Flinders
UNIVERSITY

Abstract

Objective: A series of experiments examined human operators' strategies for interacting with highly (93%) reliable automated decision aids in a binary signal detection task.

Background: Operators often interact with automated decision aids in a suboptimal way, achieving performance levels lower than predicted by a statistically ideal model of information integration. To better understand operators' inefficient use of decision aids, the current study compared participants' automation-aided performance levels to the predictions of seven statistical models of collaborative decision making.

Method: Participants performed a binary signal detection task that asked them to classify random dot images as either blue- or orange-dominant. They made their judgments either unaided or with assistance from a 93%-reliable automated decision aid that provided either graded (Experiments 1 and 3) or binary (Experiment 2) cues. Analysis compared automation-aided performance to the predictions of seven statistical models of collaborative decision making, including a statistically optimal model (Sorkin & Dai, 1994) and Robinson and Sorkin's (1985) contingent criterion model.

Results and conclusion: Automation-aided sensitivity hewed closest to the predictions of the two least efficient collaborative models, well short of statistically ideal levels. Performance was similar whether the aid provided graded or binary judgments. Model comparisons identified potential strategies by which participants integrated their judgments with the aid's.

Application: Results lend insight into participants' automation-aided decision strategies, and provide benchmarks for predicting automation-aided performance levels.

Keywords: human-automation interaction, signal detection theory, decision-making strategies, contingent criterion model

Benchmarking Aided Decision Making in a Signal Detection Task

Human operators in everyday and professional contexts work with the assistance of automated decision aids. The assisted tasks often take the form of binary signal detection judgments, which ask a decision maker to classify potentially ambiguous states of the world into either of two discrete categories (Green & Swets, 1966; Macmillan & Creelman, 2005). A credibility assessment aid, for instance, might help organizational decision makers distinguish deceptive from honest responses when questioning interviewees in negotiations or investigations (Jensen, Lowry, & Jenkins, 2011). Analogously, a combat identification system might help soldiers distinguish friends from foes on the battlefield (Wang, Jamieson, & Hollands, 2009). Ideally, assistance from an automated aid will help the human operator to achieve higher levels of *sensitivity*, the ability to distinguish between states of the world. But like the human operator, an automated decision aid performing a signal detection task is typically required to render judgments based on incomplete or uncertain data. The aid's sensitivity will therefore be imperfect, just as the human operator's is, and the aid's judgments will sometimes be wrong.

Imperfect sensitivity does not render an aid inherently useless. Even if the automation errs in occasional judgments, the human operator may be able to achieve a higher sensitivity with the aid's assistance than without it (Wickens & Dixon, 2007). In practice, unfortunately, people often interact with automated aids in a suboptimal way. This may manifest as either *misuse*, a tendency to act on the aid's judgments uncritically, or *disuse*, a tendency to disregard or underweight the aid's judgments (Parasuraman, 2000; Parasuraman & Riley, 1997). These effects compromise the benefits of automated assistance, and in the worst case, operators may even perform a task more poorly when assisted by a decision aid than when unassisted (e.g., Alberdi, Povyakalo, Strigini, & Ayton, 2004).

An important goal of automation design is therefore to encourage more efficient

human-automation interaction, allowing the automation-aided operator to achieve higher levels of sensitivity. Notably, automation-aided performance in a signal detection task can be conceptualized as a form of collaborative decision making in which two agents, the human and the aid, reach separate judgments about the state of the world and then combine their judgments to reach a joint decision (Sorkin & Dai, 1994; Sorkin, Hays, & West, 2001). Understanding the process by which the operator integrates his or her own judgment with that of the aid may thus allow practitioners to better tailor the design of automated aids, to encourage efficient human-automation collaboration. In the worst case, it will allow system designers to better predict automation-aided performance levels.

Using binary cues: The contingent criterion model

Many studies of automation-aided decisions have specifically considered the case in which an aid provides the human operator binary judgments (e.g., Botzer, Meyer, Pak, & Parmet, 2010; Dzindolet, Pierce, Beck, Dawe, & Anderson, 2001; Rice & McCarley, 2011). Robinson and Sorkin's (1985) *contingent criterion* (CC) model has become the modal account of human-automation interaction under these circumstances (e.g., Elvers & Elrif, 1997; Maltz & Meyer, 2001; Meyer, 2001). The model is built on the framework of signal detection theory (SDT) (Green & Swets, 1966; Macmillan & Creelman, 2005). SDT assumes that to render a signal detection judgment, the decision maker first encodes evidence for or against either of two exhaustive and mutually exclusive potential states of the world, one of which is conventionally termed *signal* and the other *noise*. The evidence values are distributed continuously, and unless the task is trivially easy, the evidence distributions corresponding to the two states of the world overlap at least partially.

The decision maker transforms continuous evidence values into discrete judgments by comparing them to a response criterion. Values below the criterion value lead to a judgment of signal absent, and values above it lead to a judgment of signal present. The decision

maker's criterion may be *conservative*, biased toward judgments of noise; *liberal*, biased toward judgments of signal; or *unbiased*. Assuming that the signal and noise evidence distributions are Gaussian with a common standard deviation (Macmillan & Creelman, 2005), sensitivity can be measured by the statistic *d'*,

$$d' = \mathrm{z}\,(HR) - \mathrm{z}\,(FAR),$$

and bias by the statistic *c*,

$$c = -0.5 \times [\mathrm{z}\,(HR) - \mathrm{z}\,(FAR)].$$

A value of $d' = 0$ indicates chance performance, and a value of $d' = 5$ indicates near perfect performance. Negative values of *c* indicate liberal bias, positive values indicate conservative bias, and a value of 0 indicates unbiasedness.

The CC model views the aid and the human operator as operating in sequence, with the aid rendering its judgment first and the operator establishing his or her own response criterion contingent on the aid's judgment. The operator is thus presumed to operate with a relatively liberal response criterion following a judgment of signal present from the aid, and with a relatively conservative criterion following a judgment of signal absent from the aid. For ease of exposition, we will refer to a signal present judgment as*Yes* and a signal absent judgment as *No*. Team hit rate under the CC model, HR_{CC}, is,

$$HR_{CC} = HR_{aid}\ (HR_{operator|\text{“Yes”}}) + (1 - HR_{aid})\ HR_{operator|\text{“No”}},$$

where HR_{aid} is the hit rate of the automated aid, $HR_{operator|\text{“Yes”}}$ is the hit rate of the unaided human operator given a *Yes* judgment from the aid, and $HR_{operator|\text{“No”}}$ is the hit rate of the unaided human operator given a *No* judgment from the aid. Team false alarm rate under the CC model, FAR_{CC}, is,

$$FAR_{CC} = FAR_{aid}\ (FAR_{operator|\text{“Yes”}}) + (1 - FAR_{operator|\text{“No”}})\ FAR_{operator|\text{“No”}},$$

where FAR_{aid} is the false alarm rate of the automated aid, $FAR_{operator|\text{“Yes”}}$ is the false alarm rate of the unaided human operator given a *Yes* judgment from the aid, and $FAR_{operator|\text{“No”}}$ is

the false alarm rate of the unaided human operator given a *No* judgment from the aid. Team sensitivity under the CC model, d'_{CC}, is thus,

$$d'_{CC} = z\,(HR_{CC}) - z\,(FAR_{CC}).$$

The operator's optimal criterion setting following an aid's judgment is determined by the aid's predictive value (Robinson & Sorkin, 1985). Assuming an unbiased payoff matrix, normative bias following a response *i* from the aid, as measured by the statistic β, is,

$$\beta_{optimal} = [1\text{-}p(\text{signal}|i)] \,/\, p(\text{signal}|i),$$

where *i* is either a *Yes* or a *No*. Normative behavior thus entails larger bias shifts in response to more reliable automated aids. Data have shown that operators' response criteria in fact shift in the expected direction following a *Yes* or *No* judgment from an aid, but that the magnitude of these shifts is smaller than predicted by the normative CC model (Elvers & Elrif, 1997; Meyer, 2001; Wang et al., 2009). These findings have been taken as evidence that operators employ a CC strategy in automation-aided decision tasks, but choose their criteria suboptimally (cf., Botzer et al., 2010).

But while evidence for suboptimal automation use is incontestable, evidence that this is the result of a CC process is more tentative. Bias shifts in the direction of an aid's recommendation are consistent with a CC strategy. Other information integration strategies, however, will also produce differences in response bias conditional on the aid's decision. In fact, any collaborative strategy under which the operator tends to agree with the aid will engender differences in the operator's bias conditional on the aid's decision. Differences in conditional operator bias therefore do not necessarily implicate the decision process postulated by the contingent criterion model. Additionally, the suboptimal CC model by itself offers little help in anticipating the performance benefits that an automated aid will produce. While aided performance will be less than statistically ideal, the model does not specify just how far short of that standard it will fall. Phrased differently, whereas the operator's cued

criterion settings are fixed in the optimal CC model, the suboptimal model makes them free parameters, providing little *a priori* basis for predicting the operator's automation-aided sensitivity. Comparing automation-aided performance to the predictions of alternative, fixed-parameter or parameter-free models may therefore be useful both to identify strategies that provide plausible alternative accounts of human-automation decision making, and to establish benchmarks that help designers predict the performance levels automation-aided operators might attain.

