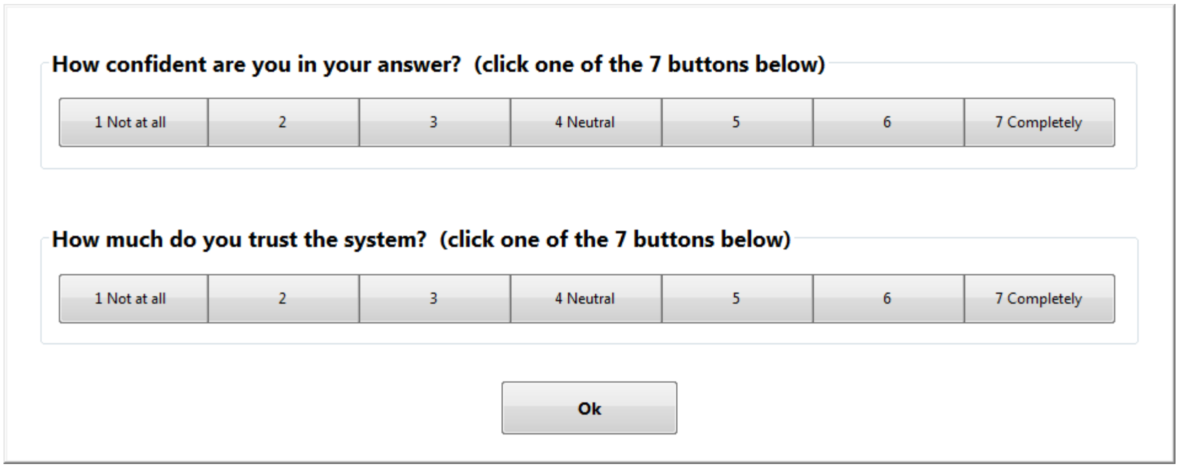
Overall score:

System Confidence:

My advice is:
Plan D

✓ A peek at your other options. You can AGREE or DISAGREE only.

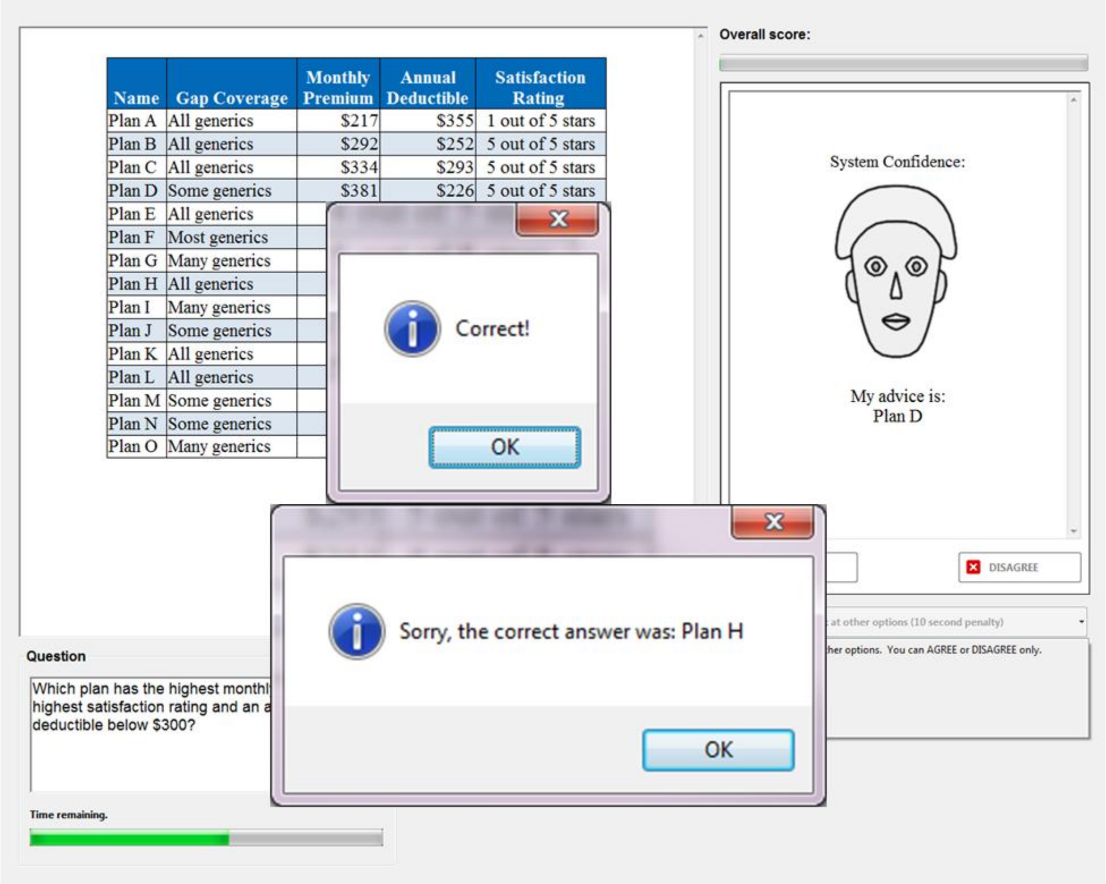
Figure 14. Experiment screen when the user chooses “Peek”.



How confident are you in your answer? (click one of the 7 buttons below)

How much do you trust the system? (click one of the 7 buttons below)

Figure 15. Trust and confidence scales after each trial.



Name	Gap Coverage	Monthly Premium	Annual Deductible	Satisfaction Rating
Plan A	All generics	$217	$355	1 out of 5 stars
Plan B	All generics	$292	$252	5 out of 5 stars
Plan C	All generics	$334	$293	5 out of 5 stars
Plan D	Some generics	$381	$226	5 out of 5 stars
Plan E	All generics			
Plan F	Most generics			
Plan G	Many generics			
Plan H	All generics			
Plan I	Many generics			
Plan J	Some generics			
Plan K	All generics			
Plan L	All generics			
Plan M	Some generics			
Plan N	Some generics			
Plan O	Many generics			

Correct!

Sorry, the correct answer was: Plan H

Question

Which plan has the highest monthl
highest satisfaction rating and an a
deductible below $300?

System Confidence:

My advice is:
Plan D

at other options (10 second penalty)

her options. You can AGREE or DISAGREE only.

Figure 16. Feedback screen after each trial.

Procedure

After reading the informational letter, participants completed the propensity to trust survey and the insurance experience questionnaire. Participants were then told that their job was to find the best prescription drug plan based on the criteria in the question, using the aid. Participants were told that they were using two different systems, system A and system B but were not explicitly told the difference in reliability-system confidence relationship between the two systems (or blocks). Participants were informed that both systems were mostly reliable, but no other explanation was given. Next, the experimenter oriented participants to the computer program and explained the decision task through 3 practice trials in which they agreed, disagreed or peeked to answer the question. Once participants no longer had questions, they began the first block of 30 trials. After completing block one, the experimenter started block two.

For each trial, participants were shown the question, the plan table, and the DSS. The timer bar began counting down 45 seconds as soon as the question was presented. Once the participant responded, the decision task screen disappeared and participants rated their trust in the DSS and their confidence in their own answer. The decision task screen then reappeared and provided feedback on the correctness of their answer.

At the conclusion of the decision task, participants completed computerized versions of a demographics questionnaire and abilities tests. The protocol for study 2 is in Appendix D.

Independent variables

Type of display (between subjects)

The 5 display types were 1) No system confidence (None), 2) Number, 3) Pie Graph (Pie), 4) Evaluative Categories (EvalCat), and 5) Anthropomorphic faces (Anthro). The None condition was analyzed as a control group to determine whether participants were using system confidence information when it was provided.

Reliability-System confidence coupling (within subjects)

We manipulated the extent to which reliability of the DSS and system confidence reported by the DSS corresponded. When they did correspond, reliability of the system was calculated such that *system confidence* was directly related to *system reliability*. For example, in the 25% confidence display condition, the automation will be correct in its advice for 25% of the trials and incorrect for 75% of the trials (see Table 3).

In the uncoupled condition, the relationship between actual reliability and reported system confidence was not related; reliability was 67% in each of the 3 conditions (Table 3). The purpose of this manipulation was to examine whether participants were using system confidence information and if it affected trust independently from reliability information. An additional purpose of this manipulation was to examine whether trust changes as a result from the mismatch in expectancy. This variable was within subjects and conditions and was counterbalanced so that half of the participants receive the coupled condition first, and the other half will receive the uncoupled condition first. Participants were not informed about the overall reliability of the system.

Table 3.
System Confidence and Reliability Manipulations

	Actual System Reliability	
System confidence level	**Coupled**	**Uncoupled**
25%	25% correct	67% correct
50%	50% correct	67% correct
75%	75% correct	67% correct

System confidence (within subjects)

The same levels of confidence used in study 1 were used in study 2: 25% (low), 50% (neutral), and 75% (high). Experimenters explained system confidence as “a system generated estimation of whether it is providing a correct suggestion that is based on the quality of the information it (the system) is using to provide a suggestion, on a trial by trial basis. If the system has low confidence, it is likely that it does not have much information to help it make a decision, the information is degraded, or there are other choices that are very similar and hard to choose between. If the system has high confidence, it is likely that the system has plenty of good information and that one prescription drug plan is much better than the others”.

Dependent Variables

Three categories of dependent measures were analyzed in this study: 1) Participant characteristics (propensity to trust and insurance experience), 2) Trust and dependence (behavioral trust, subjective trust, participants’ confidence in their answer), and 3) Performance (decision accuracy and task time). Trust calibration was assessed by comparing participants’ compliance with the DSS’s suggestion and actual system accuracy (reliability), similar to the analysis of McGuirl & Sarter (2006). If compliance was higher than actual system accuracy, this was considered mistrust (using the DSS when one shouldn’t). If compliance was lower than actual system accuracy, this was categorized as distrust (not using the DSS when one should).

Propensity to trust

A measure of participants' propensity to trust machines was chosen because of its high correlation with dispositional trust and history-based trust (Fiske & Neuberg, 1990) .The Complacency-Potential Rating Scale (CPRS) was developed by Singh, Molloy, & Parasuraman (1993) and has four factors: Confidence-Related, Reliance-Related, Trust-Related, and Safety-Related complacency. The CPRS has high internal consistency (r >.98) and test-retest reliability (r=.90) (Singh, et al., 1993). Participants rated each one of the 20 items on a 5 point Likert scale from Strongly Agree to Strongly Disagree. The CPRS was scored by adding up the scores for items 1-16 and excluding the last 4 filler items.

Insurance purchasing experience

An insurance purchasing experience questionnaire used in Price & Pak (accepted) was used to control for differences in insurance knowledge. The brief questionnaire asks 3 questions: 1) Have you ever purchased health insurance?, 2) If #1 is yes, how many years of experience do you have purchasing health insurance, 3) Have you ever purchased a prescription drug plan?, 4) If #3 is yes, how many years of experience do you have purchasing prescription drug plans? The answer choices for questions 1 and 3 were yes or no, and for 2 and 4 the answer choices were 1)Less than 6 months, 2) 6 months but less than 1 year, 3) 1 year but less than 5 years, 4) 5 years but less than 10 years, 5) At least 10 years.

Trust and Performance Dependent Variables

Subjective trust

After each trial, participants answered the question, “How trustworthy is the automation?” on a 7 point Likert scale used in Pak, Fink, Price, Bass, & Sturre (2012). The end points of the Likert scale ranged from (1) Not at all trustworthy to (7) Completely trustworthy.

Behavioral trust

In addition to subjective trust, an objective measure of trust will be based on participants’ actual behavior with the DSS. This measure of behavioral trust was significantly correlated with subjective trust (r=.35) in Pak, Fink, Price, Bass, & Sturre (2012). Participants had the option of agreeing, disagreeing, or peeking at other suggested answers. If participants unconditionally agreed with the DSS, this represented high trust and was coded with a value of 4. If participants peeked and then agreed, this represented an attitude of “trust but verify” and was coded with a value of 3. Peeking and then disagreeing represented moderate distrust and was coded with a value of 2. Immediately disagreeing with the DSS represented distrust and was coded with a value of 1.

