AN ABSTRACT OF THE THESIS OF

Shannon P. Gyles for the degree of Master of Science in Psychology, presented on June 13, 2019.

Title: Metacognition, Numeracy, and Automation-aided Decision-making.

Abstract approved: _____

Jason S. McCarley

Automated decision aids can improve human decision-making but the benefits are often compromised by inefficient use. The current experiment examined whether *metacognition*—the ability to assess self-performance—and *numeracy*—the ability to understand and work with numbers—predict the efficiency of automation use in a signal detection task. Two-hundred twenty-one participants classified random dot images as blue or orange dominant, receiving assistance from an 84% reliable decision aid on some trials. Type 1 and metacognitive signal detection measures were estimated from participants' confidence ratings, and numeracy was measured using a subjective scale. The inefficiency of automation use was assessed by measuring the deviation from optimal bias following cues from the aid (*bias error*). Data gave strong evidence that metacognition was not associated with bias error, and anecdotal evidence that numeracy and suboptimality were weakly negatively correlated. These results suggest that operators used a strategy of combining the aid's judgments with their own that is not metacognitively driven, but may depend on numeracy.

Metacognition, Numeracy, and Automation-aided Decision-making

by
Shannon P. Gyles

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented June 13, 2019
Commencement June 2019

Master of Science thesis of Shannon P. Gyles presented on June 13, 2019.

APPROVED:

Major Professor, representing Psychology

Director of the School of Psychological Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Shannon P. Gyles, Author

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

1 INTRODUCTION

In many decision-making contexts, automated decision aids assist human operators making signal detection judgments. These judgments require operators to assess whether a specified pattern of information (i.e., a signal) is present among background noise. To this end, automated aids provide operators with additional assessments of the environment on which they may base decisions. For example, an air conflict detector might help an air traffic controller identify potential collisions, a decision support system might help a nuclear power plant operator monitor whether a situation is safe or dangerous (Lee & Seong, 2007), and a combat identification system might help soldiers distinguish friends from foes on the battlefield (Wang, Jamieson & Hollands, 2009). Automated aids are not exclusive to safety-critical systems; they also support routine decision-making in many aspects of everyday life. Google maps might help commuters to identify the fastest route to work, park assist might help drivers find parking spaces that will fit their vehicle, and Netflix recommendations might help viewers decide what TV show to watch next.

1.1 Automated Decision Aids

Automated systems and devices carry out tasks, either electronically or mechanically, that would normally be performed by humans (Parasuraman & Riley, 1997; Parasuraman, Sheridan & Wickens, 2000). Tasks fall into four classes corresponding to stages of information processing: 1) information acquisition, 2) information analysis, 3) decision selection, and 4) action

implementation (Parasuraman et al., 2000). Tasks across these stages can be automated to varying degrees, ranging from fully manual to fully autonomous control (see Endsley & Kaber, 1999; Endsley & Kiris, 1995; Parasuraman et al., 2000; Riley, 1989; Sheridan & Parasuraman, 2005; and Sheridan & Verplank, 1989 for various automation level taxonomies).

Developments in sensory technology and data-processing algorithms have facilitated a shift away from automation as a means of replacing physical labor and simple open-loop control systems (Parasuraman, 1997), toward its involvement in higher-level cognitive processes (Bahner, Hüper & Manzey, 2008; Dietrich, Fodor, Zucker & Bruckner, 2010). Automated decision aids, for example, are intended to support human decision-making in domains such as air traffic control (e.g., Metzger & Parasuraman, 2005), medicine (e.g., Anand, Biondich, Liu, Rosenman & Downs, 2004), military command and control (e.g., Cesar, 1995), and driving (e.g., Maltz, & Shinar, 2004).

Automated aids support decision-making at the information acquisition, information analysis, and decision selection stages of processing (Clamann & Kaber, 2003; C.D. Wickens, 2000). At low levels of automation, decision aids may simply monitor the environment and highlight potentially important information. Automated weather forecasting systems, for example, track changes in atmospheric conditions and provide warnings when severe weather is detected (Bally, 2002; Joe et al., 2012). At higher levels of automation, decision aids not only alert operators to information but provide judgments or recommended courses of action (Parasuraman et al., 2000). For

example, medical algorithms evaluate patient history and risk factors to suggest diagnoses (Anand et al., 2004), and combat identification systems interrogate transponder signals to identify soldiers as ‘friendly’ or ‘unknown’ (Dzindolet, Pierce, Pomranky, Peterson & Beck, 2001).

1.1.1 Automation-aided decision-making

Automation has the potential to improve signal detection sensitivity, reduce mental workload, and improve system efficiency (e.g., Dixon & Wickens, 2006; Maltz & Shinar, 2003; Metzger & Parasuraman, 2005). Unlike fully automated systems, in which humans assume a supervisory role to monitor automated processes (Moray, 1986; Sheridan & Verplank, 1978) or are removed from the control loop altogether (Kaber & Endsley, 2004), automated decision aids are designed to support human operators. In other words, both human and automation-generated judgments contribute to automation-aided performance, making it possible for human-automation teams to exceed the performance of either individual (Corcoran, Dennett, & Carpenter, 1972; Dalal & Kasper, 1994; Parasuraman, 1987; Thackray & Touchstone, 1989).

Unless a distinction is trivially easy, decisions pertaining to the presence or absence of a signal are based on probabilistic data. That is, there is variation in the way that signal and noise events present that makes it almost impossible to tell them apart with perfect accuracy. Like those of their human counterparts, an automated aid’s judgments are constrained by the uncertainty

inherent in the data and by noise introduced by data-gathering instruments (C.D. Wickens, 2000). Ideally, assistance from an aid will improve the operator's decision making even if the aid is not perfectly reliable (e.g., Bartlett & McCarley, 2017; 2019; Wickens & Dixon, 2007). Automated target detection systems, for instance, have been found to enhance human performance even with relatively low automation reliability (i.e. hit rates) of 70 to 75% (Reiner, Hollands & Jamieson, 2017; Yeh & Wickens, 2001). However, operators often disregard or underweight an imperfect aid's judgments, producing automation-aided performance that falls short of achievable levels (Bartlett & McCarley, 2017; 2019; Parasuraman, 2000; Parasuraman & Riley, 1997; Wickens & Dixon, 2007).

1.2 Signal Detection Theory

The efficiency of an operator's automation use can be assessed by examining the sensitivity of the human-automation team relative to statistically ideal levels. Given that automation-aided performance constitutes a form of collaborative decision-making, signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005) provides a suitable framework for this analysis. SDT models decision-makers' ability to discriminate between potential states of the world, termed *signal* and *noise*. Precise definitions are task-specific, but generally, noise is random background variation with no discernible pattern (e.g., normal fluctuation in power plant temperature gauges), whereas a signal is an information-bearing pattern (e.g.,

a dangerously high temperature gauge reading) added to the noise. The terms can also be applied to two discernible but different patterns of information—one arbitrarily labeled signal and the other noise—as in discriminations between truths and lies, healthy and diseased patients, and guilty and innocent defendants (Stanislaw & Todorov, 1999).

Signal and noise events are associated with separate evidence distributions. SDT assumes that upon observing a stimulus, a decision-maker assesses the evidence to decide whether it originated from the signal or noise distribution. On average, larger evidence values are observed when a signal is present than when it is absent, but so long as the evidence distributions overlap, the decision-maker's judgments will be uncertain. The decision-maker's ability to distinguish between the two alternatives, as determined by the overlap between signal and noise distributions, is termed *sensitivity*. Under the assumption that signal and noise distributions are normal and of equal variance, sensitivity can be measured with the statistic *d'*.

Whether a decision-maker renders a signal or noise judgment depends on the decision criterion adopted. Evidence values that exceed the criterion elicit signal judgments and values that fall below the criterion elicit noise judgments. An unbiased criterion position is halfway between the signal and noise distributions such that signal and noise judgments are equally likely. A decision-maker's criterion may be *liberal*, biased toward signal judgments, c*onservative*, biased toward noise judgments, or unbiased. The distance from the decision-maker's criterion to the unbiased position halfway between the

signal and noise distributions provides a measure of *response bias*, measured with the statistic *c*.

Response bias can also be described in terms of the relative likelihood of obtaining the evidence value on a signal versus a noise trial. The alternative response bias measure, β, corresponds to the likelihood ratio of the height of the signal distribution to that of the noise distribution at the observer's criterion. Likelihood ratios that exceed β elicit signal present judgments, whereas ratios that fall below β elicit noise judgments. The natural logarithm of β is analyzed in place of β to convert response bias to a symmetrical scale on which negative values indicate liberal bias, and positive values indicate conservative bias. Log β is given by the equation,

$$\ln \beta = d' \times c, \tag{1}$$

where *d'* and *c* are measures of the decision-maker's sensitivity and criterion placement, in units of the standard deviation of the signal and noise evidence distributions. Assuming a symmetrical payoff matrix, the optimal response bias is determined by the base rates of signal and noise events and is given by,

$$\beta^* = \frac{p(noise)}{p(signal)}.$$

When the true state of the world is signal, signal responses are correct and are called *hit*s. Conversely, when the true state of the world is noise, signal responses are incorrect and are called *false alarms* (Green & Swets, 1966). The combination of two stimulus classes (signal present, signal absent) and two response classes (signal, noise) produces four stimulus-response events that describe signal detection performance, as shown in Figure 1. The proportions of signal trials judged as signal and noise are termed hit rate (HR) and false-alarm rate (FAR), respectively, and are used to calculate *d'* and *c*.