Alternative models of binary cue use

A very simple strategy for interacting with an automated aid, proposed as a potential strategy for collaborative decision making between pairs of human decision makers, is the *best decides* (BD) model (Bahrami et al., 2010; Denkiewicz, Rączasek-Leonardi, Migdal, & Plewczynski, 2013). This model assumes that the human operator knows whether he or she is more or less sensitive than the aid. If more sensitive, the operator ignores the aid entirely and makes a judgment each trial for him or herself. If less sensitive, the operator defers to the aid's judgments by default. Team sensitivity under the BD model, d'_{BD}, is thus,

$$d'_{BD} = \max (d'_{operator}, d'_{aid}).$$

Although simpler than the CC strategy, the BD strategy makes far less efficient use of the paired decision makers' judgments, producing lower levels of automation-aided sensitivity. Nonetheless, observed automation-aided performance is often poorer still than predicted by the BD model (e.g., Meyer, 2001; Rice & McCarley, 2011).

Another pair of strategies, the *yes/yes* (YY) and *no/no* (NN) decision models proposed by Pollack and Madans (1964), are also inefficient, but again seem to outperform human-automation teams. Under the YY model, both the operator and the aid must report "signal present" for the team to produce a collaborative signal present judgment. Conversely, under the NN model, both the operator and the aid must report "signal absent" to produce a

collaborative signal absent judgment. Since the YY and NN decision models make symmetrical predictions, we will only discuss and report the predictions of the NN model. Team hit rate under the NN model, HR_{NN}, is,

$$HR_{NN} = 1- (1-HR_{operator}) (1-HR_{aid}),$$

where $HR_{operator}$ is the hit rate of the unaided human operator. Team false alarm rate under the NN model, FAR_{NN}, is,

$$FAR_{NN} = 1- (1-FAR_{operator}) (1-FAR_{aid}),$$

where $FAR_{operator}$ is the false alarm rate of the unaided human operator. Team sensitivity under the NN model, d'_{NN} is thus,

$$d'_{NN} = z (HR_{NN}) – z (FAR_{NN}),$$

and team criterion under the NN model, c_{NN}, is,

$$c_{NN} = -1/2 [z (HR_{NN}) + z (FAR_{NN})].$$

Pollack and Madans (1964) found that automation-aided participants achieved sensitivity levels lower than predicted by the NN and YY models.

Adapted to the context of human-automation decision making, Bahrami et al.'s (2010) *coin flip* (CF) model might provide a more plausible and better-fitting process model of human-automation performance. The model assumes that if the human operator and aid agree on a yes-or-no judgment, that's the judgment of the team. If they reach different decisions, the disagreement is effectively resolved by coin flip, that is, by selecting among the two response options randomly and with equal probability. The model thus posits discrete states in which the operator either ignores the model's judgment or defers to it fully. Predictions for the CF model in the current work can be made by estimating team hit rate (*HR*) and false alarm rate (*FAR*) from the individual team member's HR and FAR, then transforming those scores using the standard equation for calculating *d'*. Assuming that the human operates with the same response bias under individual and automation-aided conditions, team hit rate under

the CF model, HR_{CF}, is,

$$HR_{CF} = (HR_{operator})\,(HR_{aid}) + 0.5\,(HR_{operator})\,(1\text{-}HR_{aid}) + 0.5\,(1\text{-}HR_{operator})\,(HR_{aid})$$

$$= 0.5(HR_{operator} + HR_{aid}),$$

Team false alarm rate under the CF model, FAR_{CF}, is,

$$FAR_{CF} = (FAR_{operator})\,(FAR_{aid}) + 0.5\,(FAR_{operator})\,(1\text{-}FAR_{aid}) + 0.5\,(1\text{-}FAR_{operator})\,(FAR_{aid}),$$

Team sensitivity under the CF model, d'_{CF}, is,

$$d'_{CF} = z\,(HR_{CF}) - z\,(FAR_{CF}),$$

and team criterion under the CF model, c_{CF}, is,

$$c_{CF} = -1/2\,[z\,(HR_{CF}) - z\,(FAR_{CF})].$$

Because the CF model reflects a highly inefficient strategy for combining agents' judgments (Bahrami et al., 2010), it may offer a more plausible account of human-automation collaboration than the models discussed above. Alternatively still, we may consider a model that is similar but potentially more consonant with empirical findings in the study of decision making. Like the CF model, the *probability matching* (PM) model, posits that yes-or-no disagreements between agents are resolved randomly. The PM model, however, assumes that the operator defers to the aid's judgment with a probability equal to the aid's average reliability, mimicking a strategy that participants use in probabilistic choice tasks (see Koehler & James, 2014; Vulkan, 2000, for reviews), including automation-aided decision tasks in which operators have no access to raw data (Bliss, Gilson, & Deaton, 1995; Wiegmann, 2002). Team hit rate under the PM model, HR_{PM}, is,

$$HR_{PM} = R_{aid} \times HR_{aid} + (1\text{-}\,R_{aid}) \times HR_{operator},$$

where R_{aid} is the aid's average reliability rate. Team false alarm rate under the PM model, FAR_{PM}, is,

$$FAR_{PM} = R_{aid} \times FAR_{aid} + (1\text{-}\,R_{aid}) \times FAR_{operator},$$

Team sensitivity under the PM model, d'_{PM} is,

$$d'_{PM} = z\,(HR_{PM}) - z\,(FAR_{PM}),$$

and team criterion under the PM model, c_{PM} is,

$$c_{PM} = -0.5 \times [z\,(HR_{PM}) - z\,(FAR_{PM})].$$

The CF and PM models can be considered variants of the same discrete-state model, differing only in the fixed probability with which the operator defers to the aid. Assuming the automated aid's decisions are more accurate on average than the operator's, the PM model offers an aided decision strategy more efficient than the CF model but nonetheless suboptimal.

Strategies for using direct evidence values

As noted, the models discussed above presume an aid rendering yes-or-no judgments. Phrased differently, they presume an aid that measures the strength of evidence for a signal and then applies a decision rule to transform that strength estimate into a binary judgment. Some empirical studies have examined variations on this design in which the aid renders confidence-graded judgments on a scale of more than two levels, providing a more fine-grained assessment of the evidence for or against a signal, but even in these cases the aid's judgments have been discretized. Automated aids in one study, for example, provided participants alarms on a 4-level scale, where the lowest level was the absence of a signal and the highest level denoted an urgent alarm (Sorkin, Kantowitz, & Kantowitz, 1988). A visual search aid in another study ranked potential target locations on a 5-level scale (St. John & Manes, 2002). Both of these studies found evidence for better human performance with graded than with binary automated cues, as have some others (Andre & Cutler, 1998; Gupta, Bisantz, & Singh, 2002; McCarley, 2009; Wiczorek & Manzey, 2014). Other research, however, has failed to replicate this benefit (Wickens & Colcombe, 2007; Wiczorek, Manzey, & Zirk, 2014).

An alternative and less-explored design option is to allow the aid to share its evidence

estimates directly. By preserving information that is lost when responses are discretized, such direct evidence sharing offers the potential of better human-automation performance than is achievable with standard, discrete judgments from an aid (Bahrami et al., 2010). The *optimal weighting* (OW) model (Bahrami et al., 2010; Sorkin & Dai, 1994; Sorkin et al., 2001), built on the assumption of direct evidence sharing from the aid, in fact offers the strategy for best-possible automation-aided performance. The model assumes that the human and the automated aid both operate as equal-variance Gaussian signal-detectors (Macmillan & Creelman, 2005). Each trial, both agents assess the stimulus independently and estimate the likelihood that it contains a signal. The automated aid reports its likelihood estimate to the human operator, with the automation-aided decision based on a weighted average, Z, of these estimates,

$$Z = \Sigma a_i X_i,$$

where i indexes the agent, human or automation, a_i is the weight applied to agent i's estimate, and X_i is that estimate. Assuming the human and aid's judgments are stochastically independent, the optimal weight for agent i is proportional to the agent's sensitivity, d'_i. In the context of automation-aided decision making, team sensitivity under the OW model, d'_{OW}, is,

$$d'_{OW} = (d'_{operator} + d'_{aid})^{1/2}.$$

Another model for using direct evidence judgments from the aid, the *uniform weighting* (UW) model, is identical to the OW model except that it assumes that the operator assigns equal weights to the two estimates of signal likelihood when averaging them, i.e., that $a_{human} = a_{aid}$ (Sorkin et al., 2001). In this case, team sensitivity under the UW model, d'_{UW}, is,

$$d'_{UW} = (d'_{operator} + d'_{aid})/2^{1/2}.$$

If the aid and operator are equally sensitive, the UW model is equivalent to the OW model. Otherwise, d'_{UW} is lower than d'_{OW}.