Participants’ confidence in their decision

Participants also answered the question, “How confident are you in your answer?” on 7 point Likert scale used in Pak, Fink, Price, Bass, & Sturre (2012) after each trial. The end points of the Likert scale ranged from (1) not at all confident to (7) completely confident. Previous literature shows that when trust exceeds confidence, participants’ compliance with automation is much higher than when trust falls below confidence (Pak, et al., 2012; Lee & Moray, 1992, 1994).

Trust calibration

Trust calibration was analyzed using a Pearson correlation between mean behavioral trust and a created system reliability score. The system reliability score matched the system reliability on the same 4 point scale as behavioral trust, such that when system reliability was 25%, the appropriate behavioral trust score was equal to 1 (disagree with the system), 50% was 2.5 (either peek and disagree or peek and agree), and 75% was 4 (agree with the system). The higher the correlation, the better calibration there was between actual system reliability and behavioral trust.

Decision accuracy

Decision accuracy was a measure of decision performance. This was a binary measure; either the participant chose the best answer (scored as 1) or they did not (scored as 0). An overall accuracy score was computed by summing the number correct.

Task time

Task time was measured in seconds from the time the decision task was presented until the time the participant indicates their decision. Task time was restricted to a maximum of 45 seconds per question. In a previous study, Price & Pak (accepted) found that for this task, younger subjects' task time mean was 38.5 seconds (*SD*=2.87). The time constraint was reduced for this study to impose a higher workload.

HYPOTHESES

H1: A 3-way interaction was predicted such that in the low system confidence condition (25%) and neutral system confidence condition (50%) but not in the high system confidence condition (75%), subjective trust (DV) was higher in the anthropomorphic faces and evaluative categories conditions than the other 3 conditions (i.e., pie condition, number only condition, and no system confidence condition), but only when system confidence was coupled with reliability. In the uncoupled condition, no effect of system confidence level or system confidence display was expected.

H2. Higher correlations between system reliability and average behavioral trust are predicted for participants in the faces and evaluative categories conditions, in comparison to the non-affective conditions, but only when system confidence is coupled with reliability. Participants in the affective conditions (i.e., faces and evaluative categories) were expected to better trust the DSS's system confidence value. In other words, participants were expected to have a higher overall trust (subjective trust) in the DSS's ability to report an accurate system confidence value and would thus engage more often in verification behavior (i.e., peeking or disagreeing with the DSS) when system confidence was low (25% or neutral 50%). When system confidence was high, participants would be more likely to comply (i.e., agree) with the DSS.

H3. A 3-way interaction was predicted such that in the low system confidence condition (25%) and neutral system confidence condition (50%) but not the high confidence condition (75%), decision accuracy (DV) would behigher in the anthropomorphic faces and evaluative categories conditions than the other 3 conditions (i.e., graph condition, number only condition, and no system confidence conditions), but only when system confidence was coupled with reliability. In the uncoupled condition, no effect of system confidence level or system confidence display was expected. Participants in the affective conditions (i.e., anthropomorphic faces and evaluative categories) were expected to have higher overall subjective trust in the DSS. Similar to the justification of H3, participants would trust that a low (25%) and neutral (50%) system confidence rating requires verification and a high system confidence rating (75%) does not. If the user's behavioral trust is appropriately calibrated with system reliability (and if system reliability and system confidence are coupled), then accuracy was expected to be higher.

H4: A main effect of system confidence display on task time (DV) was predicted such that participants in the anthropomorphic faces condition will perform faster across all levels of system confidence and system confidence-reliability relationship, followed by the evaluative categories condition, graph condition, number only condition, and then the no system confidence condition. This prediction was based on the differences in the automaticity of processing feelings of trustworthiness between each of these conditions, as explained in the introduction and independent variables section of study 1. Furthermore, if users trusted the DSS and complied with its suggestion, there would not be a time cost associated with verification behavior (i.e., peeking and checking other answers).

RESULTS

Participants

One-hundred younger adults (49 male, 51 female) between the ages of 18 and 27 (*M*=21.5, *SD* =3.09) participated in study 2. No significant differences (*p*<.05) were between display conditions (the only between subjects variable) on propensity to trust, technology experience, perceptual speed abilities, working memory abilities, spatial orientation abilities, spatial visualization abilities, health status, or prescription drug plan insurance purchasing experience. There were significant differences in the mean age between display type groups (*F*(4, 95)=13.28, *p*<.000), such that subjects in the Pie (*M* =18.15, *SD*=.366) condition were significantly younger than those in all other conditions (i.e., Anthro (*M*=21.55, *SD*=2.86), EvalCat (*M*=21.4, *SD*=2.98), None (*M*=23.2,*SD*=2.58), and Number (*M*=23.2, *SD*=2.8). Due to this difference in age between display groups all analyses include analyses include age as a covariate for effects of display type (the only between subjects variable). The remaining participant characteristics are displayed in Table 4.

Table 4.

Table 4.
Experiment 2 Participant Characteristics (N=100)

Category	Frequency	Percentage
Gender		
Female	51	51%
Male	49	49%
Race/Ethnicity		
American Indian/Alaskan	1	1%
Asian	17	17%
Native Hawaiian/Pacific Islander	1	1%
Black/African American	12	12%
White	59	59%
Hispanic	6	6%
Multiracial	1	1%
Other	3	3%
Marital status		
Single	93	93%
Married	7	7%
Experience with computers?		
Yes	100	100%
Computer experience (years)		
At least 5 years	100	100%
Purchased health insurance?		
Yes	20	20%
No	80	80%
If yes, how long?		
< 6 months	12	60%
6 months but <1 year	4	20%
1 year but < 5 years	2	10%
5 years but <10 years	2	10%
Purchased prescription drug insurance?		
Yes	9	9%
No	91	91%
If yes, how long?		
< 6 months	3	33%
6 months but < 1 year	3	33%
1 year but < 5 years	3	33%

Trust and Performance Measures

The following analyses are organized by the specific hypotheses outlined in the previous section. A 4 (display type: anthro, nPie, evalcats, number) x 3 (system confidence) x 2 (system confidence reliability-coupling) mixed factors ANOVA on subjective trust, behavioral trust, time, accuracy and confidence was conducted. In addition, to test the control condition (no system confidence display), an additional 3 (system confidence) x 2 (system confidence reliability-coupling) ANOVA on the same 5 dependent measures was conducted. Post-hoc analyses were conducted for significant effects using Bonferroni corrections. The results are structured so that the results from the first ANOVA for display conditions is presented first, followed by the second ANOVA on the control (no display) condition.

Subjective Trust

Mauchly's test indicated that the assumption of sphericity had been violated for System confidence ($\chi 2$ (2) = 28.473, p < .001) and thus the degrees of freedom were corrected using Huynh-Feldt estimates of sphericity, ε = .839 (Huynh & Feldt, 1976). System-confidence-reliability coupling and the interaction between system confidence and coupling both met the assumptions of sphericity and thus no degrees of freedom corrections were made for those analyses.

The first ANOVA for display type conditions revealed significant main effects of coupling (F(1, 76)=9.04, p<.004, η_p^2 =.11) and system confidence (F(1.64, 124.6)=56.63, p<.000, η_p^2 =.43) on subjective trust. There was not a significant main effect of display type (p=.23). Subjective trust was higher when system reliability and confidence were uncoupled (*M*=4.57, *SD*=1.22) than coupled (*M*=4.3, *SD*=1.08). Subjective trust ratings increased as system confidence display level increased; trust was rated lower in the low system confidence level (25% *M*=3.09, *SD*=1.06*)* than neutral (50% *M*=4.36, *SD*=1.11) and high (75% *M*=4.66, *SD*=1.14) levels of system confidence. System confidence level had the greater effect (η_p^2 =.43) on subjective trust than coupling (η_p^2 =.11).

Main effects were qualified by a significant interaction between coupling and system confidence (*F*(5.99, 146)=15.83, p<.000, η_p^2 =.17). The interaction between coupling and system confidence (Figure 17) indicated that participants rated their trust in the low (25%) and neutral (50%) confidence conditions significantly higher when reliability and confidence were uncoupled than coupled (25% confidence: coupled *M*=3.71, *SD*=1.08, uncoupled *M*=4.1, *SD*=1.26; 50% confidence: coupled *M*=4.14, *SD*=1.11, uncoupled *M*=4.6, *SD*=1.27). There was no difference in subjective trust between coupled and uncoupled conditions with the 75% system confidence level display. In the coupled condition, subjective trust increased as system confidence increased from 25% (*M*=3.71, *SD*=1.08) to 50% (*M*=4.14, *SD*=1.11) to 75% (*M*=4.68, *SD*=1.16). In the uncoupled condition, trust was lower in the 25% (*M*=4.1, *SD*=1.25) condition than in the 50% (*M*=4.59, *SD*=1.26) and 75% (*M*=4.62, *SD*=1.29). However, in the uncoupled condition, there was no difference between the 50% and 75% conditions.

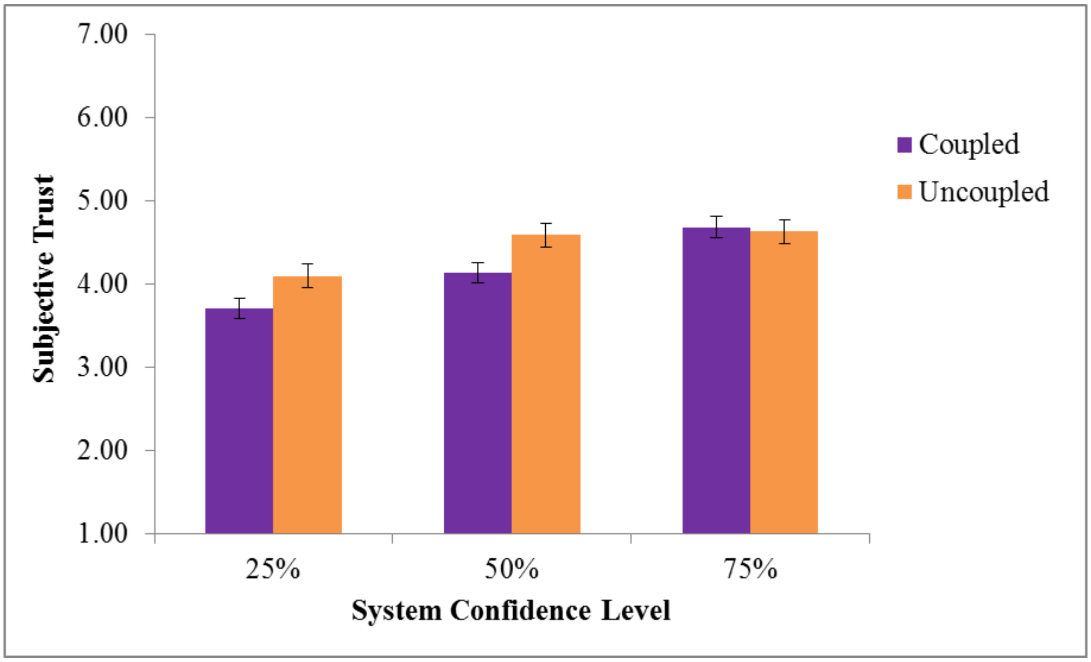


Figure 17. Mean subjective trust by system confidence for all displays. Error bars represent standard error of the mean.

For the no display condition (None), there was a significant main effect of system confidence (*F*(1.32, 5.03)=4.96, p=.02, η_p^2 =.21) and a significant interaction between system confidence and coupling (*F*(1.88, 8.73)=4.09, p=.03, η_p^2 =.18), but no main effect of coupling (*p*=.09; see Figure 18). In the uncoupled condition, participants rated their trust lower in the 25% confidence condition (*M*=4.95, *SD*=1.14) than in the 50% confidence condition (*M*=5.28, *SD*=1.29). In the coupled conditions, there were no differences in trust between system confidence display levels (*p* >.05). At 50% system confidence (though not displayed), trust was higher in the uncoupled conditions (*M*=5.28, *SD*=1.29) than the coupled conditions (*M*=4.7, *SD*=1.29), but there were no effects of coupling on subjective trust in the 25% and 75% system confidence conditions (*p*>.05).