	Signal present	Signal absent
Respond signal	Hit	False alarm
Respond noise	Miss	Correct rejection

Figure 1. Stimulus-response events for signal detection judgments.

Just as human operators make signal detection judgments by comparing the strength of observed evidence to a criterion, decision aids reach diagnoses by comparing data to a designer-specified criterion (Rice & McCarley, 2011). A human operator receiving assistance from an automated decision aid can therefore be thought of as a team of two agents who

independently evaluate a stimulus, estimate the strength of evidence, then integrate their judgments to arrive at a joint decision.

As discussed in section 1.1.1, inherently ambiguous data and noisy sensors mean that neither the operator nor the aid will be able to make correct judgments all the time. However, so long as both the human and the aid have sensitivity greater than 0, appropriately combining judgments makes it possible to achieve automation-aided sensitivity beyond what either agent could achieve alone (Macmillan & Creelman, 2005).

1.2.1 Using graded automation cues

Assuming that an automated aid reports its evidence values for or against either state of the world directly—that is, that the aid provides an assessment of signal strength on a continuous scale (Bartlett & McCarley, 2017; 2019)—the operator's ideal strategy for using the aid is to average his or her own assessment of signal strength with the aid's, weighting each assessment by the decision maker's average *d'* (Bahrami et al, 2010; Sorkin & Dai, 1994). Automation-aided sensitivity under the optimal weighting (OW) model, d'_{OW}, is

$$d'_{OW} = \sqrt{{d'_{operator}}^2 + {d'_{aid}}^2}.$$

An alternative strategy, and one that is generally less efficient, is for the operator to average his or her estimate of signal strength with the aid's in an

unweighted manner (Sorkin et al., 2001). Automation-aided sensitivity under the unweighted (UW) model, d'_{UW}, is

$$d'_{UW} = \frac{\sqrt{d'_{operator} + d'_{aid}}}{2^{1/2}}.$$

The graded cues used in the OW and UW models preserve information about uncertainty that is lost when judgments are discretized. This information allows the operator to lend more credence to judgments accompanied by higher evidence values, which predicts better automation-aided sensitivity than models of binary cue use (Bartlett & McCarley, 2017; 2019).

1.2.2 Using binary automation cues

Instead of sharing estimates of signal strength directly, automated aids are often designed to convert their judgments to binary diagnoses (Dzindolet, Pierce, Beck, Dawe, & Anderson, 2001; Rice & McCarley, 2011). In these circumstances, researchers typically infer a suboptimal *contingent criterion* (CC) strategy (Robinson & Sorkin, 1985) from participants' performance (Bartlett & McCarley, 2017; Elvers & Elrif, 1997; Maltz & Meyer, 2001, Meyer, 2001, Wang et al., 2009). Under the CC model, an operator's response bias is contingent on the aid's judgment. Specifically, the operator is presumed to adopt a more liberal β when the aid judges signal, and a more conservative β when the aid judges noise. The difference between the two

values of β reflects the extent to which the operator relies on the aid's judgments.

Under a suboptimal CC model, operators adjust β in the direction of the aid's judgments, but to an inadequate degree. Thus, operators tend to be more conservative than they should be following a signal cue from the aid and more liberal than they should be following a noise cue (see Figure 2). This suboptimal cue-contingent bias is consistent with the *sluggish beta* phenomenon, whereby humans tend to set β closer to 1 than is optimal (Chi & Drury, 1998; Wang et al., 2009; Wickens & Hollands, 2000).

The criterion position that corresponds to β^* is determined by an operator's underlying signal and noise distributions. Rearranging equation 1 shows that optimal criterion placement, c^*, changes as d' increases or decreases (Lynn & Barrett, 2014):

$$c^* = \frac{\ln \beta^*}{d'}.$$

This formulation implies that to achieve the optimal criterion placement, operators must have knowledge of their own sensitivity as well as the aid's.

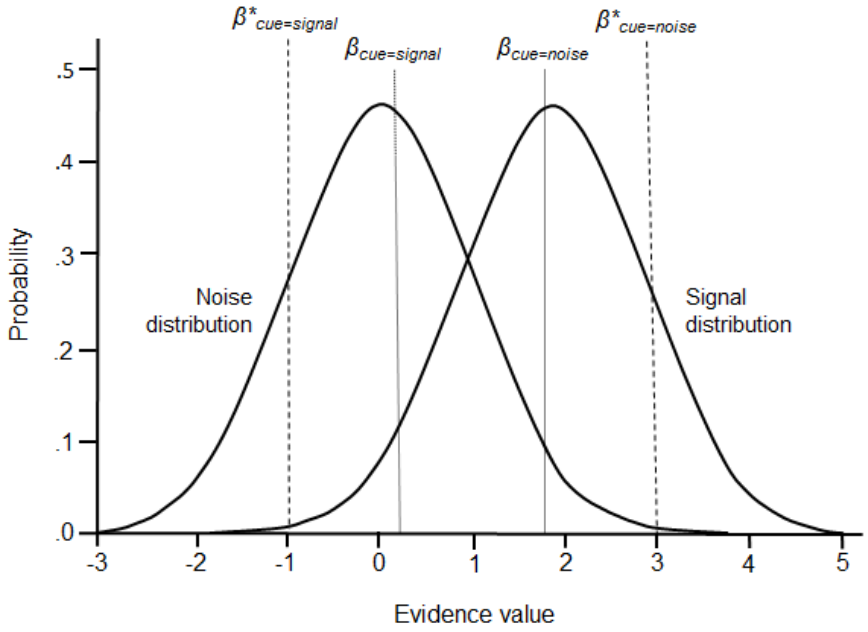

Figure 2. An example of cue contingent criteria that are less extreme than optimal. *β*cue = signal* indicates the hypothetical optimal *β* following a signal cue from the aid and *β*cue = noise* indicates the optimal *β* following a noise cue. *β*cue=signal* and *β*cue=noise* reflect hypothetical observed *β*s following their respective cues from the aid.

A less efficient strategy for using binary automation cues is the *best decides* (BD) model (Bahrami et al., 2010; Denkiewicz, Rączasek-Leonardi, Migdal, & Plewczynski, 2013). Under the BD model, an operator who is more sensitive than the aid will ignore the aid entirely, while an operator less sensitive than the aid will defer to the aid's judgments. Like the CC model, the BD model requires operators to know their own sensitivity, to the extent that they can determine whether they are more or less sensitive than the aid.

Other strategies are less efficient, still. The *probability matching* (PM; Bliss, Gilson, & Deaton, 1995; Wiegmann, 2002) and *coin flip* (CF; Bahrami

et al., 2010) models, for instance, assume that when the human and aid agree, the agreed-upon judgment becomes the team decision. Under the *PM* model, disagreements are resolved by deferring to the aid's judgment with a probability equal to the aid's average reliability. For example, if an aid correctly identifies signals 80% of the time, the *PM* model assumes that the operator relies on the aid's judgment for 80% of the trials on which they disagree. Under the *CF* model, disagreements are resolved using a coin flip strategy to decide between alternatives. The PM and CF models offer highly inefficient strategies for using cues from an aid, but so long as the aid is more accurate than the operator, the PM model confers a slight advantage.

1.2 Metacognition

The ability to judge their self-performance may influence how well human operators use automated decision aids. The OW and UW models of combining judgments, for instance, assume that on a trial-to-trial basis, operators average their own assessment of signal strength with the aid's assessment. Accurately judging the strength of an observed signal requires an operator to have an accurate internal representation of the underlying signal and noise distributions. The OW model also requires operators to weight their judgments according to their average sensitivity, which assumes some awareness of the extent to which the distributions overlap. Similarly, the CC model assumes that operators set their criteria according to their perceived

sensitivity, and the BD model assumes that operators know, at the very least, whether they are more or less sensitive than the aid.

In general, the awareness of one's own cognitive abilities, processes, and resources is termed *metacognition* (Garafalo & Lester, 1985). Decisions are not made in isolation but are accompanied by assessments of decision quality. These metacognitive assessments allow people to monitor their task performance and adjust their behavior accordingly. The capacity for accurate introspection varies across individuals (Fleming, Weil, Nagy, Dolan & Rees, 2011; Kelemen, Frost & Weaver, 2000) and can be operationalized as the extent to which an observer's confidence ratings predict the accuracy of his or her judgments. Observers with good metacognition are more likely to be correct when they are confident and less likely to be correct when they are not confident. Conversely, observers with poor metacognition are worse at discriminating between their own correct and incorrect decisions, producing weaker associations between decision accuracy and confidence. Poor metacognition may present, for example, as overconfidence in incorrect judgments and underconfidence in correct judgments.

1.2.1 Type 2 signal detection theory

Conceptualizing a metacognitive judgment as a secondary discrimination task allows analysis within the framework of SDT (Maniscalco & Lau, 2012). While *Type 1* tasks require operators to discriminate between signal and noise events, *Type 2* tasks require them to discriminate between

their own correct and incorrect Type 1 judgments—an ability termed *metacognitive sensitivity* (Maniscalco & Lau, 2012). Metacognitive sensitivity is estimated from a decision maker's confidence ratings for Type 1 judgments and is measured with the statistic meta-*d'*.