Comparing the performance of the OW and UW models to the performance of the

models discussed above suggests that human operators may benefit more from an aid that shares its evidence assessments directly, without discretizing responses. As yet, though, this possibility apparently has not been tested empirically.

The current experiments

The models above span a range of performance levels, from perfectly efficient to highly inefficient. The present series of experiments tested the performance of automation-aided decision makers in a two-alternative forced choice (2AFC) task against the models to investigate human operators' strategies for interacting with an automated decision aid, and to benchmark empirical automation-aided performance. Participants viewed orange and blue random-dot images, and were asked to determine each trial which color was dominant (Voss, Rothermund, & Voss, 2004). They performed the task alone or with assistance from an automated decision aid. The aid rendered its judgment either in the form of a binary diagnosis accompanied by an estimate of signal strength (Experiments 1 and 3), or simply as a binary diagnosis (Experiment 2). The predictions for each collaborative model were calculated from the participant's unaided sensitivity and the sensitivity of the aid. Observed collaborative sensitivity values were then compared to the statistically optimal values predicted by each model.

This research complied with the tenets of the Declaration of Helsinki and was approved by the Social and Behavioural Research Ethics Committee at Flinders University. Informed consent was obtained from all participants.

Experiment 1

Method

Participants. Participants were 40 adults (mean age = 20.97 years, *SD* = 3.76, range = 17-35; 34 females, 6 males) recruited from the Flinders University of South Australia. All participants were compensated with $10.00 AUD for an experimental

session that lasted approximately 45 min. Participants were fluent in English, had normal color vision, and normal or corrected-to-normal visual acuity.

Apparatus and Stimuli. The experimental task was controlled by E-prime (Psychology Software Tools, Inc., Pittsburgh, PA), and stimuli were presented on a 23-inch Samsung monitor with a resolution of 1,920 x 1,080 pixels and a 120 Hz refresh rate. Participants were seated approximately 60 cm from the monitor, with viewing distance unconstrained.

Stimuli were 300 blue and orange random dot images (256 x 256 pixels). Figure 1 shows a sample orange-dominant stimulus image. Each stimulus was either blue-dominant or orange-dominant. In the blue-dominant stimuli, each pixel was randomly assigned the color blue with a probability of 0.52 or the color orange with a probability of 0.48. In the orange-dominant stimuli, those probabilities were reversed.

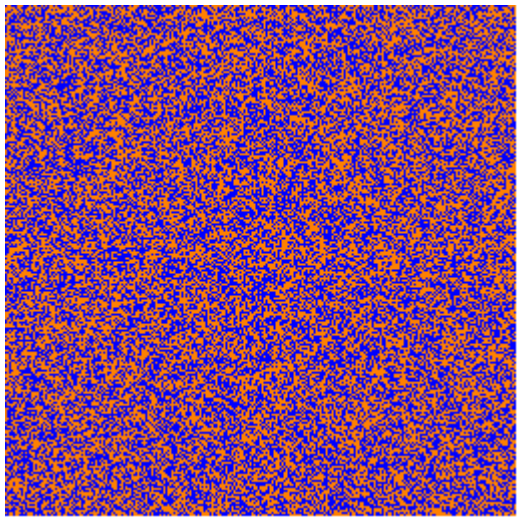

Figure 1. A sample orange-dominant stimulus image.

Procedure. Participants performed a 2AFC task requiring them to classify stimulus images as blue- or orange-dominant. A cover story asked the participants to imagine themselves as geologists sorting samples of a fictional mineral Vibranium into blue and orange strains. The instructions informed them, “Unfortunately, the two

strains are difficult to tell apart. Both are speckled blue and orange. The only difference visually is that one strain tends to have a little more orange, and the other tends to have a little more blue. For simplicity, we will call them VBN-ORANGE and VBN-BLUE. However, there is a lot of overlap in their appearance, and it is almost impossible to sort them with 100% accuracy by eye." Participants were asked to press the number 1 on the keyboard if they thought the image was mostly orange, and to press the number 3 on the keyboard if they thought the image was mostly blue.

Participants were also told that on some trials, they would be assisted by an automated decision aid that would provide a binary blue or orange judgment along with an estimate of signal strength. Instructions read, "The aid works by testing the chemical properties of the sample, and then assessing whether the sample is more likely to be VBN-ORANGE or VBN-BLUE. However, just like a human judge, the aid can sometimes make mistakes; testing has shown that on average, the aid is correct 93% of the time and incorrect 7% of the time. To help you predict whether it is right or wrong, the aid will give its assessment along with a numeric rating each trial. A higher rating means that the aid is more likely to be correct. The aid will provide its assessment and rating at the start of each trial. You should use the aid to help you make your decisions, but be aware that you are free to disagree with it any time you wish. Use your own best judgement."

The aid's judgments were calculated using an equal-variance Gaussian signal detection model. Evidence values for blue-dominant images were sampled from a Gaussian distribution with a mean of -1.5 and a standard deviation of 1, and evidence values for orange-dominant images were sampled from a Gaussian distribution with a mean of 1.5 and a standard deviation of 1. Thus, the *d'* of the aid was 3. The aid transformed evidence values into binary judgments using an unbiased response

threshold, offering a judgment of blue-dominant if the evidence value sampled for a given trial was less than 0 and a judgment of orange-dominant if the evidence value sampled was greater than 0. Given the aid's *d'* of 3, the unbiased criterion produced an average accuracy rate of 93%. The aid's estimate of signal strength was simply the absolute value of the sampled evidence value. As noted, participants were informed that a higher value indicated stronger evidence. Because they were generally not expected to have had extensive formal training in statistics, however, they were not provided any additional information about the distribution of evidence values.

Figure 2 shows the sequence of events within an automation-aided trial for Experiment 1. Each trial was initiated with a key press from the participant. This was followed by a 1,000-ms fixation screen, a 1,000-ms screen displaying the automated aid's diagnosis, and then the stimulus display. On aided trials participants were provided with the aid's diagnosis, e.g., "Aid judges: Orange 2.14." On unaided trials, participants were instead provided with a neutral message, "Waiting for sample." Presentation of the aid's diagnosis before the stimulus display allowed participants time to attend to the diagnosis carefully, and ensured that the diagnosis and stimulus arrived in the same order in which the CC model presumes they are processed (though see Wiegmann, McCarley, Kramer, & Wickens, 2006, for evidence that automation dependence is similar regardless of the order in which cue and stimulus are presented). Other models make no presumption as to the order of processing. The neutral message served to match the sequence and timing of events across the aided and unaided blocks. The stimulus display remained onscreen until the participant's response. At the end of each trial, participants received a 1,500-ms feedback message of either "Correct!" or "Incorrect!"

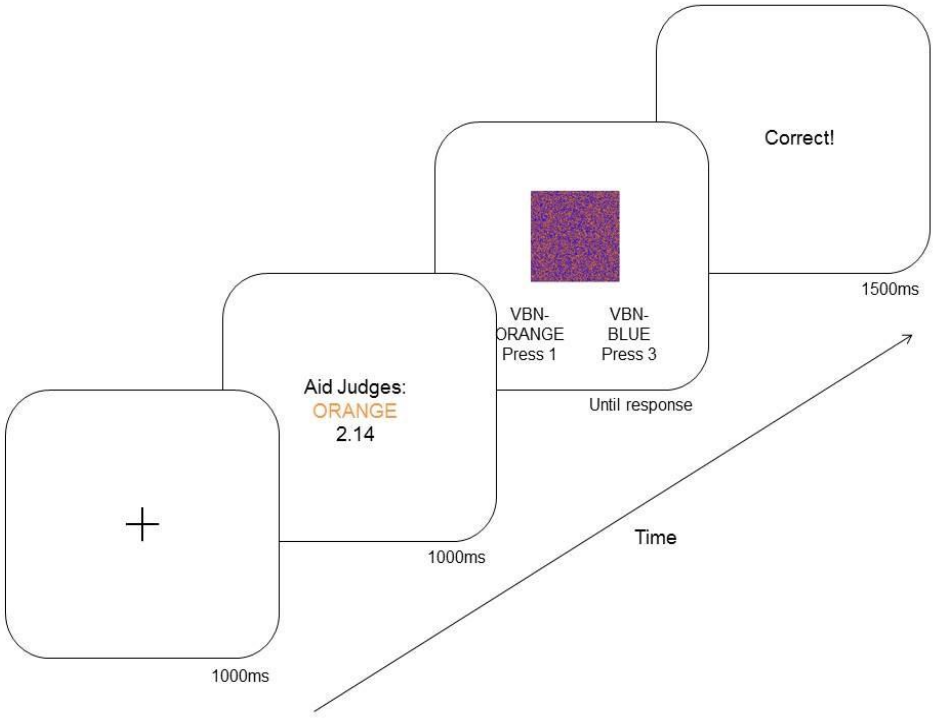

Figure 2. The sequence of events within an automation-aided trial for Experiment 1.