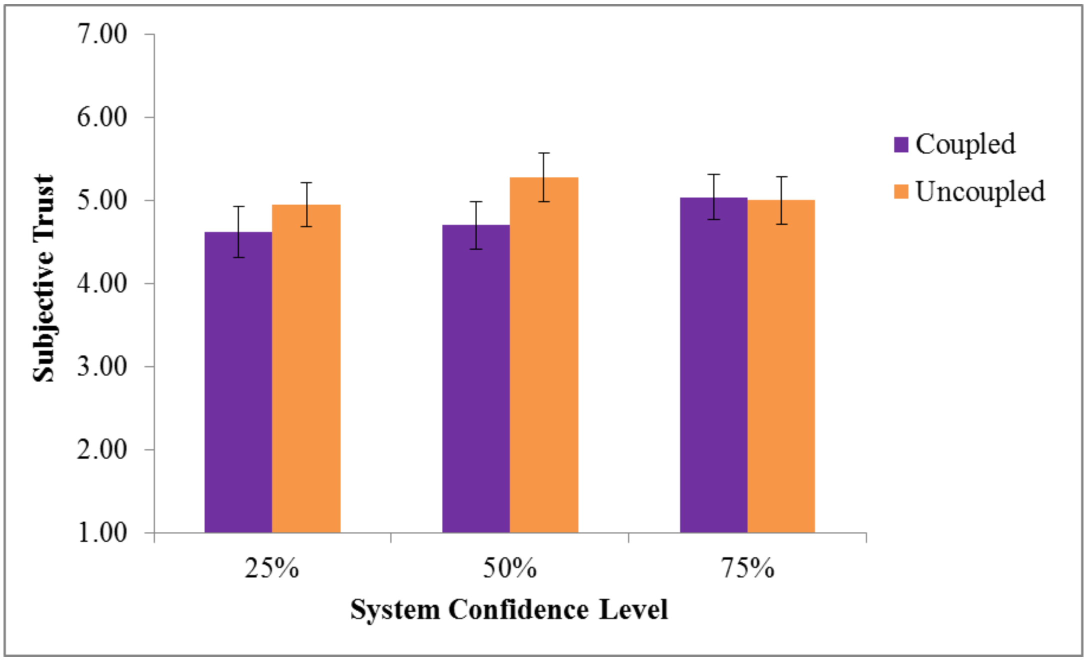


Figure 18. Mean subjective trust by system confidence for no display. Error bars represent standard error of the mean.

Behavorial trust and trust calibration

Mauchly's test indicated that the assumption of sphericity had been violated for System confidence ($\chi 2$ (2) = 11.2, p = .004) and thus the degrees of freedom were corrected using Huynh-Feldt estimates of sphericity, ε = .954 (Huynh & Feldt, 1976). System-confidence-reliability coupling and the interaction between system confidence and coupling both met the assumptions of sphericity and thus no degrees of freedom corrections were made for those analyses.

The first analysis of display types resulted in no effect of display condition (*p*=.95) on behavioral trust. There were main effects of coupling (*F*(1,76)=85.72, *p*<.000, η_p^2 =.53), system confidence (*F*(1.86,141.33)=149.7, *p*<.000, η_p^2 =.66), and a significant system confidence and coupling interaction (*F*(2,152)=57.04, *p*<.000, η_p^2 =.43). The results are graphed in Figure 19. Both system confidence (η_p^2 =.66) and coupling (η_p^2 =.43) had strong effects on behavioral trust.

When system confidence was 25% or 50%, participants were more likely to agree with the DSS in the uncoupled conditions (25% *M*=2.64, *SD*=.51; 50% *M*=3.38, *SD*=.52) than coupled conditions (25% *M*=2.1, *SD*=.55; 50% *M*=2.62, *SD*=.62). There was not a significant difference in the 75% condition between coupled and uncoupled conditions (*p*>.05). In the coupled condition, behavioral trust was significantly higher as system confidence increased from 25% (*M*=2.1, *SD*=.55) to 50% (*M*=2.62, *SD*=.62) and to 75% (*M*=3.15, *SD*=.53). In the uncoupled condition, although all 3 levels were significantly different, the 50% condition had higher behavioral trust (*M*=3.38, *SD*=.52) than 25% (*M*=2.65, *SD*=.51 and 75% conditions (*M*=3.13, *SD*=.52). Behavioral trust was also higher overall in the uncoupled conditions (*M*=3.05, *SD*=.41) than the coupled conditions (*M*=2.62, *SD*=.47).

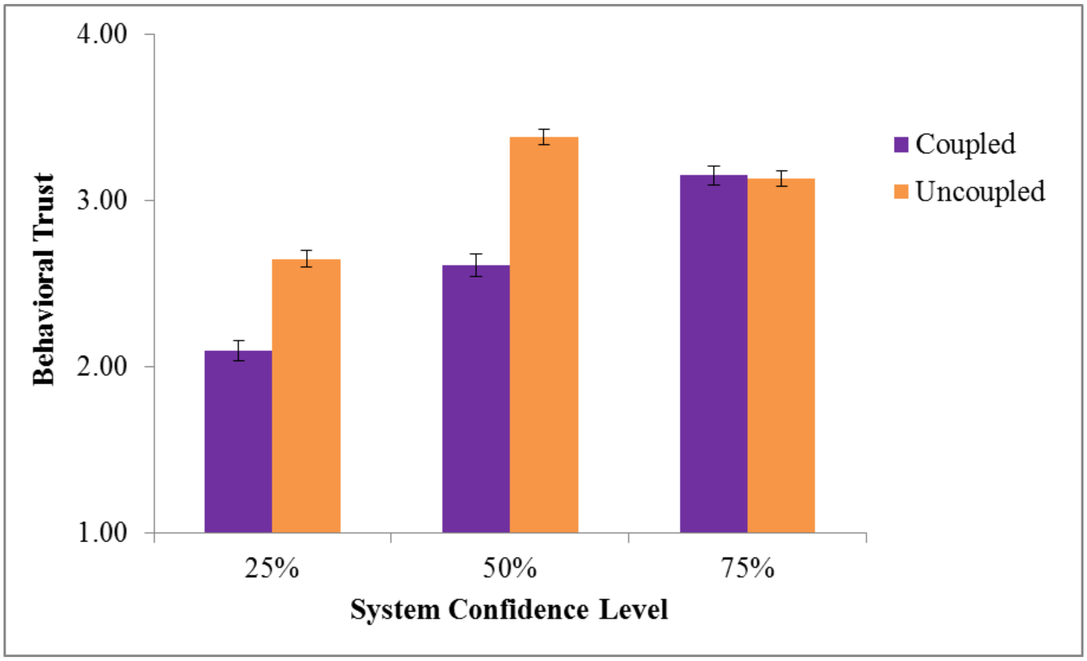


Figure 19. Mean behavioral trust by system confidence for all displays. Error bars represent standard error of the mean.

For the no display condition analysis, there were significant main effects of coupling (*F*(1,19)=127.29, p<.000, $\eta_p^{\,2}$ =.87), system confidence (*F*(2,38)=36.18, p<.000, $\eta_p^{\,2}$ =.66), and a significant interaction between coupling and system confidence (*F*(1.9, 36.1)=26.96, p<.000, $\eta_p^{\,2}$ =.59; see Figure 20). The same pattern emerged as in the display condition analysis; participants were more likely to agree with the DSS when coupled than uncoupled in the low (25% coupled *M*=2.19, *SD*=.41; 25% uncoupled *M*=2.78, *SD*=.34) and neutral system confidence conditions (50% coupled *M*=2.49, *SD*=.43; 50% uncoupled *M*=3.54, *SD*=.25). Additionally, overall trust was higher in the uncoupled conditions (*M*=3.1, *SD*= .21) than coupled conditions (*M*=2.54, *SD*=.32). In the coupled condition, behavioral trust was significantly higher as system confidence increased from 25% (*M*=2.18, *SD*=.41) to 50% (*M*=2.49, *SD*=.45) and to 75% (*M*=2.94, *SD*=.45). However, in the uncoupled condition, although trust in all 3 system confidence levels were significantly different from each other, the 50% condition had higher behavioral trust (*M*=3.54, *SD*=.25) than 25% (*M*=2.78, *SD*=.34) and 75% (*M*=2.99, *SD*=.29).

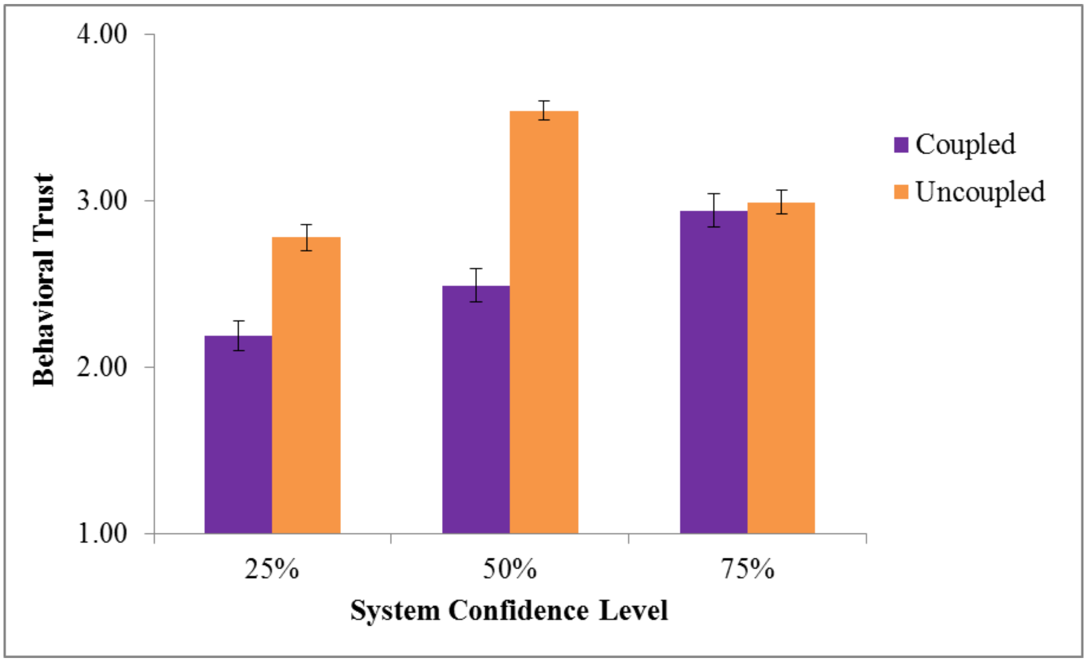


Figure 20. Mean behavioral trust by system confidence for no display. Error bars represent standard error of the mean.

Trust calibration was analyzed using Pearson correlations between system reliability score and behavioral trust score across display types and system confidence values. Although the planned analysis included running the 3 levels of system confidence separately to determine if there was greater mistrust or distrust at different levels of system confidence, this was not feasible. A system reliability score was assigned to the 3 levels of system confidence (25%=1, 50%=2.5, 75%=4) in order to map what should have been the appropriate behavioral trust response (i.e., 1=disagree, 2=peek and disagree, 3=peek and agree, 4=agree). If correlations were run for each level of system confidence, the system reliability score would be constant and thus a correlation is not possible. Thus, the correlation analysis was run between coupling conditions and across display conditions only and represents an overall trust calibration collapsed across system confidence conditions.

Table 5 represents the Pearson correlation coefficients in order of magnitude for the 5 display types and the coupled vs. uncoupled conditions. A Chi square test of equality of independent correlations was conducted between display types and found no significant differences (*p*>.05).

Table 5.
Correlations Between Behavioral Trust and System Reliability Score

Display Type	R	*p*
Coupled (*n*=720)		
Anthro	.320	<.000
Number	.290	<.000
Pie	.289	<.000
EvalCat	.281	<.000
None	.207	<.000
Uncoupled (*n*=720)		
Anthro	.220	<.000
Pie	.127	.001
Number	.123	.001
EvalCat	.113	.001
None	.064	n.s.

Decision accuracy

System confidence violated the assumption of sphericity (Mauchly's Test, ($\chi 2$ (2) = 18.28, p < .000) and thus the degrees of freedom were corrected using Huynh-Feldt estimates of sphericity, ε = .90 (Huynh & Feldt, 1976) for that variable only. All others are reported with sphericity assumed values.