Analogous to Type 1 signal detection judgments, performance can be described by the combination of two stimulus classes (correct decision, incorrect decision) and two response classes (high confidence, low confidence), producing four stimulus-response events (see Figure 3).

	Correct decision	Incorrect decision
High confidence	Hit	False alarm
Low confidence	Miss	Correct rejection

Figure 3. Stimulus-response events for Type 2 signal detection judgments.

Unless metacognitive judgments incorporate additional information not used in the Type 1 judgment, *d'* places an upper limit on the decision maker's meta-*d'*. To control for Type 1 performance, the ratio of meta-*d'* to *d'* is taken as a measure of *metacognitive efficiency*. Given that many strategies of integrating human and aid judgments rely on metacognition, metacognitive efficiency may predict the efficiency of automation use.

1.3 Numeracy

Automated aids that share their sampled evidence values directly provide operators with the information necessary to achieve optimal automation-aided performance (Sorkin & Dai, 1994). In practice, however, the evidence that graded cues improve performance is mixed. Although some studies have found better human-automation sensitivity with graded than with binary cues (e.g., McCarley, 2009; Ragsdale, Lew, Dyre, & Boring, 2012; Sorkin, Kantowitz, & Kantowitz, 1988; St. John & Manes, 2002; Wiczorek & Manzey, 2014), others have shown no benefit (Bartlett & McCarley, 2017, Wickens & Colcombe, 2007; Wiczorek, Manzey, & Zirk, 2014). In fact, Bartlett & McCarley (2017) found that even when the aid reported graded cues, participants relied only on the binary judgments, ignoring the fine-grained assessments of signal strength available to them.

Perhaps, operators do not use graded automation cues due to an unwillingness or inability to interpret numeric information. Performance is consistent with this interpretation regardless of whether an aid provides its estimate of signal strength as a raw value, likelihood ratio, confidence rating, or verbal expression of confidence (Bartlett & McCarley, 2019). This suggests that the issue might not be simply an aversion to interpreting numbers (e.g., ratios and percentages), but with comprehending probabilistic information more generally.

Probabilistic or statistical reasoning is one of four categories that comprise general numeracy (among basic, computational, and analytical;

Golbeck, Ahlers-Schmidt, Paschal & Dismuke, 2005)—a strong predictor of general decision-making skill (Garcia-Retamero & Cokely, 2013; 14; Ghazal, Cokely & Garcia-Retamero, 2014; Peters et al., 2006). Highly numerate individuals, for example, are more likely to infer stronger and more precise meanings from numerical information (Peters et al., 2006), are less susceptible to framing effects (Peters et al., 2006), and are more willing to engage in shared decision-making (Galesic & Garcia-Retamero, 2011) than those who are less numerate. Beyond interpreting and understanding the aid's confidence estimates, operators must also use this information to integrate the aid's judgment with their own. If individual differences in numeracy affect an operator's ability to interpret and weight cues from an aid, then in addition to metacognition, numeracy may predict the efficiency of automation use.

1.4 The Current Study

The current study examined whether metacognition and numeracy are associated with the efficiency of automation use in a signal detection task. Participants viewed a series of blue and orange random-dot images and were asked to judge the dominant color on each trial. Participants performed the task alone or with assistance from an automated decision aid that provided binary judgments with confidence estimates. Type 1 and Type 2 signal detection measures were estimated for unaided and aided conditions, and numeracy was measured with the Subjective Numeracy Scale (SNS; Fagerlin

et al., 2007). The efficiency of automation use was assessed using a response bias error approach adapted from Wang and colleagues (2008).

1.5.1 Subjective Numeracy Scale

The SNS (Fagerlin et al., 2007) is an 8-item self-report measure of perceived mathematical ability and preferences (see Appendix for scale). The scale consists of two 4-item subscales: the ability subscale, on which respondents rate their ability to perform mathematical tasks involving fractions and percentages; and the preferences subscale, on which respondents rate their preference for information presented in tables, graphs, and numbers. The SNS has high internal reliability (Cronbach's $\alpha = 0.82$) and correlates highly ($r = 0.68$) with scores on the Objective Numeracy Scale (ONS; Lipkus, Samsa & Rimer, 2001). The ONS is an 11-item numeracy test that assesses respondents' ability to perform mathematical tasks involving frequencies, proportions, and probabilities. Further, scores on the SNS predict the likelihood of correctly recalling and interpreting risk information and eliciting utilities in a medical decision-making context (Zikmund-Fisher, Smith, Ubel, Fagerlin).

1.5.2 Measuring automation use

The correlation between an automated aid's cues and an operator's responses appears to provide an intuitive measure of the extent to which the operator relies on the automation (e.g., Bisantz & Pritchett, 2003; Brunswik,

1956). However, a highly sensitive operator will tend to arrive at the same conclusion as a highly sensitive aid irrespective of reliance, potentially leading to erroneous inferences. Alternatively, reliance can be measured as the extent to which *misuse* (incorrectly agreeing with the aid's judgment) exceeds *disuse* (incorrectly overriding the aid's judgment; e.g., Dzindolet, Pierce, Beck, et al., 2001; Dzindolet, Pierce, Pomranky, et al., 2001; Parasuraman & Riley, 1997). The problem with this method, however, is that the appropriateness of over- or underrelying on an aid varies depending on the aid's reliability (Wang, Jamieson & Hollands, 2008).

To address this, Wang, Jamieson and Hollands proposed a response bias difference approach, in which the difference between observed and optimal bias on trials on which the aid provides signal and noise cues indicates how much an operator over- or under-relies on the aid (Wang, Jamieson & Hollands, 2008; 2009). The extent to which this difference of differences is suboptimal will be referred to as *bias error*, and will serve as the measure of the efficiency of automation use.

1.5.3 Hypotheses

We predicted that:

1. Participants' *d'* would be higher on aided trials than unaided trials. This served as a manipulation check for whether the aid improved sensitivity.
2. Metacognitive efficiency (meta-*d'*/*d'*) would be higher for aided trials than for unaided trials.

3. Unaided metacognitive efficiency would be negatively correlated with bias error.
4. Numeracy scores would be negatively correlated with bias error.

2 METHOD

2.1 Preregistration

Two experiments were separately preregistered on Open Science Framework prior to any observation of the data. Experiment 1 (see https://osf.io/56uxj) was directly replicated by Experiment 2 (see https://osf.io/huj7g/) and the data were aggregated to provide greater statistical sensitivity. All data, analyses, and results are available at https://osf.io/zye9r/, both for the separate and combined experiments.

2.2 Participants

Participants were 256 undergraduate students (M_{age} = 19.84 years, *SD* = 3.52; 70 males, 186 females) recruited from Oregon State University. Preregistrations for Experiments 1 and 2 specified target sample sizes of 100 and 120 participants who met all inclusion criteria, respectively. Samples of this size were selected to allow us to detect minimum correlations of 0.2 between metacognition and the efficiency of automation use, and numeracy and the efficiency of automation use.

As preregistered, data were excluded from participants who failed to achieve *d'* scores of at least 0.25 in both the unaided and aided conditions, and from participants who failed to use at least three levels of the four-level confidence rating scale. Exclusions left 221 participants (98 from Experiment 1, 123 from Experiment 2) for analysis. Note that the sample size for Experiment 2 was larger than preregistered because more participants became

available than expected. Data were not observed before running the additional participants. All participants gave informed consent and received course credit for an experimental session that lasted approximately 60 minutes. Participants were fluent in English and screened for normal color vision (Ishihara, 1918) and normal or corrected-to-normal visual acuity (Snellen, 1862).

2.3 Apparatus and stimuli

The experimental task was controlled by software written in PsychoPy (Peirce, 2007; 2009) and run on an Apple Mac Mini computer. Stimuli were presented on a 24-inch monitor with 1920 x 1080px resolution and a 60Hz refresh rate. Participants viewed the monitor at distance of approximately 65cm, with head position unconstrained. Stimuli were randomly generated blue and orange dot images (256 x 256px) that were either blue or orange dominant. Each pixel was assigned the dominant color with a probability of 0.51 (see Figure 4) and the alternative color with a probability of 0.49.

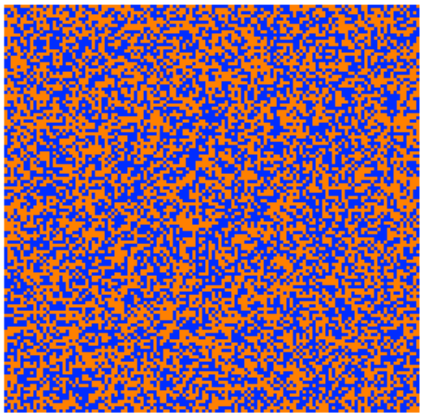

Figure 4. A sample orange-dominant stimulus.