Each session comprised a block of 50 unaided practice trials followed by a block of 50 aided practice trials, then a block of 100 unaided experimental trials and a block of 100 aided experimental trials, with the order of the experimental blocks counterbalanced across participants. The order of stimulus images viewed within blocks was randomized across trials. Participants were allowed to rest between blocks. An experimental session lasted approximately 45 min.

Analysis

For analysis, orange-dominant stimuli were treated as signal events and blue-dominant stimuli as noise events. For clarity of exposition below, we refer to orange and blue judgments as *yes* and *no* judgments, respectively. Hit rates and false alarm rates were calculated from the participants' responses, and data were converted to signal detection measures of sensitivity and bias, *d'* and *c* (Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). A prior of 0.5 was added

to the raw response frequency value in each cell of the 2 x 2 SDT matrix for each participant to correct for perfect hit and false alarm rates (Hautus, 1995). Data from practice trials were excluded from analysis.

Data analysis employed Bayesian parameter estimation using a Markov chain Monte Carlo (MCMC) sampling procedure (Kruschke, 2013, 2015; Lee & Wagenmakers, 2014). This approach begins by assuming a prior distribution on a parameter value of interest, then updates the prior through probabilistic sampling to approximate the posterior distribution on parameter values in light of the observed data.

Analyses were conducted using sampling functions from the package JAGS (Plummer, 2015) in the R programming language (http://www.r-project.org). All parameters were assumed to follow normal distributions, with vague priors on their means and standard deviations (means ~ $N[0, 1 \times 10^6]$; standard deviations ~ $1/\Gamma[.001, .001]$). The use of vague priors ensures that the analysis does not commit *a priori* to strong conclusions, and allows the observed data to dominate the posterior distribution. Each estimate was based on four MCMC chains, run for 10,000 burn-in steps followed by 100,000 sample steps each. Chains were thinned to every fourth step in order to reduce sample autocorrelation, leaving a total of 100,000 samples for analysis. All estimated parameters showed values of the Gelman-Rubin statistic (Gelman & Rubin, 1992) of 1.01 or less, indicating satisfactory convergence of the MCMC chains (Kruschke, 2015).

Descriptive statistics reported include the mean and 95% highest density intervals (HDI) for the estimated posterior distributions (Kruschke, 2013). The 95% HDI is the region that contains 95% of the posterior distribution mass, and within which all values have higher probability than any values outside the region. If the

distribution is unimodal and symmetrical, the 95% HDI is equivalent to the central 95% region of the posterior (Gelman et al., 2013). Where it is useful to compare measures to a value of 0—for example, when examining differences between aided and unaided performance, or between observed data and model predictions—the reported statistics also include the proportion of the estimated posterior distribution that lies above or below 0 (Kruschke, 2013). Values are reported with the nomenclature $x\% < 0 < y\%$. For example, $1\% < 0 < 99\%$ indicates that 1% of the posterior distribution lies below 0, and 99% lies above. We describe an effect as credible if the 95% HDI on the difference between conditions does not overlap 0, and we describe an effect as decisive if more than 99% of the posterior distribution on difference scores falls to one side of 0 (cf. Jeffreys, 1961; Wetzels et al., 2011).

Results

Table 1 presents participants' mean hit and false alarm rates for the unaided and aided conditions of Experiments 1–3. The gray bars of Figure 3 present the corresponding mean values of *d'*. The gray bars of Figure 4 present participants' mean values of the bias measure *c* in the automation-aided conditions of Experiments 1–3, contingent on the aid's binary judgment. Dotted lines in Figures 3 and 4 present model-predicted values. Results for Experiment 1 appear in the left data column of the table and left panels of the figures.

Data were excluded from four participants in Experiment 1 who failed to achieve an unaided *d'* score of at least 0.5, suggesting a failure to understand or comply with the instructions. Including these participants' data in the analyses below did not change the pattern of results.

Sensitivity. Automation-aided *d'* decisively exceeded unaided *d'*, M_{diff} = 0.48, 95% HDI [0.20, 0.75], 0% < 0 < 100%, confirming that assistance from the aid improved participants' sensitivity.

To assess model performance, analyses compared observed *d'* scores from the automation-aided conditions to the model-predicted scores based on the participants' unaided sensitivity. Mean model error scores (predicted scores minus observed scores) are presented in the text, with 95% HDIs. The two models that took into account the aid's graded evidence judgments, the OW model, M_{err} = 1.06, 95% HDI [0.86, 1.28], 0% < 0 < 100%, and the UW model, M_{err} = 0.97, 95% HDI [0.75, 1.19], 0% < 0 < 100%, both decisively overestimated participants' automation-aided sensitivity, as did the three most efficient of the binary-cue models, M_{err} = 0.75, 95% HDI [0.54, 0.96], 0% < 0 < 100% for the optimal CC model, M_{err} = 0.52, 95% HDI [0.31, 0.72], 0% < 0 < 100% for the NN model, and M_{err} = 0.37, 95% HDI [0.18, 0.56], 0% < 0 < 100% for the BD model. In contrast, the CF model decisively underestimated participants' aided sensitivity, M_{err} = -0.23, 95% HDI [-0.43, -0.03], 99% < 0 < 1%. Observed sensitivity did not differ credibly from the predictions of the PM model, M_{err} = 0.16, 95% HDI [-0.04, 0.36], 6% < 0 < 94%.

Table 1
Mean Hit and False Alarm Rates and 95% HDIs (in brackets) for the Unaided and Aided Conditions of Experiments 1, 2, and 3.

	Experiment 1		Experiment 2		Experiment 3	
	Unaided	Aided	Unaided	Aided	Unaided	Aided
Hit rate	0.82 [0.78, 0.87]	0.90 [0.87, 0.92]	0.83 [0.80, 0.87]	0.89 [0.86, 0.91]	0.82 [0.79, 0.86]	0.90 [0.86, 0.93]
False alarm rate	0.14 [0.11, 0.17]	0.10 [0.08, 0.12]	0.13 [0.10, 0.15]	0.08 [0.06, 0.10]	0.14 [0.10, 0.18]	0.10 [0.07, 0.13]

Note. HDI = Highest-density interval.

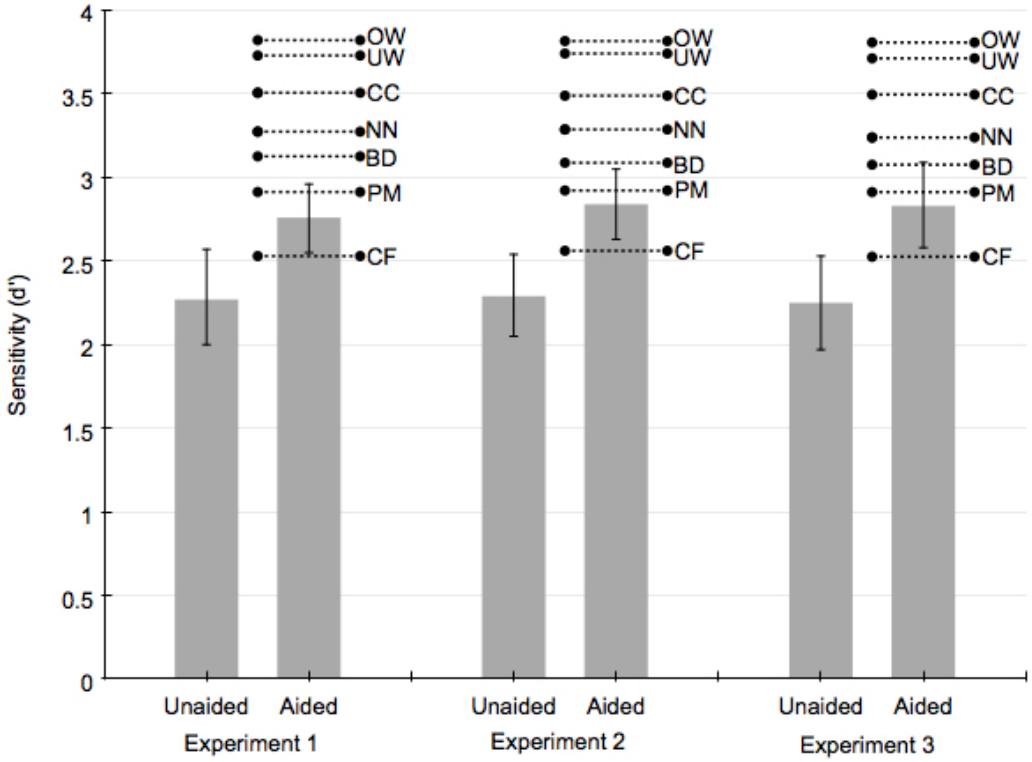

Figure 3. Mean *d'* values (gray bars) and model predictions (dotted lines) for Experiments 1, 2, and 3. Error bars indicate 95% highest-density intervals.

Bias. Observed levels of automation-aided sensitivity fell closest to the predictions of the PM model, which holds that the operator defers to the aid with a probability equal to the aid's average reliability. One interpretation of this finding is that participants were in fact using a PM strategy. An alternative possibility, however,

is that participants were using a different strategy, but one which happened to mimic the sensitivity of the PM model. As a further test of the models, analyses compared the participants' automation-aided response bias, contingent on the aid's judgment, to the predictions of the NN, CF, optimal CC, and PM models. Note that the predicted bias for trials on which the aid provided a *Yes* judgment is negative infinity under the NN model, and is therefore not shown in Figure 4.