For the display type analysis, there were significant main effects of coupling (*F*(1, 95)=30.28, *p*<.000, η_p^2 =.29) and system confidence (*F*(1.84, 139.9)=86.38, *p*<.000, η_p^2 =.53). There was not a significant main effect of display type (*p*>.05). The significant main effects were qualified by a significant interaction between coupling and system confidence (*F* (2, 190)=21.42, *p*<.000, η_p^2 =.22). System confidence had the greater effect on accuracy (η_p^2 =.53). These results are graphed in Figure 21.

Participants were more accurate when system confidence and reliability were uncoupled than coupled, but only in the 25% (coupled *M*=.59, *SD*=.15; uncoupled *M*=.74, *SD*=.15) and 50% (coupled *M*=.74, *SD*=.14; uncoupled *M*=.79, *SD*=.10) conditions. There was no difference in accuracy between coupling conditions in the 75% condition (*p*>.05). In the coupled condition, accuracy increased as system confidence increased from 25% (*M*=.59, *SD*=.15) to 50% (*M*=.74, *SD*=.14) to 75% (*M*=.81, *SD*=.09). In the uncoupled condition, accuracy was higher in the 50% (*M*=.79, *SD*=.10) and 75% (*M*=.80, *SD*=.12) condition compared to the 25% (*M*=.74, *SD*=.15) condition, but there was no difference in accuracy between 50% (*M*=.79, *SD*=.10) and 75% (*M*=.80, *SD*=.12).

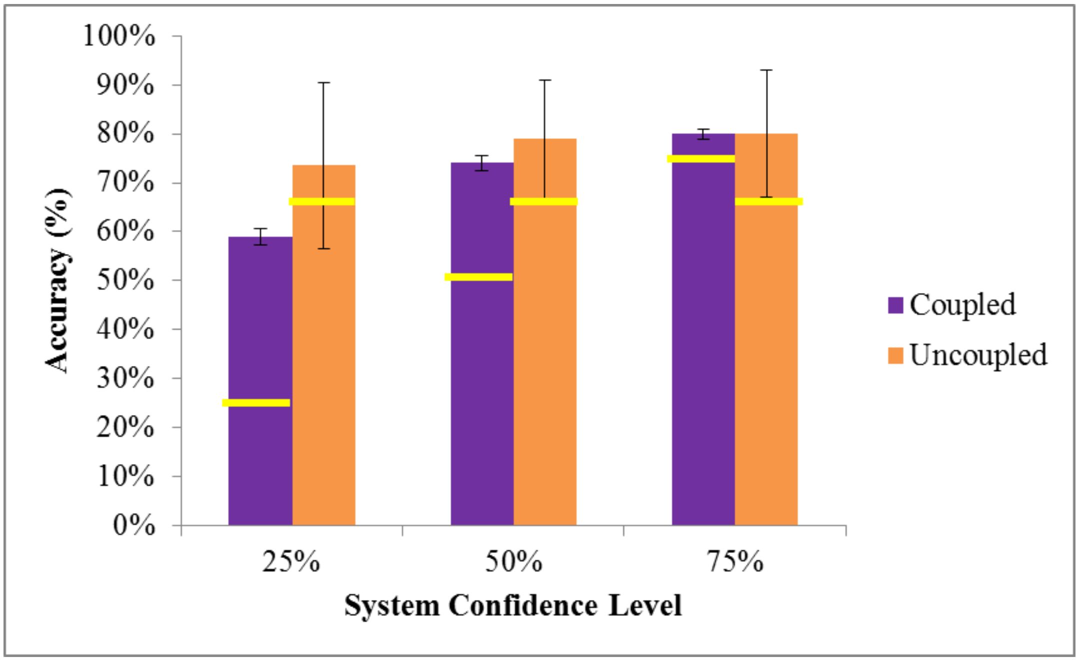


Figure 21. Percent accuracy by system confidence for all displays. Yellow bars represent actual system reliability and error bars represent standard error of the mean.

In the control analysis, there were significant main effects of coupling (F(1,19)=15.73, p=.001, η_p^2 =.87) and system confidence (F(1.73, 32.85)=4.98, p=.02, η_p^2 =.21), see Figure 22. The interaction, however, was insignificant. Coupling had a larger effect on accuracy (η_p^2 =.87) compared to system confidence level (η_p^2 =.21). Accuracy was higher in the uncoupled (*M*=.80, *SD*=.08) condition than in coupled condition (*M*=.69, *SD*=.15). Accuracy was higher in the 50% (*M*=.78, *SD*=.13) and 75% (*M*=.79, *SD*=.09) conditions than the 25% (*M*=.67, *SD*=.16) condition, but there was no difference between the 50% (*M*=.78, *SD*=.13) and 75% (*M*=.79, *SD*=.09) conditions.

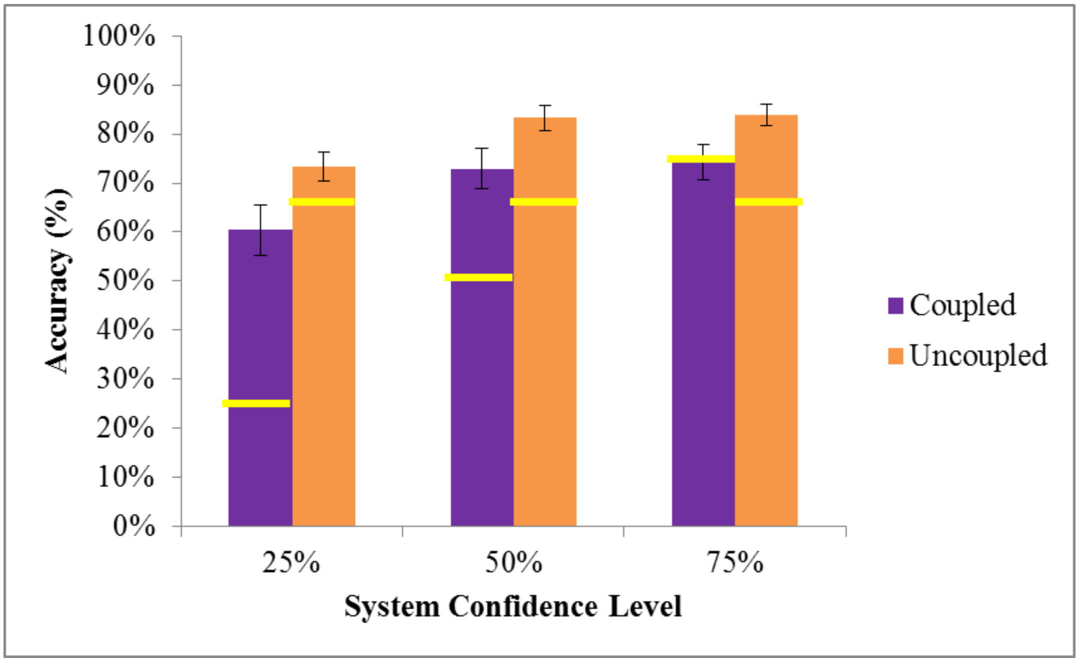


Figure 22. Percent accurate by system confidence for no display. Yellow bars represent actual system reliability and error bars represent standard error of the mean.

Decision task time

System confidence violated the assumption of sphericity (Mauchly's Test, ($\chi 2$ (2) = 39.9, p < .000) and thus the degrees of freedom were corrected using Huynh-Feldt estimates of sphericity, ε = .784 (Huynh & Feldt, 1976) for that variable only. All other results are reported with sphericity assumed values.

For the display type conditions, there was a significant main effects of system confidence (F(1.53, 116.84)=68.95, p<.000, η_p^2 =.48), but not of coupling or display type (p>.05) on task time. However, there was a significant interaction with a small effect on task time between system confidence and coupling (F(2, 150.63)=6.58, p=.002, η_p^2 =.08) and this is displayed in Figure 23. In the 50% condition, uncoupled task time (M=27.95, SD=1.4) was significantly faster than coupled task time (M=30.94, SD=1.07). No differences were found in coupling within the 25% and 75% conditions (p>.05). In the coupled condition, time was faster as system confidence increased from 25% to 50% to 75% and all differences between levels were significant. In the uncoupled condition, time was faster in the 50% and 75% conditions compared to 25%, however task time was not significantly different between 50% and 75%.

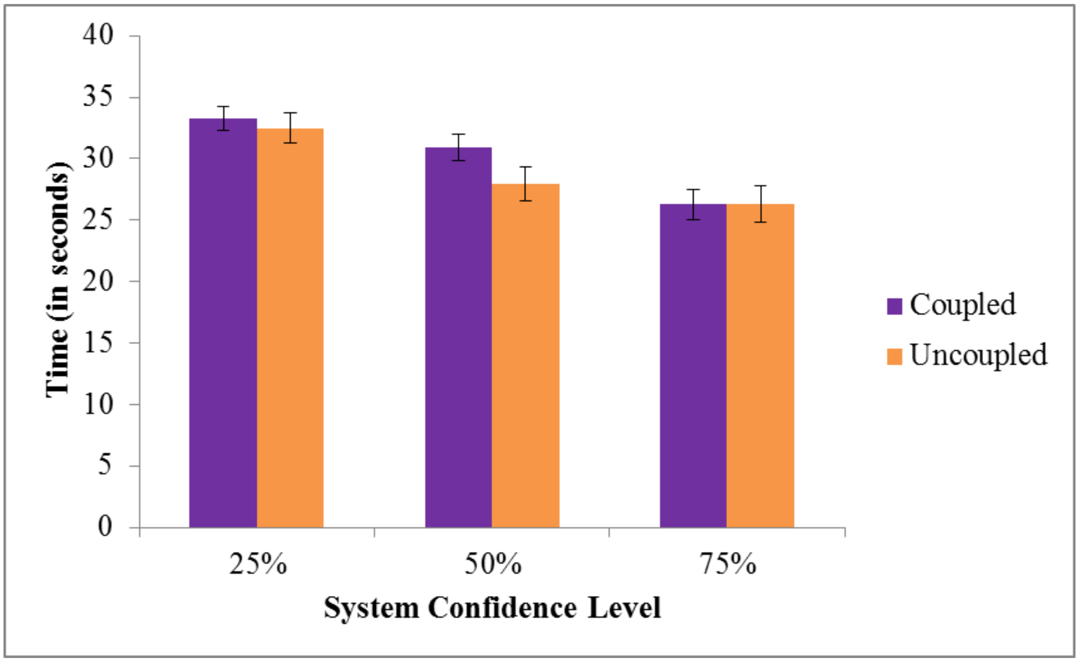


Figure 23. Mean task time by system confidence for all displays (in seconds). Error bars represent standard error of the means.

In the control condition, there was a significant main effect of system confidence ($F(2, 38)=16.43$, $p<.000$, $\eta_p^2 =.46$), but not of coupling ($p>.05$) on task time. None of the interactions were significant ($p>.05$). Figure 24 shows the effect of system confidence. Task time was faster in the 50% (*M*=28.22, *SD*=5.4) and 75% (*M*=28.7, *SD*=4.69) system confidence conditions compared to the 25% (*M*=31.28, *SD*=5.4) condition but the difference in task time between 50% (*M*=28.22, *SD*=5.4) and 75% (*M*=28.7, *SD*=4.69) was not significant.

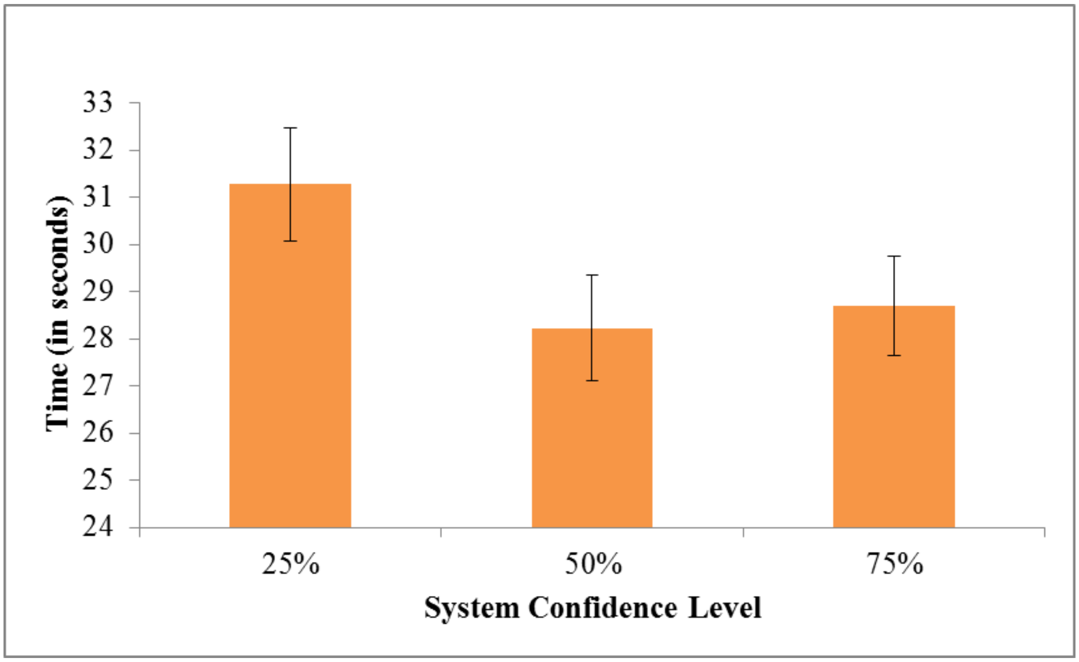


Figure 24. Mean task time by system confidence for no display (in seconds). Error bars represent standard error of the means.