2.4 Procedure

Participants performed a two-alternative forced-choice task modeled after that used by Bartlett and McCarley (2017, 2019). Instructions asked participants to imagine they were geologists sorting samples of a mineral into blue and orange strains. The only difference between strains was that the blue strain appeared slightly more blue, and the orange strain appeared slightly more orange, but participants were informed that the strains were impossible to sort with 100% accuracy. Participants viewed each sample and judged its dominant color.

On some trials, participants were assisted by a decision aid that provided a binary blue or orange diagnosis of the stimulus, accompanied by a confidence estimate. The aid's judgments were calculated using a standard equal-variance normal signal detection model (Macmillan & Creelman, 2005; C.D. Wickens, 2002). Evidence values were sampled from a Gaussian distribution with a mean of -1 for blue-dominant stimuli or 1 for orange-dominant stimuli, and a standard deviation of 1. The aid therefore had a *d'* of 2—the absolute difference between the means of the signal and noise distributions.

The aid transformed evidence values into binary judgments by comparing them to an unbiased criterion ($c = 0$) for distinguishing signal (orange) from noise (blue). If the sampled evidence value was less than 0, the aid judged *blue*. If the evidence value was greater than zero, the aid judged *orange*. The aid's *d'* of 2, coupled with an unbiased criterion, produced an

average accuracy of 84%. The aid's confidence estimates were calculated by converting the likelihood ratio of the aid's sampled evidence value to a posterior probability, then converting it to a percentage. The likelihood ratio of the aid's sampled evidence value is given by the equation,

$$LR = \frac{P(evidence\ value \mid diagnosis = truth)}{P(evidence\ value \mid diagnosis \neq truth)},$$

where *diagnosis = truth* denotes that the aid's binary judgment (blue- or orange-dominant) was correct, and *diagnosis ≠ truth* indicates that the aid's binary judgment was incorrect. If the likelihood ratio favors the signal response, the posterior probability that the aid's judgment is correct is given by

$$P(diagnosis = truth) = \frac{P(evidence\ value \mid diagnosis = truth)}{P(evidence\ value \mid diagnosis = truth) + P(evidence\ value \mid diagnosis \neq truth)}.$$

The probability of the selected response was always ≥ 0.5, owing to equal base rates of signal and noise events and the aid's unbiased criterion. Posterior probabilities were scaled to a 0-1 range and converted to a percentage, such that a probability of 0.5 produced a confidence estimate of 0%. The aid's reported confidence estimates were given by,

$$Conf = \frac{P(diagnosis = truth) - 0.5}{0.5} \times 100.$$

A higher confidence estimate indicated stronger evidence in favor of the aid's diagnosis.

Figure 3 shows the sequence of events in an automation-aided trial. Each trial was initiated by a mouse click from the participant, followed by the stimulus display. On aided trials, participants were provided with the aid's assessment, for example, 'Orange 31%'. On unaided trials, participants were provided with the neutral message, 'No reading'. A rating scale with the options 'Definitely blue', 'Probably blue', 'Probably orange', and 'Definitely orange' appeared onscreen beneath the stimulus. The stimulus display remained onscreen until the participant responded by clicking a rating. New stimulus images were generated each trial and the dominant stimulus color was selected randomly.

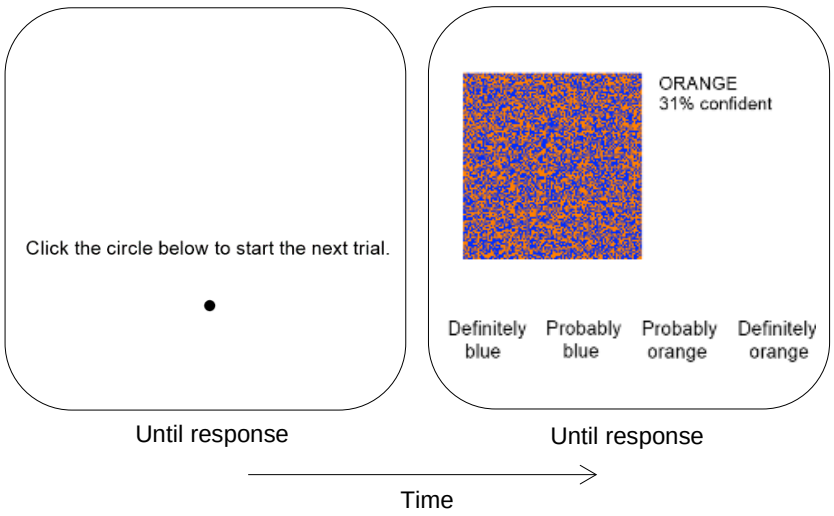

Figure 5. The sequence of events within an automation-aided trial.

Before performing the experimental trials, participants completed 50 unaided practice trials, followed by 50 aided practice trials. During practice, each response was followed by a 1,500ms feedback message of either ‘Correct!’ or ‘Incorrect!’ to allow participants to gauge their own task performance and that of the aid. After practice, unaided and aided experimental trials were run in separate blocks of 100 trials, with the order of blocks randomized across participants. Feedback was withheld on experimental trials to minimize changes in performance due to learning.

At the end of the task, participants completed the 8-item Subjective Numeracy Scale (see Appendix A), which gauged their self-assessed numerical ability and preferences for the presentation of numerical and probabilistic information (Fagerlin et al., 2007). Participants responded on a 6-point Likert-type scale. Responses were reverse coded where appropriate and summed to produce a subjective numeracy score for each participant.

3 ANALYSIS

For signal detection analysis, orange-dominant stimuli were arbitrarily treated as signal events and blue-dominant stimuli as noise events. Analyses used Bayesian estimation procedures to reallocate credibility from prior probability distributions to posterior distributions that are consistent with the observed data. In line with recommended practice for estimating continuous parameters with little prior knowledge about their values (Kruschke, 2013, 2015; Kruschke, Aguinis, & Joo, 2012, Kruschke & Liddell, 2017), analyses specified vague, noncommittal prior distributions on the parameters. Vague priors place similar credibility across the range of possible values, and therefore have trivial influence on the posterior distribution (Edwards, Lindman & Savage, 1963; Kruschke, 2010; 2014, Lindley, 1961).

Prior distributions were updated through probabilistic sampling using the JAGS package (Plummer, 2015) in R to approximate the posterior distributions. All parameter estimates were based on four Markov chain Monte Carlo (MCMC) simulations, run for 10,000 burn-in steps, followed by 250,000 sample steps each. MCMC chains were thinned to every fifth step to reduce autocorrelation, leaving 50,000 samples for analysis. 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).

To begin, Type 1 sensitivity, *d'*, and metacognitive sensitivity, *meta-d'*, were estimated for each subject. Next, a hierarchical procedure was used to

estimate group mean *d'* and meta-*d'* for unaided and aided task conditions. Finally, correlations between metacognitive efficiency, subjective numeracy, and bias error were estimated.

3.1 Single-subject signal detection measures

Participants' hit and false-alarm rates were transformed into measures of Type 1 sensitivity, *d'*, and bias, *c* (Green & Swets, 1966; Macmillan & Creelman, 2005). Calculated manually, *d'* is the difference of the standardized hit rate (HR) and false-alarm rate (FAR; Stanislaw & Todorov, 1999),

$$d' = z(HR) - z(FAR),$$

where *HR* = *p*(signal response | signal) and *FAR* = *p*(signal response | noise). The bias measure *c* corresponds to the distance from the decision criterion to an unbiased position halfway between the signal and noise distributions, given by

$$c = -0.5 \times [z(HR) - z(FAR)].$$

Using Fleming's (2017) single-subject Bayesian estimation procedure, values of *d'* and *c* parameters were inferred from the observed hit and false alarm counts and the total number of signal (*S*) and noise (*N*) trials. Counts of hits and false alarms were assigned binomial likelihood distributions, such that

Hits ~ Binomial (*HR*, *S*)

False alarms ~ Binomial (*FAR*, *N*).

The bias measure, log β, was calculated at each step of the MCMC chain using the formula (T.D. Wickens, 2000),

$$\ln \beta = d' \times c.$$

Confidence ratings conditional on correct and incorrect decisions were transformed into estimates of Type 2 sensitivity, meta-*d'* (Maniscalco & Lau, 2012). Meta-*d'* reflects the *d'* that corresponds to the observed confidence ratings assuming a metacognitively optimal observer. Analogous to Type 1 signal detection measures, meta-*d*' was estimated from Type 2 hits (high confidence | correct decision) and false alarms (high confidence | incorrect decision). Signal detection measures for individual participants were obtained using Fleming's single-subject default prior distributions,

d' ~ dnorm(0, .5),

c ~ dnorm(0, 2),

meta-*d'* ~ dnorm(1, 0.5).

Meta-*d'* measures how much information confidence ratings carry about Type 1 performance, but is itself confounded by Type 1 performance

(Galvin, Podd, Drga & Whitmore, 2003). Of two observers who each make optimal use of the information available to their Type 1 judgment, the observer with the higher *d'* will achieve a higher meta-*d'*. In other words, meta-*d'* not only reflects differences in the quality of metacognitive evaluation, but in the quality of information being metacognitively evaluated.