As expected, observed bias was decisively more liberal when the aid gave a *Yes* judgment than when it gave a *No* judgment, M_{diff} = 1.26, 95% HDI [1.03, 1.49], 0% < 0 < 100%, confirming that participants' responses were biased in the direction of the aid's judgments. The magnitude of the observed shifts, however, did not closely match the predictions of any of the models under consideration. For trials on which the aid issued a *Yes* judgment, observed bias was decisively more conservative than predicted by the PM model, M_{err} = -1.38, 95% HDI [-1.54, -1.22], 100% < 0 < 0%, the optimal CC model, M_{err} = -0.73, 95% HDI [-0.95, -0.50], 100% < 0 < 0%, or the CF model, M_{err} = -0.24, 95% HDI [-0.41, -0.07], 100% < 0 < 0%. For trials on which the aid issued a *No* judgment, observed bias was decisively more liberal than predicted by the PM model, M_{err} = 1.36, 95% HDI [1.20, 1.51], 0% < 0 < 100%, the optimal CC model, M_{err} = 0.67, 95% HDI [0.47, 0.86], 0% < 0 < 100%, or the CF model, M_{err} = 0.24, 95% HDI [0.08, 0.40], 0% < 0 < 100%, and decisively more conservative than predicted by the NN model, M_{err} = -0.61, 95% HDI [-0.78, -0.44], 100% < 0 < 0%.

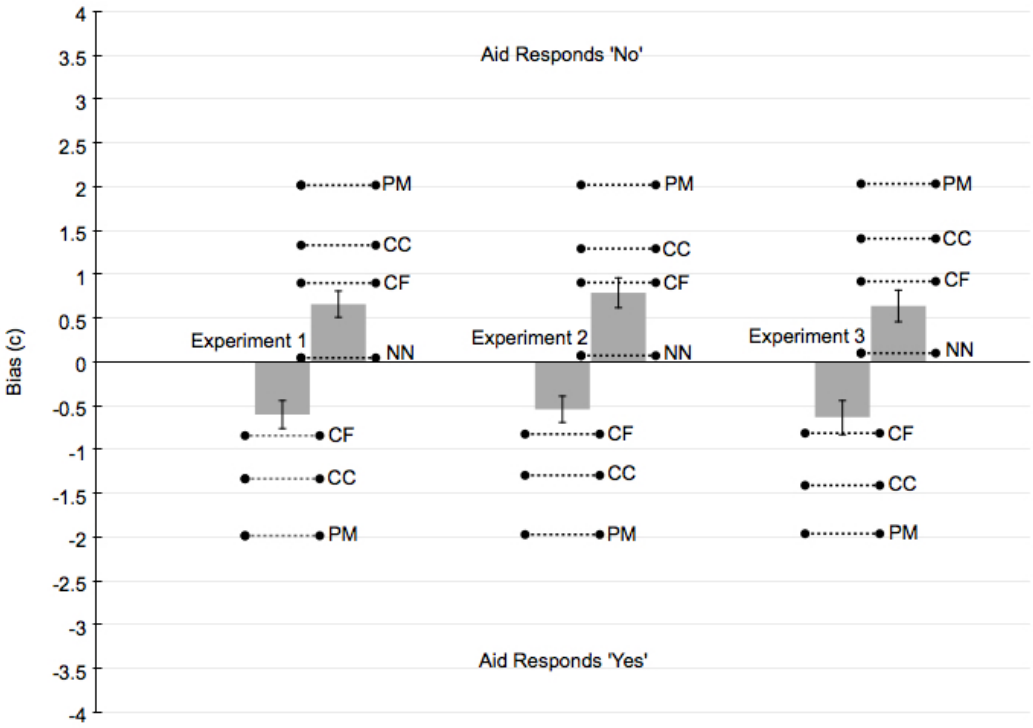

Figure 4. Mean of the observed (gray bars) and model-predicted (dotted lines) *c* values for Experiments 1, 2, and 3, contingent on the aid's judgment. The left bar within each panel corresponds to trials on which the aid provided a *Yes* judgment, and the right bar corresponds to trials on which the aid provided a *No* judgment. Error bars indicate 95% highest-density intervals.

Discussion

Automation-aided sensitivity fell closest to the predictions of the PM model. Aided values of *c*, however, were far less extreme than the PM model predicted, and did not match any of the models' predictions closely. But what's perhaps most surprising is that participants appear to have made little or no use of the aid's graded evidence values, substantially underperforming both the OW and UW models. In other words, aided performance was no better than could have been obtained even if the aid had provided only binary judgments. Experiment 2 pursued this result.

Experiment 2

Automation-aided participants in Experiment 1 far underperformed the OW and UW models, suggesting they made little use of the automated aid's graded evidence outputs. Experiment 2 tested this possibility by replicating the procedure of Experiment 1, but only providing participants with a binary judgment from the aid each trial. If participants derived no benefit from the aid's signal strength ratings in the first experiment, performance in Experiment 2 should match that of Experiment 1.

Method

Participants. Participants were 37 adults (mean age = 22.45 years, *SD* = 5.41, range = 17-39; 26 females, 11 males) recruited from the Flinders University of South Australia, none of whom had taken part in Experiment 1. All participants were compensated with $10.00 AUD for an experimental session that lasted approximately 45 minutes. Participants were fluent in English, had normal color vision, and normal or corrected-to-normal visual acuity.

Apparatus and Stimuli. Apparatus and stimuli were identical to those of Experiment 1, except that participants received only a binary, orange-or-blue judgment from the aid each trial.

Procedure. Experimental procedure and data treatment were similar to those of Experiment 1. Instructions were identical to those of Experiment 1, but modified to omit any mention of continuous values from the aid. Participants were advised simply, "The aid will provide its assessment at the start of each trial."

Results

Results for Experiment 2 appear in the middle data column of Table 1 and the middle panels of Figures 3 and 4. Data were excluded from one participant who failed

to achieve an unaided *d'* score of at least 0.5. Including that participants' data in the analyses below did not change the pattern of results.

Sensitivity. Automation-aided sensitivity exceeded unaided sensitivity by a mean of M_{diff} = 0.55, 95% HDI [0.35, 0.76], 0% < 0 < 100%, with a credible interval that clearly excluded 0, indicating that assistance from the aid again improved participants' sensitivity.

As in the first experiment, however, the participants' aided performance was poor relative to the predictions of the most efficient decision models under consideration. Both the OW model, M_{err} = 0.98, 95% HDI [0.80, 1.15], 0% < 0 < 100%, and the UW model, M_{err} = 0.90, 95% HDI [0.73, 1.08], 0% < 0 < 100%, decisively overestimated aided sensitivity. This is unsurprising, given that OW and UW performance is unattainable based only on binary judgments from the aid. However, aided performance also fell decisively below the levels predicted by the optimal CC model, M_{err} = 0.65, 95% HDI [0.47, 0.82], 0% < 0 < 100%, and the NN model, M_{err} = 0.45, 95% HDI [0.27, 0.62], 0% < 0 < 100%, and credibly below the levels predicted by the BD model, M_{err} = 0.25, 95% HDI [0.05, 0.44], 1% < 0 < 99%. In contrast, aided sensitivity was decisively better than predicted by the CF model, M_{err} = -0.28, 95% HDI [-0.45, -0.11], 100% < 0 < 0%, and again did not differ credibly from the predictions of the PM model, M_{err} = 0.08, 95% HDI [-0.11, 0.28], 20% < 0 < 80%.

Bias. Observed bias was again decisively more liberal when the aid responded *Yes* than when it responded *No*, M_{diff} = 1.33, 95% HDI [1.09, 1.58], 0% < 0 < 100%. For trials on which the aid issued a *Yes* judgment, observed bias was more conservative than predicted by the PM model, M_{err} = -1.43, 95% HDI [-1.58, -1.28], 100% < 0 < 0%, the optimal CC model, M_{err} = -0.75, 95% HDI [-0.99, -0.50], 100% <

0 < 0%, or the CF model, M_{err} = -0.28, 95% HDI [-0.44, -0.13], 100% < 0 < 0%. For trials on which the aid issued a *No* judgment, observed bias was decisively more liberal than predicted by either the PM model, M_{err} = 1.23, 95% HDI [1.05, 1.41], 0% < 0 < 100%, or the optimal CC model, M_{err} = 0.50, 95% HDI [0.28, 0.72], 0% < 0 < 100%, and decisively more conservative than predicted by the NN model, M_{err} = -0.72, 95% HDI [-0.90, -0.53], 100% < 0 < 0%. Observed bias after a *No* from the aid did not differ credibly from that predicted by the CF model, M_{err} = 0.11, 95% HDI [-0.07, 0.30], 11% < 0 < 89%.