Confidence

For the display conditions, a significant main effect was present for system confidence (*F*(2, 76)=9.98, p<.000, $\eta_p^{\ 2}$ =.12). Graphed in Figure 25. Participants rated their confidence in their answer higher in the 75% system confidence condition (*M*=5.31, *SD*=1.17) than in the 25% system confidence condition (*M*=5.05, *SD*=1.07). There was no difference between the 50% condition and the 75% or 25% conditions (*p*>.05). For the control condition, there were no significant main effects of any variable (*p*>.05).

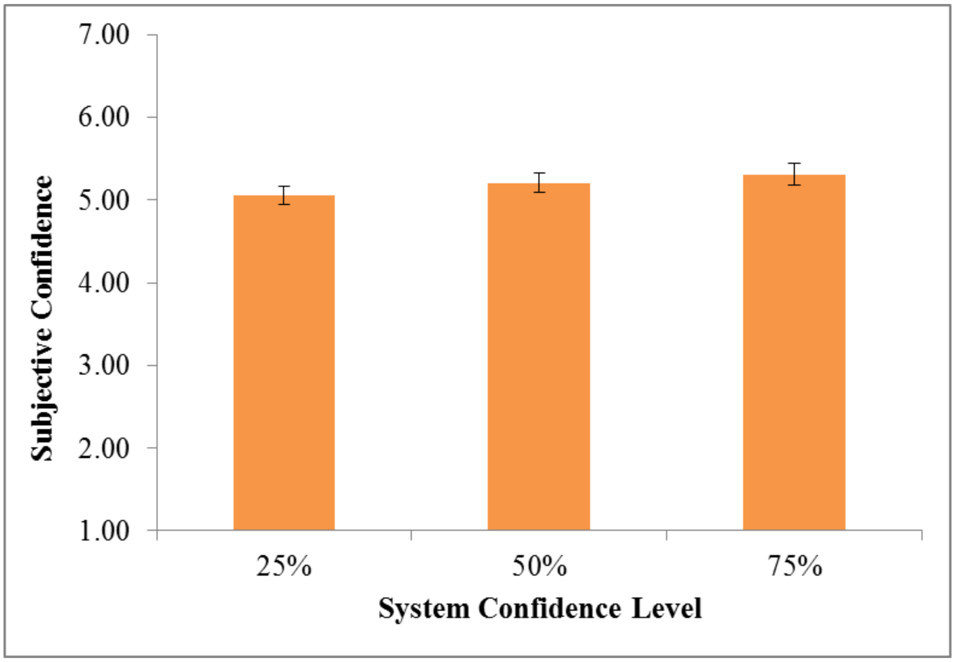


Figure 25. Mean subjective participant confidence ratings by system confidence for all displays. Error bars represent standard error of the means.

DISCUSSION

The main goals of the two experiments were to gain an understanding of trust behavior with a DSS with different system confidence displays. Specifically, affective displays in the form of an anthropomorphic face display and the use of evaluative categories was compared to more traditional displays (i.e., Pie graph, Number, or None). Affective displays were thought to evoke more automatic feelings of goodness or badness (or in this case, trustworthiness) and thus influence verification behavior when it is appropriate (i.e., in low system confidence conditions). Anthropomorphic face displays were designed to convey system confidence through affect, similar to how confidence and trust are conveyed in face to face interactions. Following this logic, trust was expected to be better calibrated when reading faces for confidence compared to a number or a more traditional display (i.e., numerical display, bar graph, or pie graph) where the user has to interpret the meaning of a number and decide whether to trust the DSS.

Study Hypotheses

H1: A 3-way interaction was predicted such that in the low system confidence condition (25%) and neutral system confidence condition (50%) but not in the high system confidence condition (75%), subjective trust (DV) was higher in the anthropomorphic faces and evaluative categories conditions than the other 3 conditions (i.e., pie condition, number only condition, and no system confidence condition), but only when system confidence was coupled with reliability. In the uncoupled condition, no effect of system confidence level or system confidence display was expected.

H2. Higher correlations between system reliability and average behavioral trust are predicted for participants in the faces and evaluative categories conditions, in comparison to the non-affective conditions, but only when system confidence is coupled with reliability. Participants in the affective conditions (i.e., faces and evaluative categories) were expected to better trust the DSS's system confidence value. In other words, participants were expected to have a higher overall trust (subjective trust) in the DSS's ability to report an accurate system confidence value and would thus engage more often in verification behavior (i.e., peeking or disagreeing with the DSS) when system confidence was low (25% or neutral 50%). When system confidence was high, participants would be more likely to comply (i.e., agree) with the DSS.

H3. A 3-way interaction was predicted such that in the low system confidence condition (25%) and neutral system confidence condition (50%) but not the high confidence condition (75%), decision accuracy (DV) would behigher in the anthropomorphic faces and evaluative categories conditions than the other 3 conditions (i.e., graph condition, number only condition, and no system confidence conditions), but only when system confidence was coupled with reliability. In the uncoupled condition, no effect of system confidence level or system confidence display was expected. Participants in the affective conditions (i.e., anthropomorphic faces and evaluative categories) were expected to have higher overall subjective trust in the DSS. Similar to the justification of H3, participants would trust that a low (25%) and neutral (50%) system confidence rating requires verification and a high system confidence rating (75%) does not. If the user's behavioral trust is appropriately calibrated with system reliability (and if system reliability and system confidence are coupled), then accuracy was expected to be higher.

H4: A main effect of system confidence display on task time (DV) was predicted such that participants in the anthropomorphic faces condition will perform faster across all levels of system confidence and system confidence-reliability relationship, followed by the evaluative categories condition, graph condition, number only condition, and then the no system confidence condition. This prediction was based on the differences in the automaticity of processing feelings of trustworthiness between each of these conditions, as explained in the introduction and independent variables section of study 1. Furthermore, if users trusted the DSS and complied with its suggestion, there would not be a time cost associated with verification behavior (i.e., peeking and checking other answers).

None of the above hypotheses were supported by the results because there were no effects of display type on any of the dependent variables. The theoretical basis for these predictions was that numbers were replaced by affective information to relay system confidence information, that it would increase trust and reduce the effort needed to make a decision. One possible explanation as to why there were no effects of the anthro display condition may be that the degree of anthropomorphism was too low to influence trust (Lee & See, 2004). A low degree of anthropomorphism was chosen for this study to eliminate confounds of age, gender, attractiveness, and expertise of the DSS. Perhaps the degree of anthropomorphism was too low and did not evoke affective feelings of goodness or badness. The anthro display may have had a greater influence on trust if a more context-appropriate anthropomorphic “agent” was used, such as a medical professional or financial professional.

Another potential explanation for a lack of subjective trust differences between the anthro display and other display types may be rooted in dispositional trust or preconceived notions about the trustworthiness of the DSS. Previous research has shown that people lack trust of insurance companies and their agents (Hunter, Whiddett, Norris, McDonald, & Waldon, 2009) and this task involved choosing a prescription drug plan from a list of tables. Previous research examining trust in a health information sharing context indicated that people have a general distrust of sharing any information with health insurance agents or companies that determine eligibility when compared to doctors and medical professionals. Thus, if the DSS that was providing the suggestion was viewed as untrustworthy due to preconceived notions of insurance companies and the agents that work for them, then perhaps participants were simply unable to trust the anthro face display an intended.

Another possible explanation is that the system confidence display was not being used by participants and thus was not influencing trust. The effects of coupling, system confidence display, and task type implications are discussed in the following sections.

Effects of Coupling

To review, two conditions were included in this study to evaluate whether system confidence information was being utilized differently than reliability information, and whether participants could detect the difference between system confidence and reliability when it was not equal (or uncoupled). Three levels of system confidence were used, 25%, 50%, and 75%. In the system confidence-reliability coupled condition, the reliability matched the system confidence value (i.e., when system confidence was 25%, the answer was correct 25% of the time). In the uncoupled condition, reliability was 67% for each of the 3 system confidence levels. Ideally, participants should not use the system confidence information in the 25% and 50% conditions and instead should ignore that value and trust the DSS most of the time. In the coupled conditions, trust should increase as system confidence increases to correctly match the reliability of the system.

In the coupled condition, participants in the condition with the system confidence display were able to detect differences between the 3 levels of system confidence, as evidenced by their high trust ratings as system confidence improved (subjective trust) and by greater verification behavior (behavioral trust) when system confidence was low and neutral versus high. As expected, this increase in verification behavior led to higher task times in the low and neutral conditions compared to the high system confidence condition. Although participants did perform more accurately than the system at all levels of system confidence, accuracy was still lowest in the 25% system confidence condition. Although lower, the system was only 25% accurate, but participants were accurate 59% of the time. Participants' accuracy did benefit from the additional verification or peeking behavior. These findings are similar to what previous studies found, system confidence displays help users detect when they should and should not trust the system and this leads to better performance when system confidence and reliability are equal (Spain & Bliss, 2009; McGuirl & Sarter, 2006).

In the uncoupled conditions, participants should have rated trust the same across all three levels of system confidence because reliability was actually the same across all levels. Instead, participants rated their trust lower in the 25% condition than the 50% or 75% conditions. Participants also engaged in more verification behavior at 50% than 25% or 75%, where there should have been no difference in verification behavior because reliability was the same across all 3 system confidence levels. However, if participants were using system confidence information instead of reliability, it would be expected that it would follow the same pattern as the coupled condition. Instead, 50% system confidence seemed to cause confusion and lead to more instances of verification. The additional verification behavior did lead to higher accuracy in the 50% system confidence level condition over the 25% system confidence condition. Interestingly, this difference in the 50% did not lead to greater task time in this condition, and instead participants were faster in the 50% and 75% conditions than the 25% condition, eliminating the possibility of speed accuracy tradeoff in the 25% condition.

The results indicate that having system confidence is helpful when system confidence matches reliability, in accordance with previous studies (Spain & Bliss, 2009; McGuirl & Sarter, 2004). However, when the relationship between system confidence and reliability are uncoupled, this can lead to problems, particularly if system confidence for a trial is low but the DSS has high reliability (at least 67%). Having a system confidence display of 25% when system reliability was 67% led to distrust and diminished accuracy, trust, and speed.

One major limitation in the comparison between coupled and uncoupled conditions exists in this study. Overall reliability was actually higher in the uncoupled condition (67%) than in the coupled condition (50%). This is the most probable explanation for why main effects of coupling always favored the uncoupled condition for each dependent variable. Even so, participants should have noticed the greater disparity between 25% reliability and 67% reliability in the uncoupled condition compared to a 25% to 50% disparity in the coupled condition, and not allowed the system confidence value to bias their trust and behavior. Instead, participants relied on the system confidence value, even when they shouldn't have.

Effects of System Confidence Display

Were participants using the system confidence display or simply learning the reliability of the system over time using the feedback? The inclusion of the control condition (which did not present any system confidence information) allows for the comparison between the patterns in trust, time, and accuracy when the display is present and when it is absent.