Metacognitive efficiency was measured by calculating the ratio of meta-*d'* to *d'* (Fleming, 2017; Maniscalco & Lau, 2012). The meta-*d'*/*d'* ratio controls for Type 1 performance by quantifying how much of the metacognitive signal available for the Type 1 detection task was captured by confidence ratings. A ratio of 1 indicates that all the information used for the Type 1 decision was also available to the Type 2 decision, whereas values below 1 indicate suboptimal metacognition. A value greater than 1 indicates that confidence ratings incorporated information beyond that used to make the Type 1 decision (Fleming & Daw, 2017).

One potential method of measuring the association between metacognitive ability and automation use would be to calculate the correlation between unaided metacognitive efficiency,

$$meta_{eff} = \frac{meta\text{-}d'_{unaided}}{d'_{unaided}},$$

and human-automation efficiency,

$$team_{eff}\frac{d'_{aided}}{\sqrt{{d'_{unaided}}^2+{d'_{aid}}^2}}.$$

This approach runs the risk of a spurious correlation, however, because both measures include a common term, d'_{unaided}, in their denominator. As an alternative approach, the suboptimality of participants' automation use was assessed by comparing observed and optimal automation-aided performance, using an approach adapted from Wang, Jamieson, and Hollands (2009). Ignoring the aid's confidence estimates, optimal bias following a binary cue from the aid, as measured by the statistic $\ln \beta$ (Macmillan & Creelman, 2005), is,

$$\ln \beta^*_{\text{Cue} = i} = \ln \frac{p(signal \mid cue = i)}{p(noise \mid cue = i)},\ i = noise, signal.$$

For a participant using the aid's binary cues optimally, the difference in bias between trials on which the aid offers a 'noise' judgment and trials on which it offers a 'signal' judgment is therefore,

$$\Delta \ln \beta^* = \ln \frac{p(signal \mid cue = noise)}{p(noise \mid cue = noise)} - \ln \frac{p(signal \mid cue = signal)}{p(noise \mid cue = signal)}.$$

Correspondingly, the observed difference in bias between trials on which the

aid offers a ‘noise’ judgment and trials on which it offers a ‘signal’ judgment is,

$$\Delta \ln \beta = \ln \beta_{\text{Cue = noise}} - \ln \beta_{\text{Cue = signal}}.$$

The absolute difference between these differences therefore gives a measure of the participants’ deviation from optimal cue usage,

$$\text{Bias error} = \text{abs}(\Delta \ln \beta^{*} - \Delta \ln \beta)$$

3.2 Group means and difference scores

Kruschke’s (2013) robust hierarchical Bayesian parameter estimation procedure was used to calculate group mean *d’*, meta-*d’*, metacognitive efficiency, and mean difference scores from individual participants’ parameters, for unaided and aided task conditions. Hierarchical procedures estimate parameters at the individual- and group-level in the same model, allowing group-level parameters to be less influenced by single-subject estimates with a high degree of uncertainty, and single-subject estimates to be constrained by the group-level fit. Participant’s individual signal detection measures were estimated with the aforementioned single-subject optimization method prior to hierarchical estimation of group-level parameters, to ensure that estimates of individual differences were independent of one another.

Scores were assumed to follow *t*-distributions, with vague priors on their means, standard deviations, and normality parameters,

μ = N(mean = 0, SD = 1000)

σ = U(min = 1/1000, max = 1000)

ν = Exp(rate = 1/29)

3.3 Correlations

To assess whether the quality of automation use varied with metacognitive ability and numeracy, analyses estimated correlations between unaided metacognitive efficiency ratios and bias error, and subjective numeracy and bias error. Correlations were estimated using Kruschke's Bayesian model (2016), with vague priors on the means and Wishart priors on the inverse covariance matrices. The mean and 95% highest-density interval (HDI) were calculated to summarize the estimated posterior distributions of all parameters of interest (Kruschke, 2013). The 95% HDI contains 95% of the posterior distribution mass, thus, an effect is considered credible if the 95% HDI on the difference between conditions does not overlap 0.

Follow-up analyses employed Savage-Dickey ratios (Wagenmakers, Lodewyckx, Kuriyal & Grasman, 2010) to assess evidence for or against the hypotheses that the correlations were null. The Savage-Dickey ratio is the height of the posterior distribution divided by height of the prior distribution at the parameter value of interest, in this case, $r = 0$. The resulting Bayes

factor, denoted B_{01}, is the ratio of the likelihood of the data under the null hypothesis versus the alternative, and therefore summarizes the strength of the evidence for or against the null. The strength of evidence is interpreted according to the descriptive guidelines suggested by Jeffreys (1961) and modified by Wetzels and Wagenmakers (2012; see Appendix B).

4 RESULTS

Experiments 1 and 2 produced effects that were qualitatively equivalent, therefore, the following results describe the aggregated data. As a manipulation check, Hypothesis 1 predicted that participants' *d'* would be higher on aided than unaided trials. Automation-aided *d'*, M = 1.91, HDI [1.85, 1.97], exceeded unaided *d'*, M = 1.36, HDI [1.31, 1.41], M_{diff} = 0.55, [0.49, 0.61], confirming that assistance from the aid helped participants achieve higher sensitivity. However, automation-aided *d'* fell short of optimal aided *d'*, M = 2.42 [2.39, 2.45], $M_{efficiency}$ = 0.63, HDI [0.59, 0.67]. See Figure 6 for a comparison of mean unaided, aided, and optimal *d'*.

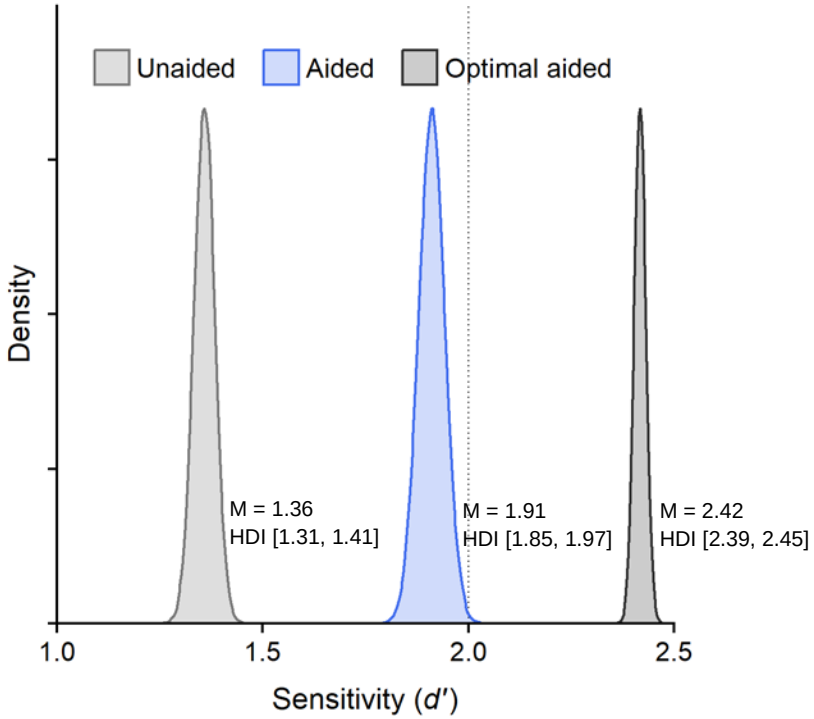

Figure 6. Estimated posterior distributions for unaided, aided and optimal *d'*. The dotted, vertical line represents the aid's *d'* of 2. HDIs reflect 95% highest density intervals.

Hypothesis 2 predicted that metacognitive efficiency would be higher for aided trials than for unaided trials. Automation-aided meta-*d'*, M = 1.29, HDI [1.21, 1.38], exceeded unaided meta-*d'*, M = 0.82, HDI [0.75, 0.90], M_{diff} = 0.55, HDI [0.49, 0.61], indicating that participants' raw metacognitive sensitivity was higher in aided blocks. Contrary to predictions, however, metacognitive efficiency did not differ credibly between the aided, M = 0.50, HDI = [0.44, 0.57], and unaided conditions, M = 0.40, HDI [0.33, 0.47], M_{diff} = 0.05, HDI [-0.02, 0.13] (see Figure 7).

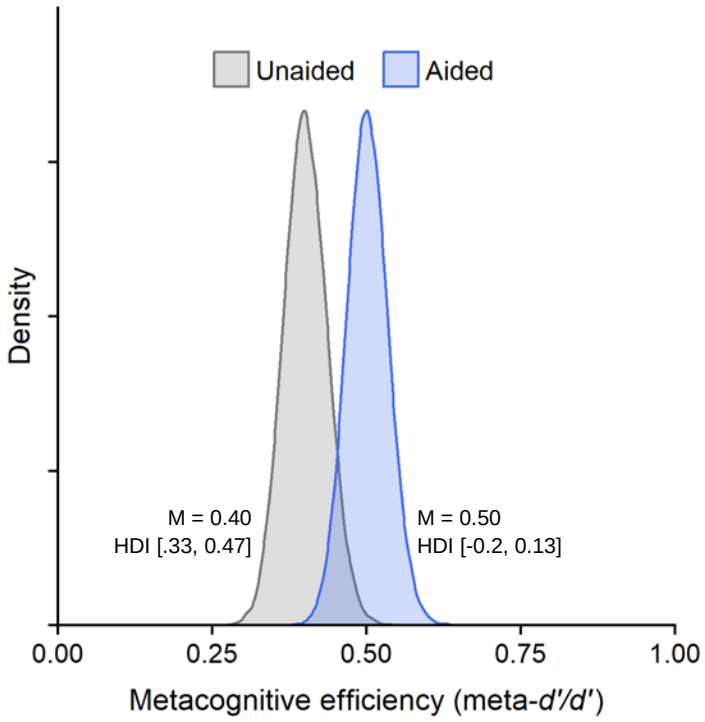

Figure 7. Estimated posterior distributions for unaided and aided metacognitive efficiency ratios. HDIs reflect 95% highest density intervals.