Cross-experiment comparison. Assistance from the automated aid increased participants' *d'* by 0.48 in Experiment 1 and 0.55 in Experiment 2, M_{diff} = 0.07, 95% HDI [-0.27, 0.41], 35% < 0 < 65%, giving no credible evidence that graded evidence values offered by the aid in Experiment 1 helped participants achieve higher sensitivity. In fact, though the difference was statistically negligible, automation-aided sensitivity trended higher in the second experiment than in the first.

Discussion

Experiment 2 produced a pattern of effects highly similar to that of Experiment 1, suggesting that participants made little use of the aid's graded evidence values in the first experiment. Results affirm more generally that automation-aided performance was highly inefficient, roughly matching the predictions of the PM model, but that participants' cue-contingent response bias did not closely match the predictions of any of the models tested.

Experiment 3

Experiment 1found highly inefficient automation use, even with graded estimates of signal strength from the aid. Experiment 3 sought to confirm this result with a close replication of the first experiment. As a modest extension, a scoring

system was incorporated to provide an overt performance incentive and help participants better track their performance over trials.

Method

Participants. Participants were 36 adults (mean age = 22.16 years, *SD* = 4.71, range = 17-35; 30 females, 6 males) recruited from the Flinders University of South Australia, none of whom had taken part in Experiment 1 or 2. All participants were compensated with $10.00 AUD for an experimental session that lasted approximately 45 minutes. Participants were fluent in English, had normal color vision, and normal or corrected-to-normal visual acuity.

Apparatus and stimuli. Apparatus and stimuli were identical to those of Experiment 1, except that a point score and running total score was provided with the feedback screen each trial.

Procedure. Experimental procedure and data treatment were similar to those of Experiment 1, except as follows. Instructions were identical to those of Experiment 1, but modified to account for the point system. Participants were advised, "You will be scored on your performance, as Marvel Mining has declared that incorrect sorting of the strains has been detrimental. You will receive 5 POINTS for every correct judgment, and you will be deducted 5 POINTS for every incorrect judgment." The total score that could be obtained in the experimental trials was 1,000 points.

At the conclusion of each trial, participants received a 1500ms feedback message of "Correct! +5, Total score = score" for all correct responses, and "Incorrect! -5, Total score = score" for all errors.

Results

Results for Experiment 3 appear in the right data column of Table 1 and the right panels of Figures 3 and 4.

Sensitivity. As in the first two experiments, assistance from the automated aid decisively improved participants' *d'*, M_{diff} = 0.58, 95% HDI [0.33, 0.83], 0% < 0 < 100%. Again, though, aided performance was highly inefficient. The OW, M_{err} = 0.97, 95% HDI [0.76, 1.19], 0% < 0 < 100%, UW, M_{err} = 0.88, 95% HDI [0.66, 1.10], 0% < 0 < 100%, optimal CC, M_{err} = 0.66, 95% HDI [0.44, 0.87], 0% < 0 < 100%, and NN models, M_{err} = 0.40, 95% HDI [0.19, 0.62], 0% < 0 < 100%, all decisively overestimated aided sensitivity, and the CF model was once more the only model to decisively underestimate it, M_{err} = -0.31, 95% HDI [-0.52, -0.10], 100% < 0 < 0%. Although the BD model again tended to overestimate aided sensitivity, M_{err} = 0.24, 95% HDI [0.00, 0.49], 3% < 0 < 97%, the difference between its predictions and observed performance in this case just failed to reach 95% credibility. As in the earlier experiments, however, observed performance fell closest to the predictions of the PM model, M_{err} = 0.08, 95% HDI [-0.17, 0.31], 26% < 0 < 74%.

Bias. As expected, observed bias was decisively more liberal when the aid responded *Yes*, than when it responded *No*, M_{diff} = 1.27, 95% HDI [0.98, 1.57], 0% < 0 < 100%. For trials on which the aid issued a *Yes* judgment, observed bias was decisively more conservative than predicted by either the PM, M_{err} = -1.33, 95% HDI [-1.52, -1.13], 100% < 0 < 0%, or optimal CC model, M_{err} = -0.78, 95% HDI [-1.13, -0.42], 100% < 0 < 0%. Observed bias trended more liberal than predicted by the CF model, M_{err} = -0.18, 95% HDI [-0.38, 0.01], 96% < 0 < 4%, though the difference was just short of credible. For trials on which the aid issued a *No* judgment, observed bias was decisively more liberal than predicted by either the PM, M_{err} = 1.39, 95% HDI [1.21, 1.57], 0% < 0 < 100%, the optimal CC, M_{err} = 0.77, 95% HDI [0.43, 1.10], 0% < 0 < 100%, or CF model, M_{err} = 0.28, 95% HDI [0.09, 0.47], 0% < 0 < 100%, and

decisively more conservative than predicted by the NN model, M_{diff} = -0.54, 95% HDI [-0.75, -0.34], 100% < 0 < 0%.

Cross-experiment comparison. Assistance from the automated-aid increased participants' *d'* by 0.48 in Experiment 1 and 0.58 in Experiment 3, M_{diff} = 0.10, 95% HDI [-0.27, 0.47], 29% < 0 < 71%, giving little evidence that the point system of Experiment 3 improved participants' automation use. This does not imply that with more data the modest performance difference between experiments might not become credible, or that stronger incentives or different feedback might not induce more efficient automation use, but it does lend confidence that the effects seen in Experiment 1 are generally robust.

Meta-analysis

To estimate the discrepancies between observed data and model predictions more precisely, we combined the data of all three experiments and repeated the analyses reported above on the aggregated data.

Consistent with the conclusions above, aggregated sensitivity was decisively higher in the aided condition than in the unaided condition, M_{diff} = 0.54, 95% HDI [0.40, 0.67], 0% < 0 < 100%, but was nonetheless highly inefficient. The five most efficient models under consideration all decisively overestimated automation-aided sensitivity, M_{err} = 1.00, 95% HDI [0.89, 1.12], 0% < 0 < 100% for the OW model, M_{err} = 0.92, 95% HDI [0.80, 1.03], 0% < 0 < 100% for the UW model, M_{err} = 0.69, 95% HDI [0.57, 0.80], 0% < 0 < 100% for the optimal CC model, M_{err} = 0.46, 95% HDI [0.34, 0.57], 0% < 0 < 100% for the NN model, and M_{err} = 0.29, 95% HDI [0.17, 0.41], 0% < 0 < 100% for the BD model, and only the CF model decisively underestimated it, M_{err} = -0.27, 95% HDI [-0.38, -0.16], 100% < 0 < 0%. As above, the PM model came closest to matching observed performance levels. With the

additional statistical resolution allowed by the aggregated data set, however, the discrepancy between the model's predictions and observed performance approached 95% credibility, M_{err} = 0.11, 95% HDI [-0.01, 0.22], 4% < 0 < 96%.

For trials on which the aid issued a *Yes* judgment, aggregated bias data were decisively more conservative than the predictions of either the PM, M_{err} = -1.38, 95% HDI [-1.47, -1.28], 100% < 0 < 0%, optimal CC, M_{err} = -0.75, 95% HDI [-0.91, -0.59], 100% < 0 < 0%, or CF model, M_{err} = -0.23, 95% HDI [-0.33, -0.14], 100% < 0 < 0%. For trials on which the aid issued a *No* judgment, aggregated bias data were decisively more liberal than the predictions of either the PM, M_{err} = 1.33, 95% HDI [1.23, 1.43], 0% < 0 < 100%, optimal CC, M_{err} = 0.64, 95% HDI [0.50, 0.79], 0% < 0 < 100%, or CF model, M_{err} = 0.21, 95% HDI [0.11, 0.31], 0% < 0 < 100%, and decisively more conservative than the predictions of the NN model, M_{err} = -0.62, 95% HDI [-0.73, -0.52], 100% < 0 < 0%.

In summary, when data were aggregated across experiments, aided sensitivity fell closest to the predictions of the PM model, but differed from them with borderline credibility. Conditionalized bias data remained inconsistent with any of the models under consideration.

Model comparisons

The results above imply that the PM model may be useful as a heuristic for roughly predicting automation-aided sensitivity, but that participants likely did not employ the PM strategy, or any of the other parameter-free or fixed-parameter strategies under consideration, to make aided decisions. This allows that the data may instead be most compatible with a suboptimal CC model (Robinson & Sorkin, 1985), under which participants make automation-assisted judgments by shifting their response criterion in the direction stipulated by the aid's decision, but to an inadequate degree. However, the analyses above did not test

the performance of the suboptimal CC model, and thus provide no direct evidence in support of the model. They also considered sensitivity and bias data separately, rather than jointly. We therefore conducted a model-fitting analysis to compare the performance of the CC model to that of the other models discussed above.