The initial planned analysis did not show any differences between any of the display condition types and the control display. However, the control was run separately in the analysis to better understand the effects of coupling and system confidence levels. In the control condition, reliability changed in the coupled conditions, but the participant had no information to alert them to this change.   Essentially, the participant had a 67% reliable system (uncoupled) and a 50% reliable system (coupled). If participants were not able to detect the change in reliability between trials in the coupled system *without* the system confidence information (e.g., trials with 25% reliability) then it would suggest that participants did benefit from having the system confidence display. In this case, the pattern in the coupled condition would remain the same *without* the display as the pattern in performance *with* the display.

In the coupled conditions, participants did not rate their trust differently as reliability changed between 25%, 50%, and 75%, but they were more likely to engage in verification behavior as reliability decreased (i.e., more likely to peek when reliability was 25%).   Accuracy, however did not improve as peeking behavior increased. The increase in peeking behavior in the low reliability condition *without* the system confidence display means participants were using reliability information to drive behavior. The DSS provided 1 recommendation, which the participant likely first checked for accuracy against the criteria in the question, since all possible answers were shown in the table. At that point, they would know if the DSS was reliable or not. If the DSS was deemed unreliable on that trial, the participant would choose to peek or disagree. However, the participant still had to find the correct answer. If the participant peeked, they were given 3 other answers that they would still need to verify. The time constraint may have been adequate to determine if the suggestion from the DSS was reliable, but inadequate to determine the correct answer, even with the other 3 possibilities to shorten the task.

The fact that users were able to determine reliability changes in the coupled condition without the use of system confidence challenges the idea that it is beneficial to provide system confidence. In a highly reliable system, a low system confidence value may not mean that the DSS is unreliable (e.g., as in the uncoupled condition), but instead that there are similar choices that make it difficult for the system to determine which is the *best* choice. The low system confidence rating may lead the user to distrust the system, causing worse performance and lower efficiency if the user tries to make the decision on their own. This is especially true in cases where there are minimal consequences for not choosing the *best* choice, but instead choosing one that is good enough. However, in scenarios where there are high or risky consequences for not choosing the best or correct answer, perhaps the added verification behavior is warranted. In that situation, the information should be structured so that it is easier for the user to find the correct answer.

Task Type and Expertise

In comparison to previous studies, this consumer-like task was much different. The task was designed to be resource limited instead of data limited, provided multiple answers to choose from, and was non-compensatory. All of the information needed to find the correct answer was provided to the participant, unlike previous studies where information was degraded (Spain & Bliss, 2009) or expertise was needed to evaluate a complex strategy where trade-offs could be made between options (McGuirl & Sarter, 2004). Participants were able to evaluate the reliability of the suggestion quickly because the DSS narrowed down the information needed to choose whether to agree, disagree, or peek.

In this respect, having all of the information available reduces the need to rely on one’s own self-confidence the way an expert user might. An expert may decide whether to trust the automation based on their confidence in their own ability to solve the problem. In this study, novice users with little domain expertise (i.e., insurance purchasing knowledge) would have been more likely to doubt themselves and trust the automation. However, the user was able to quickly verify the DSSs suggestion and thus determine reliability on a trial by trial basis. Increased verification behavior indicates that participants were using the automation to find an answer rather than solving it on their own as an expert user might. This task however, did not require domain expertise would not have helped the participant choose the correct answer as each still had to meet the requirements in the question.

Limitations and Future Research

One possible limitation of this study was the degree of anthropomorphism used in the system confidence display. Previous research has shown that the increasing the degree of anthropomorphism leads to higher ratings of competency and trustworthiness (Gong, 2008). The faces in this study might be classified as low-level anthropomorphism (Gong, 2008) because they were more robot-like than human-like. Further research should look at the effects of higher level anthropomorphic displays (i.e., those that look more like actual human faces) and matching similar characteristics such as gender, age, and ethnicity.

Although the anthro display did not have a positive effect as predicted on trust and performance in this study, it also did not negatively affect performance. Participants were able to distinguish between low, neutral, and high levels of system confidence through faces just as easily as the numerical display. This suggests that it is safe to replace other displays with an anthro display that may be more comforting, likeable, or satisfying to a user than traditional displays. Future research should examine whether users would rate the system more enjoyable to use with an anthro display.

Future research should also examine different types of consumer tasks without time constraints. If the task in this study had been designed without the time constraint, it’s possible that participants may have also improved their accuracy, most likely at the expense of efficiency. Most consumer decisions such as choosing a prescription drug plan would not have such a strict time constraint. The decision itself would still be resource intensive without the DSS. The inclusion of the DSS changes the nature of the task to a verification task from a difficult non-compensatory task. Instead of having to examine every option and attribute, the participant had to verify the answer. This may restrict generalization to other tasks that do not list all possible options for comparison or tasks where alternative suggestions are not provided under a “peek” option.

A major limitation in this study was that the coupled and uncoupled conditions did not have the same overall reliability. Future research should examine more varied levels of reliability and system confidence and keep them constant between conditions. This is difficult to do because the number of tasks must remain feasible while providing enough to explore different ratios of accuracy (i.e., there must be enough tasks to have 40%, 50%, 60% reliability). It would be beneficial to know if there are ranges of system confidence where verification behavior is more likely (e.g., how low is system confidence) or less likely (e.g., how high is system confidence) to occur. In addition, future studies examining system confidence displays should include a control condition without a display to differentiate the effects of reliability and system confidence.

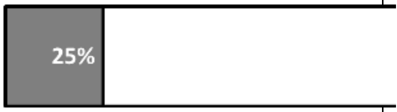
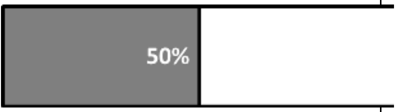
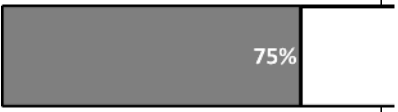
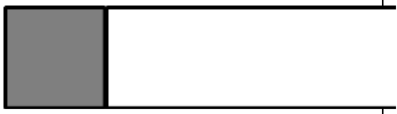
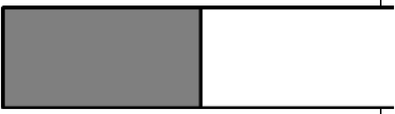
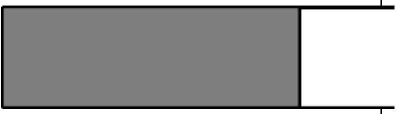
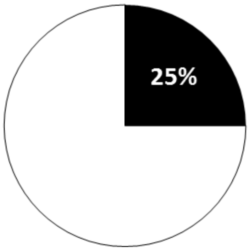
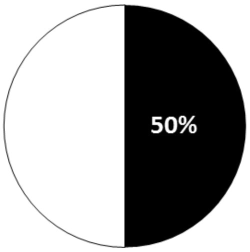
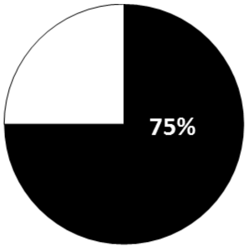
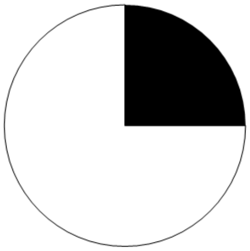
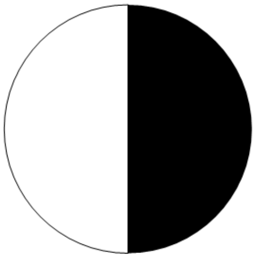
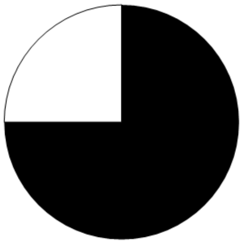
CONCLUSION

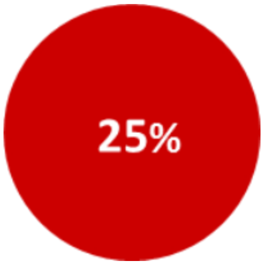
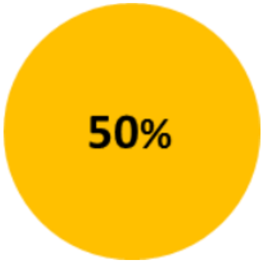
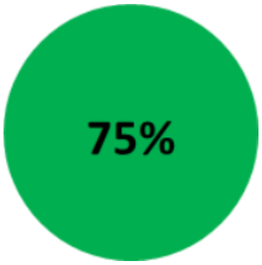
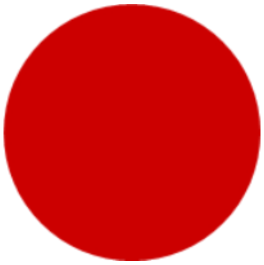
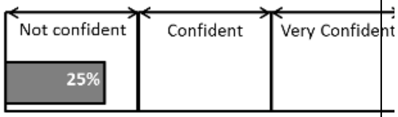
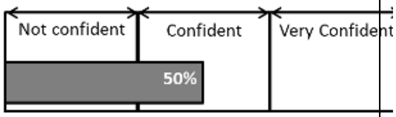
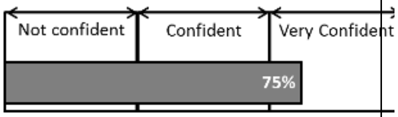
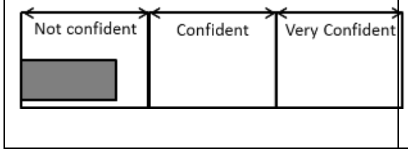
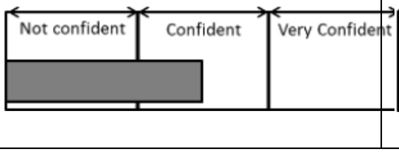
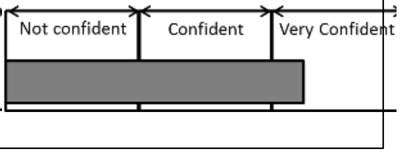
Providing system confidence information may not always provide benefits. In this study, system confidence information may have been less influential than reliability. Instead, it may be more beneficial to provide users with easy to evaluate alternative recommendations that can be verified against certain criteria. This better allows users to determine the DSS's reliability on a trial by trial basis and does not pose the potential of creating an uncoupled relationship between system reliability and system confidence. Furthermore, providing system confidence may be a costly endeavor. The amount of effort it would take to program an accurate algorithm to calculate the probability that the answer is correct may not help the user, especially if the system is already highly reliable.

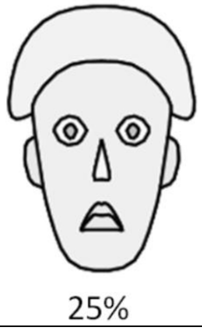
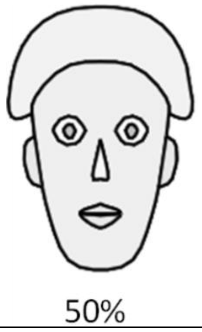
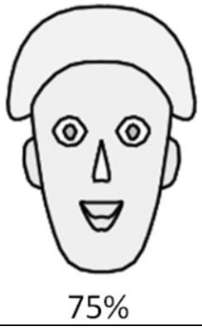
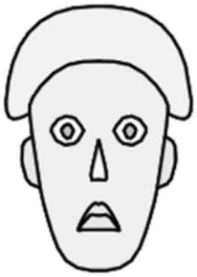
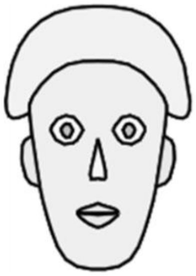
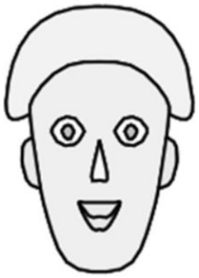
APPENDICES

Appendix A

Participant Handout of System Confidence Conditions

Image values		
25%	50%	75%
25%	50%	75%
		
		
		
		

 25%	 50%	 75%
		
 Not confident Confident Very Confident 25%	 Not confident Confident Very Confident 50%	 Not confident Confident Very Confident 75%
 Not confident Confident Very Confident	 Not confident Confident Very Confident	 Not confident Confident Very Confident

<table>
<tr><th colspan="3">Image values</th></tr>
<tr><td>25%</td><td>50%</td><td>75%</td></tr>
<tr><td> </td><td> </td><td> </td></tr>
<tr><td></td><td></td><td></td></tr>
</table>

Appendix B

Protocol for MX study 1

THINGS TO DO BEFORE SUBJECT ARRIVES

Day of:

1. Prepare the computers
   - A. Place IRB information sheet in front of each computer
   - B. Make sure the headphones are plugged in. Subjects can adjust volume during the abilities test.
   - C. On an experimenter computer, open \Desktop\Dropbox\Exp (1) then launch the dashboard program. Use the IP address that it specifies (it's the IP address of the experimenter computer)at each of the subject running computers.
   - D. Go back to the Exp1 folder, then open the "Margaux" folder
     - i. Open the file 'Margaux.exe'
       - a. Enter subject # (use next # from list)
       - b. Check to make sure the stimulus presentation time is set to 3000 & 5000
       - c. Click on the optional tab
       - d. Set the "Number of trials in Phase 1" to: 50 and "Block speed" to 100
       - e. Enter dashboard IP Address and set workstation # (on post it note at each participant station)
       - ii. Close the Dropbox folder window
   - E. Place a post-it note with the subject # at each station for your reference, also fill out the form for subject #'s.