Hypothesis 3 predicted that unaided metacognitive efficiency would be negatively correlated with the suboptimality of automation use. Analyses examined the correlation between unaided metacognitive efficiency ratios and bias error relative to optimal CC criterion placement, finding a near-zero correlation, r = -0.01, HDI [-0.14, 0.12]. The Savage-Dickey density ratio provided strong evidence that the correlation between metacognitive performance and automation use suboptimality was null, B_{01} = 11.53 (see Figure 8). Response bias error was negatively correlated with human-automation team efficiency ratios, r = 0.65, 95% HDI [-0.75, -.52], confirming that the bias error measure captured the quality of automation use.

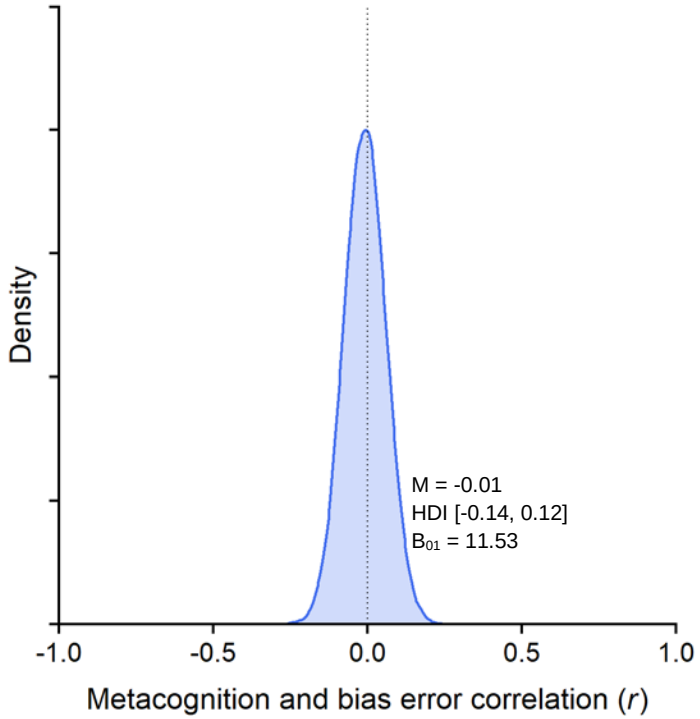

Figure 8. Estimated posterior distribution for the correlation between unaided metacognitive efficiency and bias error. HDI represents the 95% highest density interval. B01 is the Bayes factor for the null hypothesis that r = 0.

Hypothesis 4 predicted that numeracy would be negatively correlated with the suboptimality of automation use. Analyses examined the correlation between subjective numeracy scores and bias error relative to optimal CC criterion placement. Numeracy was nominally negatively correlated with suboptimality, r = -0.13, HDI = [-0.26, -0.00], but the HDI bordered on non-credibility, and the Savage-Dickey ratio provided only anecdotal evidence in favor of the null, B_{01} = 1.73 (see Figure 9).

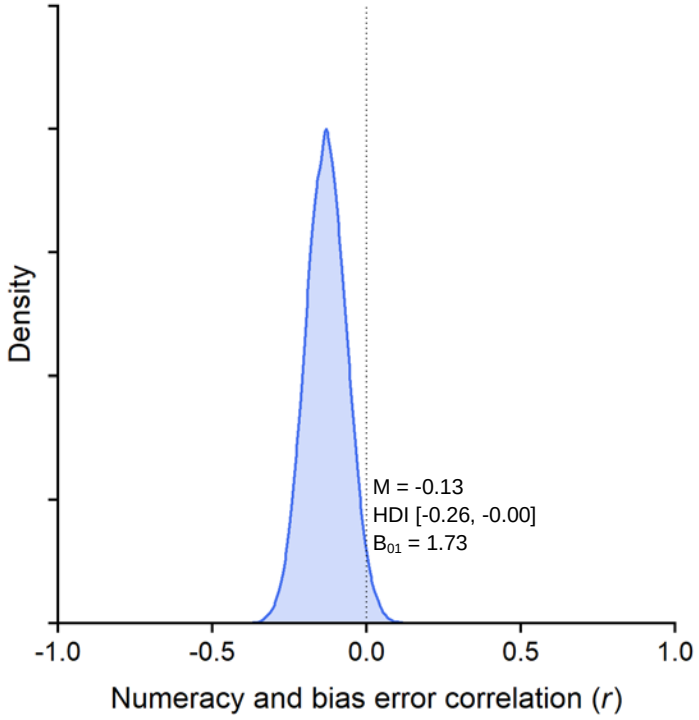

Figure 9. Estimated posterior distribution for the correlation between subjective numeracy scores and bias error. HDI represents the 95% highest density interval. B_{01} represents the Bayes factor for the null hypothesis that r = 0.

5 DISCUSSION

The current study examined the association between metacognition, numeracy, and automation-aided decision-making. As expected, participants achieved higher sensitivity when assisted by the aid. Despite improving relative to unaided performance, automation-aided sensitivity fell short of the level that was achievable based on the individual *d'*s of the operators and aid. This suggests, in line with previous findings (e.g., Bartlett & McCarley, 2017; Elvers & Elrif, 1997; Wang et al., 2009), that operators adopted a suboptimal strategy for integrating the aid's judgments with their own. Assistance from the aid also improved participants' raw metacognitive sensitivity. However, metacognitive efficiency, which controlled for Type 1 sensitivity, did not differ credibly between the unaided and aided conditions. The improvement in aided metacognitive sensitivity was therefore driven by increases in participants' Type 1 sensitivity, rather than metacognitive ability. Further contrary to predictions, neither metacognition nor numeracy were associated with the efficiency of automation use.

5.1 Metacognition and automation use strategy

The lack of association between automation use and metacognitive error could be interpreted as evidence that participants adopted a strategy of combining judgments that was not metacognitively driven. This interpretation would rule out the optimal weighting, uniform weighting, contingent criterion and best decides models, which each assume a metacognitive assessment of

some kind. However, operators may rely on metacognition in two ways when integrating judgments: estimating the strength of their sampled evidence on any given trial and estimating their average sensitivity. Meta-*d*', which assesses confidence ratings conditional on correct and incorrect judgments, is sensitive to metacognitive error in estimates of trial-to-trial signal strength but may not capture metacognitive error in estimates of average sensitivity.

It may be more appropriate, then, to conclude that performance is inconsistent with optimal and uniform weighting strategies, which produce aided decisions based on the aid's and operator's estimates of trial-to-trial signal strength. The findings do not rule out the possibility that participants adopted a strategy of automation use that assumes knowledge of average sensitivity, but not signal strength. With this is mind, performance is consistent with a suboptimal contingent criterion (CC) strategy (Robinson & Sorkin, 1985), in which an operator's criterion setting is 1) contingent on the aid's binary cue—albeit to a suboptimal extent, 2) a function of the operator's average sensitivity, but 3) unaffected by trial-to-trial signal strength.

Aided performance is also consistent with probability matching (PM) and coin flip (CF) strategies, which do not assume metacognitive awareness of average sensitivity or trial-to-trial signal strength. An earlier model-fitting analysis, however, favored a suboptimal CC model over the PM and CF models, which assumed operators did not have consistent tendencies to rely on automation (Bartlett & McCarley, 2019). Quantifying metacognitive error in

judgments of average sensitivity may provide more insight into plausible strategies in future studies.

Understanding how operators integrate an aid's judgments with their own has the potential to inform training and personnel selection strategies for better automation use. In general, individual differences in personality and cognitive ability are presumed to predict people's ability to accomplish a given task (Motowildo, Borman, Schmit, 1997), making some individuals more likely to succeed in a particular role than others. The present study hypothesized that metacognition would predict the efficiency of automation use since it is implicated in appropriately integrating human and aid judgments. However, the data suggest that metacognition is not likely to be a useful criterion for selecting operators or target of training interventions, given the strong evidence that metacognition was not associated with automation use.

5.2 Numeracy and automation use strategy

Data from analyses examining the association between subjective numeracy scores and bias error were indecisive, failing to provide strong evidence for or against the role of numeracy in automation use. Although we cannot rule out an association, that poor numeracy did not predict poor automation use suggests that numeracy is unlikely to be the limiting factor for poor automation-aided decision-making. That better numeracy did not meaningfully improve the efficiency of automation use suggests that

participants made little use of the aid's confidence estimates. Although this finding is is contrary to studies that have found graded automation cues to produce better human performance than binary cues (e.g., Sorkin, Kantowitz & Kantowitz, 1988; St. John & Manes, 2002; McCarley, 2009), it is consistent with Bartlett and McCarley's (2017) findings that participants relied only on an aid's binary judgments, even when direct estimates of signal strength were available. This finding may reflect people's tendency to favor simple heuristics over computationally effortful analyses when solving problems (Fiske & Taylor, 1991).