Method

Models were fit using an MCMC Bayesian estimation. Because the empirical data indicated that participants made little use of the aid's graded judgments, only models that relied exclusively on binary cues from the aid were considered. Four models were compared: a CC model, a variant of the CF/PM models that we will call the *discrete-state deferment* model, the BD model, and the NN model. Unaided sensitivity in all four cases was estimated using the hierarchical signal detection model described by Lee and Wagenmakers (2014). At the top level, the model assumes population distributions of sensitivity (*d'*) and criterion (*c*) values. At the level below, it assumes that individual participants render judgments using an equal-variance Gaussian signal detection model with *d'* and *c* values sampled from the population distributions. Finally, individual participants' *d'* and *c* values are reparameterized as hit and false alarm rates, and used to predict raw hit and false alarm counts from a binomial distribution. Population distributions of *d'* and *c* are assumed to be described by normal distributions, with vague priors on their means and standard deviations (means ~ N[0, .00001]; standard deviations ~ 1/Γ[.001, .001]).

Aided sensitivity was estimated differently across the four models. All four models assumed that participants made their own judgments in the aided condition with the same sensitivity as in the unaided condition, and that participants received correct judgments from the aid on 93% of all trials. The models differed in the manner by which they combined the participants' and aid's judgments. The CC model (Robinson & Sorkin, 1985) treated participants' response criteria as free parameters, estimating separate values for trials on

which the aid responded *No* and trials on which the aid responded *Yes*. It therefore subsumed the optimal and suboptimal CC models: cue-contingent criterion values that matched the normative values would signal optimal performance, and values that deviated from normative would signal suboptimal performance. The model assumed that criteria for *Yes* and *No* trials were normally distributed with the same prior distributions as the criteria for unaided trials.

In the discrete-state deferment model, participants resolved disagreements with the aid by deferring to the aid's judgment with a fixed probability. The probability of deferring to the aid was treated as a free parameter described by a beta distribution at the population level. The beta distribution is defined on the interval [0, 1], and is characterized by two parameters (Kruschke, 2015). In the parameterization used here, these parameters were the mode, ω, and concentration, κ, of the distribution. A value of $\kappa = 2$ produces a uniform distribution on the interval [0, 1]. Higher values produce more peaked distributions. The model therefore subsumed the CF and PM models discussed above: a distribution of deferment probabilities peaked tightly around a mode of 0.50 would indicate behavior consistent with the CF model, and a distribution peaked tightly around 0.93 would indicate behavior consistent with the PM model. The parameters ω and κ were assigned vague priors (both $\sim \Gamma[.001, .001]$).

Finally, the NN model assumed that an aided participant issued a *No* response only in the event that both the aid and the participant reached independent judgments of *No*, and the BD model assumed that decisions in aided blocks were made by whichever agent, human or aid, had higher sensitivity.

Each simulation employed four MCMC chains, run for 10,000 burn-in steps followed by 100,000 sample steps each. Chains were thinned to every fourth step, leaving a total of 50,000 samples for analysis. All estimated parameters showed values of the Gelman-Rubin statistic (Gelman & Rubin, 1992) of 1.01 or less. Model performance was compared using the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, & van Der Linde, 2002), a

measure that rewards a model for goodness-of-fit but penalizes it for complexity (Myung & Pitt, 1997). Smaller values denote a better-fitting model. As a rule of thumb, a difference of DIC in the range of 3 to 7 is regarded as considerable evidence in favor of the better-fitting model (Spiegelhalter et al., 2002).

Results

Of the four models under comparison, the NN produced clearly the worst performance, *DIC* = 4120, followed by the BD, *DIC* = 2652. The discrete-state deferment model performed better, producing a DIC of 2377. The mean posterior value of ω, the modal estimated rate with which participants deferred to the aid, was 0.50, 95% HDI [0.01, 0.96], a value superficially consistent with a CF strategy. However, the mean estimated posterior value of κ, the concentration of deferment rates around the modal value, was 2.00, 95% HDI [2.00, 2.00], indicating that deferment rates across participants were very close to uniformly distributed between 0 and 1. Accordingly, the 95% HDI on ω spanned almost the full range of values between 0 and 1. Model fits thus revealed no tendency for participants to cluster around any particular deferment rate, that is, no consistency in automation use across participants. This implies that fixing the value of κ at 2 should improve the model's DIC, reducing model complexity without sacrificing goodness-of-fit. Consistent with this, running the model with κ as a fixed parameter of value 2 produced a DIC of 2374, better than was achieved by treating κ as a free parameter.

The CC model produced a DIC of 2375, nearly equivalent to that for the discrete-state deferment model with fixed κ. The estimated means of the participants' cued criterion values were less extreme than optimal, both for trials on which the aid issued a *Yes* judgment, M = -0.54, 95% HDI [-0.70, -0.39] for observed c vs. M = -1.19, 95% HDI [-1.28, -1.11], for optimal c, M_{diff} = 0.65, 95% HDI [0.48, 0.81], 0% < 0 < 100%, and for trials on which the aid issued a *No* judgment, M = 0.60, 95% HDI [0.45, 0.76] for observed c vs. M = 1.19, 95%

HDI [1.10, 1.28] for optimal *c*, M_{diff} = -0.59, 95% HDI [-0.76, -0.42], 100% < 0 < 0%. HDIs around the differences between observed and optimal criterion values clearly excluded 0, indicating that a tendency toward overly conservative criterion shifts was highly consistent across participants.

In total, results suggest that data were roughly indifferent between the discrete-state deferment model with κ fixed at 2, and a suboptimal, overly-conservative CC model. As discussed below, other considerations tilt in favor of the CC model over the discrete state model.

General discussion

Of the seven fixed-parameter or parameter-free models considered above, the PM model most closely predicted participants' automation-aided sensitivity. Conditionalized on the aid's judgments, however, automation-aided response bias was inconsistent with any of the seven models. Thus, despite the rough match between the observed sensitivity data and the predictions of the PM model, participants do not seem to have used a PM strategy, or in fact to have used any of the fixed-parameter or parameter-free strategies tested.

How, then, did participants reach their automation-aided decisions? Model comparisons were effectively indifferent between a suboptimal CC model and a discrete-state deferment model that subsumes the CF and PM models as special cases. The suboptimal CC model, as explained above, assumes that participants made automation-assisted decisions by shifting their response criterion in the direction stipulated by the aid, but to an inadequate degree. The discrete state model assumes that participants resolved disagreements with the aid by deferring to the automation's judgments with some fixed probability. The models differ functionally in that the CC model implies that a decision maker is more likely to override the aid's recommendation when she is highly confident in her own judgment, for example, on trials when a signal is especially strong. In contrast, the discrete-state model

holds that the decision maker is equally likely to override the aid whether or not she is confident in her own judgment. This suggests that future work may be better able to distinguish the models empirically by examining participants' automation usage across different levels of signal strength.

Until more decisive empirical tests can be conducted, considerations of plausibility (Myung & Pitt, 1997; Spiegelhalter et al., 2002) may best adjudicate between the suboptimal CC and discrete-state deferment models, and seem to favor the suboptimal CC account. The discrete-state model achieved its best fit by assuming that deferment rates across participants were uniformly distributed between 0 and 1. In other words, it posited no consistency across individuals in the tendency to depend on the automation. The suboptimal CC model, in contrast, implied a consistent pattern of behavior across individuals, with HDIs on cue-contingent criteria indicating that participants were unanimously too conservative in their automation dependence. Although decision makers can most certainly differ in their willingness to depend on an automated aid (e.g., Szalma & Taylor, 2011), the possibility that they show no consistent tendencies at all seems unlikely, lending credence to the suboptimal CC model here. The tendency toward inadequate criterion shifts following a cue from the automated decision aid is also consistent with the more general 'sluggish beta' phenomenon (Chi & Drury, 1998; Neyedli, Hollands, & Jamieson, 2011; Wang et al., 2009), a tendency for decision makers in signal detection tasks to adjust their criterion less than they should in response to manipulations of signal rates and event payoffs. These various considerations tentatively suggest that the optimal CC model offers a more plausible account of automation usage than the discrete-state deferment model, even if both produced similar DICs.

As discussed above, other research has also inferred a suboptimal CC strategy from participants' automation-aided sensitivity and criteria (Elvers & Elrif, 1997; Meyer, 2001; Robinson & Sorkin, 1985; Wang et al., 2009). However, the present results go beyond earlier

findings by demonstrating that the participants' suboptimal criterion choice produced sensitivity that approached the predictions the PM model. Further research will be necessary to generalize this pattern across different forms of signal detection task and varying levels of aid reliability, and to identify markers of individual differences (e.g., Merritt & Ilgen, 2008) that allow some users to consistently attain higher benchmarks of automation-aided efficiency than others. But preliminarily, the data imply that, knowing the *d'* of an unaided operator and the *d'* of an automated aid, system designers can use the PM model to roughly predict the operator's aided sensitivity. These predictions can in turn inform analyses of the costs and benefits of building and deploying automated aids.

The tendency for decision makers to disuse decision aids that are not perfectly reliable is of course well-established (Parasuraman & Riley, 1997; Wickens & Dixon, 2007). The finding that participants used automated aids so inefficiently is especially notable here, though, because if used well, the aid's graded strength judgments in Experiments 1 and 3 could have enabled performance well above even the optimal CC level. In fact, aided performance was no better in the first and third experiments than in the second, which offered only binary judgments from the aid.