TIPS:

- If subject accidently presses a key before you are ready, you can skip past the tetris game by restarting the program, but in the "Options" tab, set the "Number of trials in Phase 1" to 1. They will only have to match 1 set of blocks and it will move on to phase 2. You cannot skip phase 2.

- The 2nd and 3rd task both take approximately 8-9 minutes. You can time it so you know when subjects will be ready for the next step1

2. Go ahead and read the IRB information sheet on your desk and let me know if you have any questions.
3. Please go ahead and enter your age and gender, you can then click "Begin Study" but please do not click anything else until I tell you do so.
4. Ok, please listen as I guide you through the instructions and practice tasks. In today's study you will be completing a block matching task that is similar to the game tetris. You will also be identifying levels of images as they pop up while you play the block game. You will have the opportunity to practice both tasks before the experiment begins.

Block matching game practice instructions

5. Please click **"OK" and** follow along as I read aloud the instructions for the first task, a block matching game. In this game, you must match at least three blocks vertically or horizontally of the same color. But you can only switch any two blocks horizontally. Use the cursor keys (up, down, left, right) to move your selector. Press the space bar to switch blocks. Please work as quickly as you can to increase your score. When your score reaches 50 this practice will end automatically, but please do not start the next task. Go ahead and click the "Start Practice" button to begin.

Graphic identification task

6. [When everyone finishes]Ok, in this next task, you will notice on the right side of the screen that a graphic will appear in the white box. When you see the graphic you should identify the level represented by it. You will use the keys from 1 to 3 to indicate:

   1 = 25%   2 = 50%   3= 75%

   [Give participants the handout]. This handout shows the images that you will see in this task. As you can see, sometimes there will be numbers with the images and sometimes there won't be. In the first column are all of the images that represent 25%. The second column shows the images that represent 50%, and the last column has all images that represent 75%. Please take a moment to look this over and let me know if you have any questions. Once you think you have had enough time to look this over, please turn around and look at me so I know when everyone has had enough time. You can keep the handout for a reference if you would like, however you may not have time to look up the image before it disappears on the screen.
7. Does anyone have any questions? Ok, you should use your best judgment. When you are ready to begin the task, please click "Start practice". When you complete the practice, please STOP.

Dual task instructions

8. [After the practice]Now, you will do both tasks at the same time. That is, you will have the blocks game and the graphic task occurring at the same time. Like before, you will control the blocks game by using the cursor keys (up, down, left, right) and the space bar to switch any two blocks horizontally.

You will also identify the level of the graphic in the far right. Like before,

1 = 25%, 2 = 50% 3= 75%

Doing these two tasks at the same time is very challenging. Your main focus should be the blocks game. You should try to maximize your score as quickly as possible. Any reserve attention you have available should be used for the graphic task. Do you have any questions?

9. When you finish this task, please stop and let me know when you are done. Go ahead and click "start experiment"

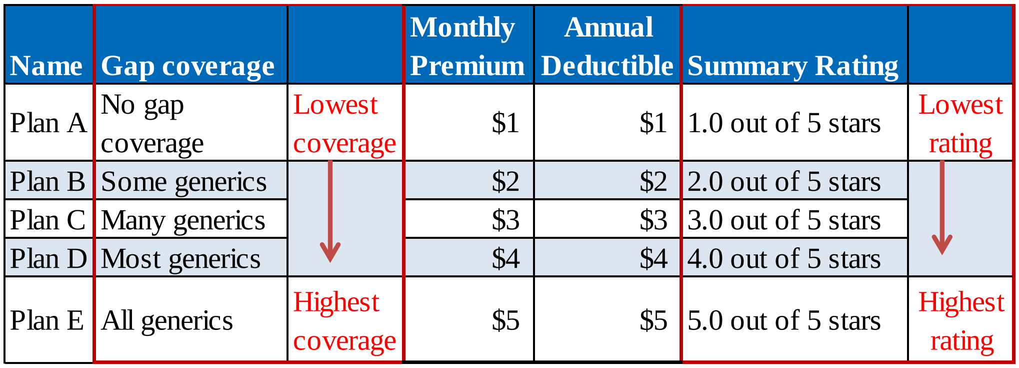
Abilities:

10. Click "Start" to get to demographics screen. After the demographics come up, click the **lower right corner.**
11. On the desktop, click on the shortcut to "Shortcut to Clip" then "run" (if it asks). Next, on the desktop, open "Shortcut to abilities.exe".
    A. Enter the subject #
    B. Enter Experiment ID as MxStudy1
    C. Make sure ONLY paper folding, cube comparison, and No RDS, MEM, Ship are checked
    D. Click ok
    E. Tell participant, "Please read through the instructions carefully and continue through until the computer tells you that you are finished. This will be the end of today's study."

Appendix C

Explanation of insurance terms used in the decision task and example table

Term	Definition
Gap Coverage	Gap coverage refers to the period of time after you and your plan have spent a certain amount of money for covered drugs when you have to pay out-of-pocket all costs for your drugs. Once you reach a set amount of out-of-pocket costs, your plan will begin coverage again. This term refers to the coverage provided during this "gap" in coverage.
Monthly Premium	The monthly premium is the set amount you must pay monthly.
Annual Deductible	The annual deductible or the amount you must pay for your prescriptions, before your drug plan begins to pay.
Satisfaction Rating	An overall score on the drug plan's quality and performance on customer service, member complaints, member experience, and pricing and patient safety



Name	Gap coverage		Monthly Premium	Annual Deductible	Summary Rating	
Plan A	No gap coverage	Lowest coverage	$1	$1	1.0 out of 5 stars	Lowest rating
Plan B	Some generics	↓	$2	$2	2.0 out of 5 stars	↓
Plan C	Many generics		$3	$3	3.0 out of 5 stars	
Plan D	Most generics		$4	$4	4.0 out of 5 stars	
Plan E	All generics	Highest coverage	$5	$5	5.0 out of 5 stars	Highest rating

Appendix D

Study #2 Protocol

Protocol for Dissertation Study #2

THINGS TO DO BEFORE SUBJECT ARRIVES

1. Determine which condition you will run (see the spreadsheet)
2. Prepare the computers
   A. Make sure all computers are set to 1280 x 1024 screen resolution (once these are set they do not need to be reset to the previous resolution)
   B. Go to the "Dropbox" folder on the desktop, click on the Margaux folder, then study 2 folder
   C. Now open the program by double clicking "mgx-study2.exe"
      i. Enter a subject number
      ii. Leave Age and Gender blank
      iii. Click "**Click to select a condition file**" button
      iv. When the open dialog box appears, select your condition file:
      v. Click the big red button before the subject arrives to hide the condition information.
   D. Place the sheet of paper with the plan information on the left of the keyboard and **[if anthro or eval cat conditions]** place the less confident/more confident sheet on the right of the keyboard.
   E. Place the consent form on top of the keyboard.

AFTER SUBJECTS HAVE ARRIVED AND HAVE BEEN SEATED <u>[ALL CONDITIONS]</u>

3. "Welcome to the study, thanks for coming in today. Please set your cell phones to silent. Go ahead and read the information sheet on your keyboard. When you are finished, please look at me so I know. Does anyone have any questions about the consent forms?
4. In today's study, you will be using the computer to make choose a prescription drug plan that best fits the criteria asked in the question. Some of the questions are designed to be very difficult. All we ask is that you try your best and guess if necessary. As you answer the questions, the computer will keep track of your score. As you get questions correctly and quickly, your score will increase. This will be indicated in a bar graph on your screen.

Your score is based on whether you get the question correct AND how quickly you respond. The quicker you make your response, the more points you get assuming you are correct.

5. Ok, first there are several terms that you may need to know in order to do the decision task. Please take a look at the sheet of paper in front of you. You can keep this out during the experiment as a reference if you need it.
    a. In the first column you will see the plan name. These are simply listed in alpha order by plan name.
    b. Gap coverage is in the second column. Gap coverage refers to the period of time after you and your plan have spent a certain amount of money for covered drugs when you have to pay out-of-pocket all costs for your drugs. Once you reach a set amount of out-of-pocket costs, your plan will begin coverage again. This term refers to the coverage provided during this "gap" in coverage. There are 5-levels of gap coverage. Here, the levels are presented in order. No gap coverage is the lowest amount, and all generics is the highest amount
    c. The "*monthly premium*" is the set amount you must pay monthly.
    d. The next is the "*annual deductible*", or the amount you must pay for your prescriptions, before your drug plan begins to pay. The dollar amounts will be in these columns.
    e. Finally, he satisfaction rating is in the last column and contains the rating out of a 5-point scale. This rating is an overall score on the drug plan's quality and performance on customer service, member complaints, member experience, and pricing and patient safety
    f. Do you have any questions about that? Ok, let's start the computer task. Please wiggle your mouse to get the screen to come up.

6. Please enter your age and gender. Next, please click the **"Begin Study"** button and then click **"YES**". This is the first of three practice trials. As you see in front of you, you see a large box on the left with a table of prescription drug plans.
7. On the lower portion of the screen you see a smaller question box that has the criteria that the plan you choose must meet.
8. On the lower left you see your timer bar. This slowly counts down 45 seconds. The more time you take in answering your questions, the fewer points you will earn toward your score even if you are correct.
9. In the upper right you see your overall score. As you answer questions correctly and quickly, your score will increase.

10. Remember, your task will be to find the prescription drug plan that best meets the criteria in the questions. There is an automated system to help you choose the correct plan.

11. You will be using two different systems today, system A and system B to answer a total of 72 questions. Both systems are mostly reliable. You will have answer 36 questions using each system and there will be a short break while we switch from one system to another. We will try three practice questions so that you can become familiar with your task. Do you have any questions before we begin the practice?"

12. Normally, you will have 45 second time limit, however these practice tasks are not timed.

[from here on follow directions for the condition you are running]

NO SYSTEM CONFIDENCE CONDITION: PAGE 3

SYSTEM CONFIDENCE DISPLAY CONDITIONS (NUMBER, GRAPH, EVALUATIVE CATEGORIES, OR ANTHRO FACE CONDITIONS): PAGE 4

[FOR NO SYSTEM CONFIDENCE CONDITION ONLY]

13. Let's go through a trial together. On the right, you see a suggestion from the system. It has analyzed the situation and has made a suggestion. Let's see how to use this system. Please read the scenario and the question but **do not** proceed. I'll give you a few seconds to read it.

    [wait to see that everyone has read it]

14. Ok. Let's say that you agree with the system's recommendation. Please press <u>AGREE</u>.
15. Next, please rate how confident you felt about the choice you just made. Since this is just practice you are probably not sure about your confidence level, but please answer anyway.
16. Also please rate how much you trust the automated system and its advice. Again, this is just practice and you may not know what you think about the aid right now, but please just answer. Using your mouse, click one of the squares that is closest to how you feel then click OK. Any of the squares can be clicked.
17. You can see that you get feedback that your choice was correct. Now, let's try another practice task. Please read the next question. Again, don't proceed until I say so.

[wait to see that everyone has read it]

Let's now disagree with the automation. Please click the disagree button.

18. Now, some multiple choices have appeared in the lower right. You can now select what you think is the correct answer. In this case, although the system suggested Plan J, let's say you think it is Plan A. Go ahead and select Plan A and click MAKE CHOICE.
19. Again, you see the rating scale about your confidence and trust. Please answer to the best of your ability.
20. Now you see feedback about your response to the question.
21. Now, let's go through one last practice question. Please read the scenario and stop after you read the question.