Graded cues from an aid make it possible for the human-automation team to achieve better sensitivity than the human or aid could achieve alone—a benchmark that observed automation-aided sensitivity falls drastically short of. The efficiency with which human operators incorporate information form an automated aid, and consequently the sensitivity of the human-automation team, may be improved by training operators to make better use of the aid's graded judgments.

An option not explored in a previous comparison of cue format (Bartlett & McCarley, 2019), is for an aid to display information graphically, rather than numerically or verbally. Graphical displays can be advantageous because they provide external representations of information, which frees up working memory for other cognitive tasks (Scaife & Rogers, 1996). In addition to offloading information storage, visual displays can offload cognitive processes onto perceptual processes (Card, Mackinlay &

Schneiderman, 1999; Scaife & Rogers, 1996). Emergent features of visual objects grouped together (Pomerantz & Pristach, 1989) can represent complex information using visual patterns (Hegarty, 2011), and replacing effortful computations with simpler pattern recognition processes might encourage operators to use cues from an automated aid more efficiently. Montgomery and Sorkin (1993) found preliminary evidence that when an automated aid presented judgments graphically, emergent features of visual displays improved signal detection sensitivity. Further research with much larger sample sizes will be necessary to examine whether the effect replicates.

Alternatively, optimal human-automation performance may be more readily achievable when each agent's judgments are integrated not by the human operator, who is prone to cognitive biases and error beyond that inherent in the task, but by the automated aid. Actuarial methods of judgment, which rely entirely on statistical algorithms, generally prove to be superior to clinical predictions made by humans (Ægisdóttir et al, 2006; Dawes, Faust & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Automated decision aids that integrate their judgments with those of a human operator may produce better performance still.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. The Counseling Psychologist, 34(3), 341-382.

Anand, V., Biondich, P. G., Liu, G., Rosenman, M., & Downs, S. M. (n.d.). *Child Health Improvement through Computer Automation: The CHICA System*. 5.

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*(9), 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Bally, J. (2004). The Thunderstorm Interactive Forecast System: Turning automated thunderstorm tracks into severe weather warnings. *Weather and Forecasting*, *19*(1), 64–72.

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking Aided Decision Making in a Signal Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(6), 881–900. https://doi.org/10.1177/0018720817700258

Bartlett, M. L., & McCarley, J. S. (2019). No Effect of Cue Format on Automation Dependence in an Aided Signal Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *61*(2), 169–190. https://doi.org/10.1177/0018720818802961

Bhana, H. S. (2009). *Correlating Boredom Proneness with Automation Complacency in Modern Airline Pilots* (Doctoral dissertation, University of North Dakota).

Bisantz, A. M., & Pritchett, A. R. (2003). Measuring the fit between human judgments and automated alerting algorithms: A study of collision detection. *Human Factors*, *45*(2), 266–280.

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics*, *38*(11), 2300–2312. https://doi.org/10.1080/00140139508925269

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Univ of California Press.

Card, S. K. (1999). Trees [M]. Readings in Information Visualization. Card SK, Mackinlay JD, Shneiderman B.

Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. *Topics in cognitive science*, *3*(3), 446-474.

Cesar, E. M. (1995). *Strategies for defining the Army's objective vision of command and control for the 21st Century*. Santa Monica, CA: Rand.

Chen, J. Y., & Barnes, M. J. (2012). Supervisory control of multiple robots in dynamic tasking environments. *Ergonomics*, *55*(9), 1043–1058.

Chen, J. Y. C., & Barnes, M. J. (2012). Supervisory Control of Multiple Robots: Effects of Imperfect Automation and Individual Differences. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(2), 157–174. https://doi.org/10.1177/0018720811435843

Chi, C.-F., & Drury, C. (1998). Do people choose an optimal response criterion in an inspection task? *IIE Transactions*, *30*(3), 257–266.

Clamann, M. P., & Kaber, D. B. (2003). Authority in adaptive automation applied to various stages of human-machine system information processing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 47, No. 3, pp. 543-547). Sage CA: Los Angeles, CA: SAGE Publications.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). *Trust in decision aids: A model and its training implications*. Presented at the in Proc. Command and Control Research and Technology Symp.

Corcoran, D., Dennett, J., & Carpenter, A. (1972). Cooperation of Listener and Computer in a Recognition Task. II. Effects of Computer Reliability and "Dependent" versus "Independent" Conditions. *The Journal of the Acoustical Society of America*, *52*(6B), 1613–1619.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, *91*(434), 883-904.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. Science, 243(4899), 1668-1674.

Dalal, N. P., & Kasper, G. M. (1994). The design of joint cognitive systems: the effect of cognitive coupling on performance. *International Journal of Human-Computer Studies*, *40*(4), 677–702.

Denkiewicz, M., Migda, 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 (Vol. 35, No. 35).

Dietrich, D., Fodor, G., Zucker, G., & Bruckner, D. (2010). *Simulating the mind: A technical neuropsychoanalytical approach*. Springer.

Dixon, S. R., & Wickens, C. D. (2006). Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 474–486. https://doi.org/10.1518/001872006778606822

Dzindolet, M. T., Beck, H. P., & Pierce, L. G. (2000). *Encouraging Human Operators to Appropriately Rely on Automated Decision Aids*.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

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*(3), 147–164. https://doi.org/10.1207/S15327876MP1303_2

Dzindolet, M. T., Pierce, L., Pomranky, R., Peterson, S., & Beck, H. (2001). Automation Reliance on a Combat Identification System. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *45*(4), 532–536. https://doi.org/10.1177/154193120104500456

Edwards, W., Lindman, H., & Savage, L. J. (19640101). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193. https://doi.org/10.1037/h0044139

Elvers, G. C. (1997). The effects of correlation and response bias in alerted monitor displays. *Human Factors; Santa Monica*, *39*(4), 570.

Elvers, G. C., & Elrif, P. (1997). The effects of correlation and response bias in alerted monitor displays. Human factors, 39(4), 570-580.Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian

hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381–394.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672–680. https://doi.org/10.1177/0272989X07304449

Fiske, S. T., & Taylor, S. E. (1991). Social cognition. Mcgraw-Hill Book Company.

Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1). https://doi.org/10.1093/nc/nix007

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science (New York, N.Y.)*, *329*(5998), 1541–1543. https://doi.org/10.1126/science.1191883

Galesic, M., & Garcia-Retamero, R. (20110509). Do low-numeracy people avoid shared decision making? *Health Psychology*, *30*(3), 336. https://doi.org/10.1037/a0022723

Gallistel, C. R. (2015). Bayes for Beginners: Probability and Likelihood. *APS Observer*, *28*(7). Retrieved from https://www.psychologicalscience.org/observer/bayes-for-beginners-probability-and-likelihood

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876.

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science*, *22*(5), 392–399.

Garcia-Retamero, R., & Cokely, E. T. (2014). The influence of skills, message frame, and visual aids on prevention of sexually transmitted diseases. *Journal of Behavioral Decision Making*, *27*(2), 179–189.

Garofalo, J., & Lester, F. K. (1985). Metacognition, Cognitive Monitoring, and Mathematical Performance. *Journal for Research in Mathematics Education*, *16*(3), 163–176. https://doi.org/10.2307/748391

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Making*, *9*(1), 15–34.

Golbeck, A. L., Ahlers-Schmidt, C. R., Paschal, A. M., & Dismuke, S. E. (2005). A Definition and Operational Framework for Health Numeracy. *American Journal of Preventive Medicine*, *29*(4), 375–376. https://doi.org/10.1016/j.amepre.2005.06.012

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. Psychological assessment, 12(1), 19.

Ishihara, S. (1918). Tests for Color blindness. *American Journal of Ophthalmology*, *1*(5), 376.

Joe, P., Dance, S., Lakshmanan, V., Heizenreder, D., James, P., Lang, P., … Yeung, H.-Y. (2012). Automated processing of Doppler radar data for severe weather warnings. In *Doppler Radar Observations-Weather Radar, Wind Profiler, Ionospheric Radar, and Other Advanced Applications*. IntechOpen.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113–153. https://doi.org/10.1080/1463922021000054335

Kaber, D. B., Omal, E., & Endsley, M. (1999). Level of automation effects on telerobot performance and human operator situation awareness and subjective workload. *Automation Technology and Human Performance: Current Research and Trends*, 165–170.

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, *28*(1), 92–107. https://doi.org/10.3758/BF03211579

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300. https://doi.org/10.1016/j.tics.2010.05.001

Kruschke, J. K. (2011). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. https://doi.org/10.1177/1745691611406925

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Kruschke, J. K. (2015). Introduction. In *Doing Bayesian Data Analysis* (pp. 15–32). https://doi.org/10.1016/B978-0-12-405888-0.00002-7

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time Has Come: Bayesian Methods for Data Analysis in the Organizational Sciences. *Organizational Research Methods*, *15*(4), 722–752. https://doi.org/10.1177/1094428112457829

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

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

Lee, S. J., Mo, K., & Seong, P. H. (2007, April). Development of an integrated decision support system to aid the cognitive activities of operators in main control rooms of nuclear power plants. In 2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making (pp. 146-152). IEEE.