Data do not make clear why decision makers used the aid's graded cues so inefficiently. Achieving optimal performance would have been challenging in multiple ways. First, participants would have had to know, implicitly or explicitly, the statistical properties of their own sensory representations corresponding to blue-dominant and orange-dominant stimuli. Second, they would have had to know the analogous statistical properties of the aid's evidence distributions. Third, they would have had to know how much better or worse their sensitivity was than the aid's. Armed with all of this knowledge, finally, the participants would have had to calculate an appropriately weighted average of their own judgment and the aid's each trial.

Given these heavy demands, the failure to match the performance of the OW model is unsurprising; researchers have long recognized that limits on information and information-processing abilities place bounds on human cognition that can prevent human decision makers from reaching putatively normative performance (Simon, 1955; Tversky & Kahneman, 1974). Nonetheless, human decision makers can at least approximate the performance of a linear cue combination rule (Einhorn, Kleinmuntz, & Kleinmuntz, 1979), and though they tend not to weight cues optimally (e.g., Johnson, Cavanagh, Spooner, & Samet, 1973; Montgomery, 1999, 2001; Montgomery & Sorkin, 1996), their deviations from normative weighting are likely to have modest effects on performance (Dawes & Corrigan, 1974; Wainer, 1976). Comparing the predictions of the OW and UW models above, for instance, shows that an equal-weighting rule for combining human and automation judgments would have approached the performance of the optimal-weighting rule. It therefore seems unlikely that participants' inefficient use of graded evidence values was caused by an inability to estimate proper weights for combining judgments. Moreover, even when high cognitive load or imperfect information make a linear decision rule difficult or impracticable, decision makers can often find nonlinear heuristic strategies that allow near-normative performance (Gigerenzer & Gaissmaier, 2011; Hogarth & Karelaia, 2007). In the current tasks, for example, a simple heuristic rule for using the aid's graded judgments to resolve disagreements between the human and aid might have been to defer to the aid when it produced a relatively high evidence value but to override it otherwise.

Despite these possibilities, the data gave little indication that participants made use of the aid's graded evidence judgments. Rather, the null differences between Experiment 2 and Experiments 1 and 3, and the highly inefficient levels of automation-aided performance seen in all three experiments, suggest that participants disregarded the aid's graded outputs entirely. This may indicate a tendency for participants to minimize effort expenditure

(Bettman, Johnson, & Payne, 1990), sacrificing decision accuracy in order to forego the short-term cognitive costs of encoding and remembering the aid's graded assessment each trial. Further research will be necessary to determine whether instruction (Sedlmeier & Gigerenzer, 2001), changes to the format in which information from the aid is presented (Bisantz, 2013; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; Todd & Benbasat, 1994), or other task and display manipulations might reduce the effort needed to use the aid's graded assessments and induce more efficient human-automation performance.

References

Alberdi, E., Povyakalo, A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology, 11*, 909-918.

Andre, A. D., & Cutler, H. A. (1998). Displaying uncertainty in advanced navigation systems. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 31-35). Santa Monica, CA: Human Factors and Ergonomics Society.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329*, 1081- 1085.

Bettman, J. R., Johnson, E. J., & Payne, J. W. (1990). A componential analysis of cognitive effort in choice. *Organizational Behavior and Human Decision Processes, 45*, 111-139.

Bisantz, A. M. (2013). Uncertainty visualization and related techniques. In J. D. Lee & A. Kirlik (Eds.), *The oxford handbook of cognitive engineering* (pp. 579-594). New York, NY: Oxford University Press.

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics, 38*, 2300-2312.

Botzer, A., Meyer, J., Bak, P., & Parmet, Y. (2010). User settings of cue thresholds for binary categorization decisions. *Journal of Experimental Psychology: Applied, 16*, 1-15.

Chi, C., & Drury, C. G. (1998). Do people choose an optimal response criterion in an inspection task? *IIE Transactions, 30*, 257-266.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Denkiewicz, M., Rączaszek-Leonardi, J., Migdal, P., & Plewczynski, D. (2013). Information-sharing in three interacting minds solving a simple perceptual task. In *Proceedings of*

the Annual Meeting of the Cognitive Science Society (pp. 2172-2176). Austin, TX: Cognitive Science Society.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology, 13*, 147-164.

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review, 86*, 465-485.

Elvers, G. C., & Elrif, P. (1997). The effects of correlation and response bias in alerted monitor displays. *Human Factors, 39*, 570-580.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-511.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451-482.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Gupta, N., Bisantz, A. M., & Singh, T. (2002). The effects of adverse condition warning system characteristics on driver performance: An investigation of alarm signal type and threshold level. *Behaviour & Information Technology, 21*, 235-248.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers, 27*, 46-51.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290*, 2261-2262.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review, 114*, 733-758.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.

Jensen, M., Lowry, P. B., & Jenkins, J. L. (2011). Effects of automated participative decision support in computer-aided credibility assessment. *Journal of Management Information Systems, 28*, 201-234.

Johnson, E. M., Cavanagh, R. C., Spooner, R. L., & Samet, M. G. (1973). Utilization of reliability measurements in Bayesian inference: Models and human performance. *IEEE Transactions on Reliability, 22*, 176-183.

Koehler, D. J., & James, G. (2014). Probability matching, fast and slow. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 103-131). San Diego, CA: Elsevier.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*, 573-603.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Waltham, MA: Academic Press.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Maltz, M., & Meyer, J. (2001). Use of warnings in an attentionally demanding detection task. *Human Factors, 43*, 217-226.

McCarley, J. S. (2009). Response criterion placement modulates the benefits of graded alerting systems in a simulated baggage screening task. In *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting* (pp. 1106-1110). Santa Monica, CA: Human Factors and Ergonomics Society.

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*, 194-210.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors, 43*, 563-572.

Montgomery, D. A. (1999). Human sensitivity to variability information in detection decisions. *Human Factors, 41*, 90-105.

Montgomery, D. A. (2001). Sampling methods for identifying differences in source reliability. *Journal of General Psychology, 128*, 5-20.

Montgomery, D. A., & Sorkin, R. D. (1996). Observer sensitivity to element reliability in a mutlielement visual display. *Human Factors, 38*, 484-494.

Myung, I.J., & Pitt, M. A. (1997). Applying Occam's razor in modelling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*, 79-95.

Neyedli, H. F., Hollands, J. G., & Jamieson, G. A. (2011). Beyond identity: Incorporating system reliability information into an automated combat identification system. *Human Factors, 53*, 338-355.

Parasuraman, R. (2000). Designing automation for human use: Empirical studies and quantitative models. *Ergonomics*, *43*, 931-951.

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

Plummer, M. (2015). JAGS Version 4.0.0 user manual. Retrieved from https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/.

Pollack, I., & Madans, A. B. (1964). On the performance of a combination of detectors. *Human Factors, 6*, 523-531.

Rice, S., & McCarley, J. S. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied, 17*, 320-331.

Robinson, D. E., & Sorkin, R. D. (1985). A contingent criterion model of computer assisted detection. In R. E. Eberts, & C. G. Eberts (Eds.), *Trends in ergonomics/human factors II* (pp. 75-82). Amsterdam: North-Holland.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*, 380-400.

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

Sorkin, R. D., & Dai, H. (1994). Signal detection analysis of the ideal group. *Organizational Behavior and Human Decision Processes, 60*, 1-13.

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review, 108*, 183-203.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors*, *30*, 445-459.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B, 64*, 583-639.

St. John, M., & Manes, D. I. (2002). Making unreliable automation useful. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 332-336). Santa Monica, CA: Human Factors and Ergonomics Society.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*, 137-149.

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The

five factor model of personality. *Journal of Experimental Psychology: Applied, 17*, 71-96.

Todd, P., & Benbasat, I. (1994). The influence of decision aids on choice strategies: An experimental analysis of the role of cognitive effort. *Organizational Behavior and Human Decision Processes, 60,* 36-74.

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

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*, 1206-1220.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*, 101-118.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*, 213-217.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors, 51*, 281-291.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6,* 291-298.

Wickens, C., & Colcombe, A. (2007). Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Human Factors, 49,* 839-850.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8,* 201-212.

Wiczorek, R., & Manzey, D. (2014). Supporting attention allocation in multitask environments: Effects of likelihood alarm systems on trust, behavior, and performance. *Human Factors, 56,* 1209-1221.

Wiczorek, R., Manzey, D., & Zirk, A. (2014). Benefits of decision-support by likelihood versus binary alarm systems: Does the number of stages make a difference? In *Proceedings of the Human Factors and Ergonomics Society 58th Annual Meeting* (pp. 380-384). Santa Monica, CA: Human Factors and Ergonomics Society.

Wiegmann. D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Human Factors, 44*, 44-50.

Wiegmann, D., McCarley, J. S., Kramer, A. F., & Wickens, C. D. (2006). Age and automation interact to influence performance of a simulated luggage screening task. *Aviation, Space, and Environmental Medicine, 77,* 825-831.