[wait to see that everyone has read it]

22. Now, say you agree with the automation but you want to make absolutely sure so you want to see what your options could be.
23. In this case, click the PEEK at options button. Here, you can see the multiple choice options without yet agreeing or disagreeing with the system. Remember, if you peek, you get 10 seconds removed from your clock. Clicking on the answers listed on the peek dropbox will not answer the question, answers can only be individually chosen after clicking disagree.
24. After peeking, you eventually agree with the aid so go ahead and click Agree.
25. Now, rate your confidence and trust and click OK.
26. Do you have any questions before you begin the study? From this point, you will now be answering questions using the computer. If you have any questions during the study, I will be sitting right here. When you are done with the first set of 36 questions the computer will notify you. At the end of the second set, there are additional surveys you'll need to complete, so please let me know so I can get you started.

[FOR THE SYSTEM CONFIDENCE CONDITIONS ONLY]

14. In the top right box you will see a display that shows the system's confidence in its suggestion. System confidence is "a system generated estimation of whether it is providing a correct suggestion" and that it "is based on the quality of the information it is using for providing a suggestion, on a trial by trial basis. If the

system has low confidence, it is likely that it does not have much information to help it make a decision, the information is degraded, or there are other choices that are very similar and hard to choose between. If the system has high confidence, it is likely that the system has plenty of good information and that one prescription drug plan is much better than the others". There is a sheet on the right side of your desk that shows the 3 different confidence images you may see. Please take a moment to look this over and let me know if you have any questions.

15. Let's go through a trial together. This is the first of three practice trials. As you see in front of you, you see a large box on the left with a table of prescription drug plans. On the lower portion of the screen you see a smaller question box about the plans presented above it.
16. On the lower left you see your timer bar. This slowly counts down 45 seconds. The more time you take in answering your questions, the less points you will earn toward your score even if you are correct.
17. In the upper right you see your overall score. As you answer questions correctly and quickly, your score will increase.
18. Please read the scenario and the question but do not proceed. I'll give you a few seconds to read it.
19. On the right, you see a suggestion from the system. It has analyzed the situation and has made a suggestion. Let's see how to use this system. Please read the scenario and the question but **do not** proceed. I'll give you a few seconds to read it.

[wait to see that everyone has read it]

20. Ok. Let's say that you agree with the system's recommendation. Please press AGREE.
21. Next, please rate how confident you felt about the choice you just made. Since this is just practice you are probably not sure about your confidence level, but please answer anyway.
22. Also please rate how much you trust the automated system and its advice. Again, this is just practice and you may not know what you think about the aid right now, but please just answer. Using your mouse, click one of the squares that is closest to how you feel then click OK. Any of the squares can be clicked.
23. You can see that you get feedback that your choice was correct. Now, let's try another practice task. Please read the next question. Again, don't proceed

until I say so.

[wait to see that everyone has read it]

Let's now disagree with the automation. Please click the disagree button.

24. Now, some multiple choices have appeared in the lower right. You can now select what you think is the correct answer. In this case, although the system suggested Plan J, let's say you think it is Plan A. Go ahead and select Plan A and click MAKE CHOICE.
25. Again, you see the rating scale about your confidence and trust. Please answer to the best of your ability.
26. Now you see feedback about your response to the question.
27. Now, let's go through one last practice question. Please read the scenario and stop after you read the question.

[wait to see that everyone has read it]

28. Now, say you agree with the automation but you want to make absolutely sure so you want to see what your options could be.
29. In this case, click the PEEK at options button. Here, you can see the multiple choice options without yet agreeing or disagreeing with the system. Remember, if you peek, you get 10 seconds removed from your clock. Clicking on the answers listed on the peek dropbox will not answer the question, answers can only be individually chosen after clicking disagree.
30. After peeking, you eventually agree with the aid so go ahead and click Agree.
31. Now, rate your confidence and trust and click OK.

[wait to see that everyone has read it]

32. Do you have any questions before you begin the study? From this point, you will now be answering questions using the computer. If you have any questions during the study, I will be sitting right here. When you are done with the first set of 36 questions the computer will notify you. At the end of the second set, there are additional surveys you'll need to complete, so please let me know so I can get you started.

[after they finish the first round]

33. Ok, do you want to take a short break while I set up the other system? I just need to enter some things into the computer to get it started.
    [look at the subject list to determine which csv file to load. Enter in subject #, and hide the options. Have the participant enter age and gender]
    Before you begin, you may get a message after trial 3 that says the practice is over. Just ignore the message and click ok to continue. When you are done, let me know and I will have you do a few surveys.

    [when done with block 2, open up the shortcut to 2.Tech Exp survey on the desktop. Enter subject #]
34. Ok, please fill this out and let me know when you are finished.

[when finished – open up abilities, enter subject # and make sure that paper folding and cub comparison, and "no RDS, MEM" is checked].

35. Ok, this is the last thing you have to do. Keep following the instructions until the computer tells you that you have finished everything. If there is a continue button, please just keep going. Let me know if you have any questions along the way.

[when finished…]

That is the end of the experiment, thank you so much for coming in today!

REFRENCES

Breslow, L.A., Ratwani, R.M., &Trafton, J.G. (2009). Cognitive models of the influence of color scale on data visualization tasks. *Human Factors*, 51 (3), 321-338.

Chernoff, H. (1973). The use of faces to represent points in k dimensional space graphically. *Journal of the American Statistical Association*, 68(342), 361-368.

Damasio, A.R., Tranel D., & Damasio, H. (1990). Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli. *Behavioural Brain Research*, 41(2), 81-94.

Ekman, P. (1999). Basic emotions. In T. Dagleish and M. Power. (Eds.), Handbook of Cognition and Emotion, New York: Wiley.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Manual for kit of factor- referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Fisk, A. D., Derrick, W. L., & Schneider, W. (1986). A methodological assessment and evaluation of dual-task paradigms. *Current Psychology*, 5(4), 315–327.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), Advances in experimental social psychology. New York: Academic Press.

Gigerenzer, G., & Goldstein, D.G. (1996). Reasoning the fast and frugal way: Models of Bounded Rationality. *Psychological Review,* 103(4), 650-669.

Gillan, D.J., Wickens, C.D., Hollands, J.G., & Carswell, C.M. (1998). Guidelines for presenting quantitative data in HFES publications. *Human Factors,* 40(1), 28-41.

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, 24, 14194-1509.

Hancock, P.A., & Parasuraman, R. (1992). Human Factors and Safety Issues in Intelligent Vehicle Highway Systems (IVHS). *Journal of Safety Research*, 23, 181-198.

Hess, U., Blairy, S., & Kleck, R. E. (1997). The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior*, 21(4), 241–257.

Hilburn, B., Jorna, P.G., Byrne, E.A., & Parasuraman, R. (1997). The effect of Adaptive Air Traffic Control (ATC) decision aiding on controller mental workload. In M. Mouloua & J.Koonce (Eds.), Human-automation interaction: Research and Practice. Mahwah, New Jersey: Erlbaum.

Hink, J.K., Wogalter, M.S., & Eustace, J.K. (1996). Display of quantitative information: Are grables better than plain graphs and tables? *Proceedings of the Human Factors and Ergonomics Society,* 40, 1155-1159.

Hunter, I.M., Whiddett, R.J., Norris, A.C., McDonald, B.W., & Waldon, J.A. (2009). New Zealander's attitudes towards access to their electronic health records: Preliminary results from a national study using vignettes. *Health Informatics Journal,* 15(3), 212-228.

Huynh, H. & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics,* 1, 69-82.

Kramer, R.M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, 50, 569-598.

Lee, J. D. (2006). Affect, attention, and automation. In A. Kramer, D. Wiegmann & A. Kirlik (Eds.), Attention: From theory to practice. New York: Oxford University Press.

Lee, J. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies,* 40, 153-187.

Lee, J.D., & See, K.A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors,* 46(1), 50-80.

McGuirl, J.M., & Sarter, N.B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors,* 48(4), 656-665.

Merritt, S.M., & Ilgen, D.R. (2008). Not all trust is created equal: Dispositional and History-based trust in human-automation interactions. *Human Factors,* 50(2), 194-210.

Meyer, J., Shinar, D., & Leiser, D. (1997). Multiple factors that determine performance with tables and graphs. *Human Factors*, 39(2), 268-286.

Muir, B.M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27, 527-539.

Norman, D.A. & Bobrow, D.G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology,* 7, 44-64.

Oh, I. & Stone, M. (2007). Understanding RUTH: Creating believable behaviors for a virtual human under uncertainty. *Digital Human Modeling*, 4561, 443-452.

Ohman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear learning. *Psychological Review*, 108(3), 483-522.

Ortony, A. & Turner, T.J. (1990). What's basic about basic emotions? *Psychological Review,*97(3), 315-331.

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence decision-support use and performance in younger and older adults. *Ergonomics*, 55(9), 1059-1072.

Parasuraman, R. & Hancock P.A. (2008). Mitigating the adverse effects of workload, stress, and fatigue with adaptive automation. In P.A. Hancock & J.L. Szalma (Eds.) Performance under stress. Aldershot, England: Ashgate Publishing.

Parasuraman, R., & Manzey, D.H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors,* 52(3), 381-410.

Parasuraman & Miller (2004). Trust and etiquette in high-criticality automated systems. C*ommunications of the ACM,* 47(4), 51-55.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors,* 39(2), 230-253.

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* 30(3), 286-297.

Peters, E., Dieckmann, N.F., Mertz, C.K., Vastfjall, D., Slovic, P., & Hibbard, J.H. (2009). Bringing meaning to numbers: The impact of evaluative categories on decisions. *Journal of Experimental Psychology: Applied,* 15(3), 213-227.

Price, M.M., & Pak, R. (accepted). Complex decision support for older adults: Effects of information visualization on decision performance. *International Journal of Industrial Ergonomics: Special Issues on Healthcare Information Technology and Ergonomics.*

Sanchez, J. (2006). Factors that affect trust and reliance on an automated aid. (Doctoral dissertation). Retrieved from smartech.gatech.edu. (http://hdl.handle.net/1853/10485).

Shipley, W. (1986). Shipley Institute of Living Scale. Los Angeles: Western Psychological Press.

Simon, H.A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics,* 69, 99-118.

Singh, I.L., Molloy, R., & Parasuraman, R. (1993). Automation-induced “complacency”: Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology,* 3(2), 111-122.

Slovic, P., Peters, E., Finucane, M.L., & MacGregor, D.G. (2005). Affect, risk, and decision making. *Health Psychology,* 24(4), S35-S40.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397-420). New York: Cambridge University Press.

Spain, R.D., & Bliss, J.P. (2009). The effects of automation expertise and system confidence on trust behaviors. *Proceedings of the Human Factors and Ergonomics Society*, 53, 344-348.

Spain, R.D., & Madhavan, P. (2009). The role of automation etiquette and pedigree in trust and dependence. *Proceedings of the Human Factors and Ergonomics Society,* 53, 339-343.

Takeuchi, A. & Nagao, K. (1993). Communicative facial displays as a new conversational modality. *Proceedings of INTERCHI '93 conference on Human Factors in computing systems*, 187-193.

Tufte, E.R. (1983). The visual display of quantitative information, Cheshire, CT: Graphic Press.

Todorov, A. (2008). Evaluating Faces on Trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences*, 1124, 208-224.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review,* 79, 281-299.

Tversky, A. & Kahneman. D. (1974). Judgment under uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124-1131.

Walker, J., Sproull, L., & Subramani, R. (1994). Using a human face in an interface. CHI'94, 85-91.

Wechsler, D. (1997). Wechsler Adult Intelligence Scale (3rd ed.). San Antonio, TX: Psychological Corp.

Wiegmann, D. A., Rich, A., and Zhang, H. (2001). Automated Diagnostic Aids: The Effects of Aid Reliability on Users' Trust and Reliance. *Theoretical Issues in Ergonomics Science*, 2(4), 352-367.

Winston, J.S., Strange, B.A., O'Doherty, J.O., & Dolan, R.J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience,* 5(3), 277-283.

Zacks, J. & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory and Cognition*, 27(6), 1073-1079.