Lindley, D. V. (1961). *The Use of Prior Probability Distributions in Statistical Inference and Decisions*.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General Performance on a Numeracy Scale among Highly Educated Samples. *Medical Decision Making*, *21*(1), 37–44. https://doi.org/10.1177/0272989X0102100105

Macmillan, N., & Creelman, C. (2005). Detection theory: a user's guide, 2nd edn New York. *NY: Lawrence Erlbaum Associates*.

Maltz, M., & Meyer, J. (2001). Use of Warnings in an Attentionally Demanding Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(2), 217–226. https://doi.org/10.1518/001872001775900931

Maltz, M., & Shinar, D. (2003). New Alternative Methods of Analyzing Human Behavior in Cued Target Acquisition. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*(2), 281–295. https://doi.org/10.1518/hfes.45.2.281.27239

Maltz, M., & Shinar, D. (2004). Imperfect In-Vehicle Collision Avoidance Warning Systems Can Aid Drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(2), 357–366. https://doi.org/10.1518/hfes.46.2.357.37348

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01455

McCarley, J. S. (2009). Response Criterion Placement Modulates the Benefits of Graded Alerting Systems in a Simulated Baggage Screening Task. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *53*(17), 1106–1110. https://doi.org/10.1518/107118109X12524443345078

Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors*, *47*(1), 35–49.

Meyer, J. (2001). Effects of Warning Validity and Proximity on Responses to Warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 563–572. https://doi.org/10.1518/001872001775870395

Montgomery, D. A., & Sorkin, R. D. (1993, October). The effects of display code and its relation to the optimal decision statistic in visual signal detection. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 37, No. 19, pp. 1325-1329). Sage CA: Los Angeles, CA: SAGE Publications.

Moray, N. (1986). Monitoring behavior and supervisory control.

Motowildo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance*, *10*(2), 71–83.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5–6), 527–539.

Oron-Gilad, T., Szalma, J., Thropp, J., & Hancock, P. (2005). Incorporating individual differences into the adaptive automation paradigm. *Human Factors in Organizational Design and Management VIII*, 581–586.

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. https://doi.org/10.1109/3468.844354

Parasuraman, R. (1987). Human-computer monitoring. *Human Factors*, *29*(6), 695–706.

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

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10.

Peters, E. (2012). Beyond Comprehension: The Role of Numeracy in Judgments and Decisions. *Current Directions in Psychological Science*, *21*(1), 31–35. https://doi.org/10.1177/0963721411429960

Prinzel III, L. J., Freeman, F. G., & Prinzel, H. D. (2005). Individual Differences in Complacency and Monitoring for Automation Failures. Individual Differences Research, 3(1).

Ragsdale, A., Lew, R., Dyre, B. P., & Boring, R. L. (2012). Fault Diagnosis with Multi-State Alarms in a Nuclear Power Control Simulator. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 2167–2171. https://doi.org/10.1177/1071181312561458

Reiner, A. J., Hollands, J. G., & Jamieson, G. A. (2017). Target Detection and Identification Performance Using an Automatic Target Detection System.

Human Factors: The Journal of the Human Factors and Ergonomics Society, *59*(2), 242–258. https://doi.org/10.1177/0018720816670768

Rice, S., & McCarley, J. S. (20110627). 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*(4), 320. https://doi.org/10.1037/a0024243

Riley, V. (1989). A General Model of Mixed-Initiative Human-Machine Systems. *Proceedings of the Human Factors Society Annual Meeting*, *33*(2), 124–128. https://doi.org/10.1177/154193128903300227

Robinson, D., & Sorkin, R. (1985). A contingent criterion model of computer assisted detection. *Trends in Ergonomics/Human Factors*, *2*, 75–82.

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. Reviews of human factors and ergonomics, 1(1), 89-129.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators:* https://doi.org/10.21236/ADA057655

Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Making*, *18*(1), 29–53.

Snellen, H. (1862). *Probebuchstaben zur Bestimmung der Sehscharfe*, *Utrecht*, *v: d*.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, *20*(4), 1787–1792. https://doi.org/10.1016/j.concog.2010.12.011

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

Sorkin, R. D., Hays, C. J., & West, R. (2001). *Signal-Detection Analysis of Group Decision Making*.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood Alarm Displays. *Human Factors*, *30*(4), 445–459. https://doi.org/10.1177/001872088803000406

St. John, M., & Manes, D. I. (2002). Making Unreliable Automation Useful. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *46*(3), 332–336. https://doi.org/10.1177/154193120204600325

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin*, *111*(1), 42.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Thackray, R. I., & Touchstone, R. M. (1989). Detection efficiency on an air traffic control monitoring task with and without computer aiding. *Aviation, Space, and Environmental Medicine*.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and Reliance on an Automated Combat Identification System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(3), 281–291. https://doi.org/10.1177/0018720809338842

Wang, L., Jamieson, G. A., & Hollands, J. G. (2008). Selecting methods for the analysis of reliance on automation. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 52, No. 4, pp. 287-291). Sage CA: Los Angeles, CA: SAGE Publications.

Wetzels, R., & Wagenmakers, E.J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064.

Weyer, J., Fink, R. D., & Adelt, F. (2015). Human–machine cooperation in smart cars. An empirical investigation of the loss-of-control thesis. *Safety Science*, *72*, 199–208. https://doi.org/10.1016/j.ssci.2014.09.004

Whitley, B. E., & Frieze, I. H. (1985). Children's causal attributions for success and failure in achievement settings: A meta-analysis. *Journal of Educational Psychology*, *77*(5), 608.

Wickens, C. D. (2000). *Imperfect and unreliable automation and its implications for attention allocation, information access and situation awareness*.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212.

Wickens, CD, & Hollands, J. (2000). Manual control. *Engineering Psychology and Human Performance*,.

Wickens, Christopher, & Colcombe, A. (2007). Dual-Task Performance Consequences of Imperfect Alerting Associated With a Cockpit Display of

Traffic Information. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(5), 839–850. https://doi.org/10.1518/001872007X230217

Wiczorek, R., & Manzey, D. (2014). Supporting Attention Allocation in Multitask Environments: Effects of Likelihood Alarm Systems on Trust, Behavior, and Performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(7), 1209–1221. https://doi.org/10.1177/0018720814528534

Wiegmann, D. A. (2002). Agreeing with Automated Diagnostic Aids: A Study of Users' Concurrence Strategies. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(1), 44–50. https://doi.org/10.1518/0018720024494847

Yeh, M., & Wickens, C. D. (2001). Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(3), 355–365. https://doi.org/10.1518/001872001775898269

Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., & Fagerlin, A. (2007). Validation of the Subjective Numeracy Scale: Effects of Low Numeracy on Comprehension of Risk Communications and Utility Elicitations. *Medical Decision Making*, *27*(5), 663–671. https://doi.org/10.1177/0272989X07303824

Appendices

Appendix A

Subjective Numeracy Scale (Fagerlin et al., 2007)

For each of the following questions, please check the box that best reflects **how good you are at doing the following things**:

1. How good are you at working with fractions?

❑1	❑2	❑3	❑4	❑5	❑6
Not at all good					**Extremely good**

2. How good are you at working with percentages?

❑1	❑2	❑3	❑4	❑5	❑6
Not at all good					**Extremely good**

3. How good are you at calculating a 15% tip?

❑1	❑2	❑3	❑4	❑5	❑6
Not at all good					**Extremely good**

4. How good are you at figuring out how much a shirt will cost if it is 25% off?

❑1	❑2	❑3	❑4	❑5	❑6
Not at all good					**Extremely good**

For each of the following questions, please check the box that **best reflects your answer**:

5. When reading the newspaper, how **helpful** do you find tables and graphs that are parts of a story?

❑1	❑2	❑3	❑4	❑5	❑6
Not at all helpful					**Extremely helpful**

6. When people tell you the chance of something happening, do you prefer that they use **words** ("it rarely happens") or **numbers** ("there's a 1% chance")?

❑1	❑2	❑3	❑4	❑5	❑6
Always Prefer Words					**Always Prefer Numbers**

7. When you hear a weather forecast, do you prefer predictions using **percentages** (e.g., "there will be a 20% chance of rain today") or predictions using only **words** (e.g., "there is a small chance of rain today")?

❑1	❑2	❑3	❑4	❑5	❑6
Always Prefer Percentages					**Always Prefer Words**

8. How **often** do you find numerical information to be useful?

❑1	❑2	❑3	❑4	❑5	❑6
Never					**Very Often**

Appendix B

Evidence categories for the Bayes factor

Bayes factor BF_{10}			Interpretation
	>	100	Decisive evidence for H_1
30	–	100	Very Strong evidence for H_1
10	–	30	Strong evidence for H_1
3	–	10	Substantial evidence for H_1
1	–	3	Anecdotal evidence for H_1
	1		No evidence
1/3	–	1	Anecdotal evidence for H_0
1/10	–	1/3	Substantial evidence for H_0
1/30	–	1/10	Strong evidence for H_0
1/100	–	1/30	Very Strong evidence for H_0
	<	1/100	Decisive evidence for H_0

(Jeffreys, 1961; Wetzels and Wagenmakers, 2012)