



Aviation Research Lab Institute of Aviation

**University of Illinois
at Urbana-Champaign
1 Airport Road
Savoy, Illinois 61874**

INTEGRATION OF INFORMATION SOURCES OF VARYING WEIGHTS: THE EFFECT OF DISPLAY FEATURES AND ATTENTION CUEING

**Christopher D. Wickens,
Heather L. Pringle, and James Merlo**

Technical Report ARL-99-2/FEDLAB-99-1

January 1999

Prepared for

**Dr. Michael Barnes, Monitor
U.S. Army Research Laboratory
Interactive Displays Federated Laboratory
National Missile Defense Subtask
Ft. Huachuca, AZ**

Contract DAAL 01-96-2-0003

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE JAN 1999		2. REPORT TYPE		3. DATES COVERED -	
4. TITLE AND SUBTITLE Integration of Information Sources of Varying Weights: The Effect of Display Features and Attention Cueing				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Army Research Laboratory, Interactive Displays Federated Lab, Fort Huachuca, AZ, 85613				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT see report					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES 28	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

ABSTRACT

This report reviews research in which multiple sources of variable reliability information are integrated for the purpose of making diagnostic judgments or allocating resources. A framework for considering these experiments is presented, and some evidence is presented regarding the extent to which humans are calibrated, in allocating processing proportionately to the ideal weights (i.e., reliability or importance) of information channels. Two generic sources of bias are identified. Attentional biases occur when more processing is given to less important channels, at the expense of more important ones (i.e., a failure to allocate attention optimally). Trust biases occur when less than fully reliable information is offered more processing than is warranted (i.e., "over trust"). A smaller number of specific studies are reviewed, and their conclusions are integrated, which have examined how multisource information processing is modulated by properties of the display of those sources. Two sources of display information are considered: attentional guidance, (e.g., cueing) directs attention to certain regions of the display. Reliability guidance explicitly displays the level of reliability of the information source(s). Each type of display can be explicitly designed to induce the appropriate behavior from the user, or can be a feature of the display that implicitly induces the relevant behavior. Generalizations regarding the effectiveness of these display features are sought from the studies reviewed.

1. INTRODUCTION

Consider any one of a number of situation assessment scenarios, represented abstractly in Figure 1: the military commander is confronted with a number of intelligence sources, and from these must assess the likelihood that the enemy will attack in a certain way. The ballistic missile defense officer is monitoring a display showing a number of target tracks, the reports from other intelligent analysts, and guidance from automation aids, and is trying to diagnose the highest priority tracks, in the effort to allocate defensive resources. The lost pilot is trying to gather information from ground sightings, airport compasses, maps, and potentially malfunctioning instrumentation, to assess where he is. The emergency room physician is trying to diagnose the condition of the patients based upon an unclear self report of symptoms, a medical history profile, a few tests, and now a set of alarms that are sounding from medical devices monitoring the patient's health.

In all of the above scenarios, the common feature is that the operator is attempting to integrate information from a variety of sources, in order to form a degree of belief, or diagnostic certainty, distributed across one or more possible hypotheses, as to the situation that is being monitored. As shown to the right of the figure, such a belief may generate three related outputs:

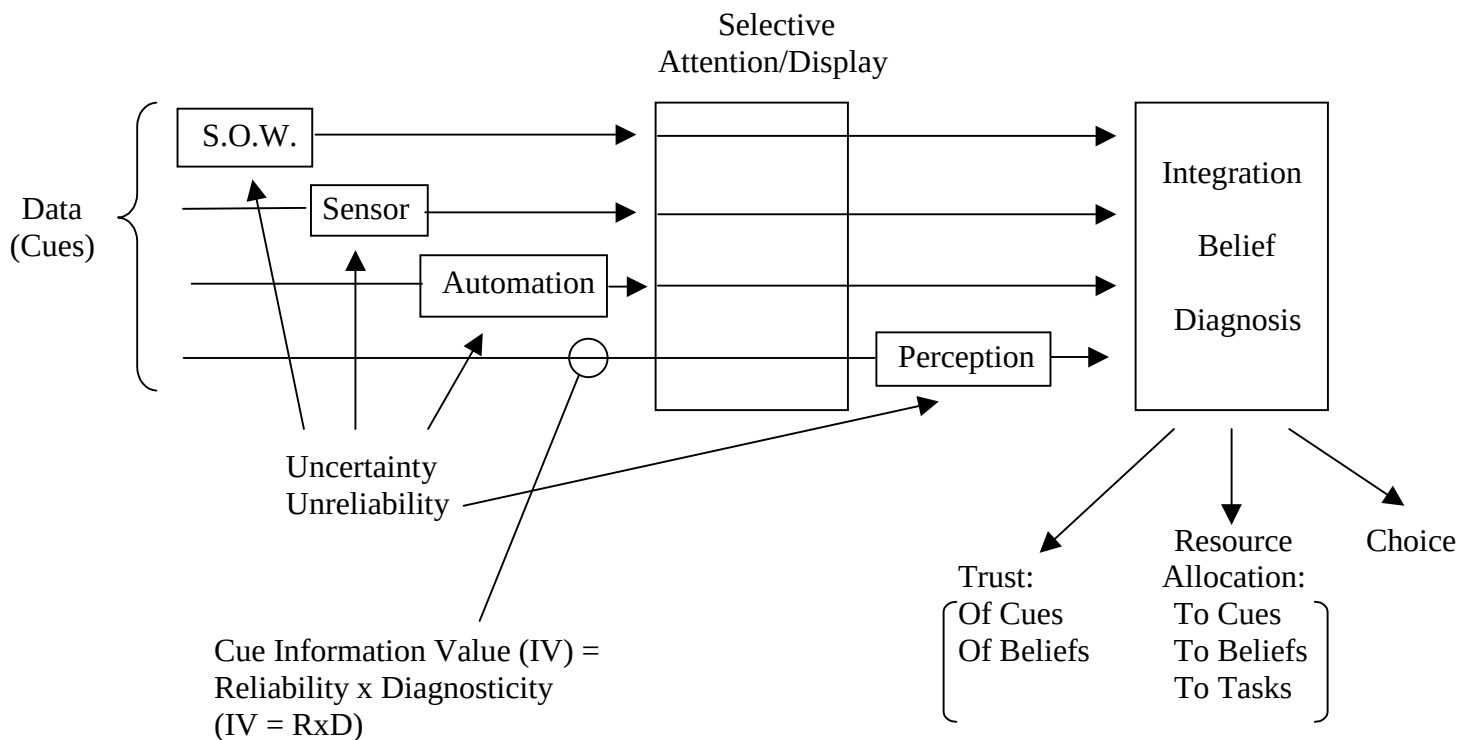


Figure 1.

1. The choice of an action, when required. For example, the military commander may choose a particular form of counterattack, the missile defense officer may choose to attack certain targets and not others, the lost pilot may decide to fly in a particular heading where he believes that a familiar landmark will be encountered, and the physician will choose a particular treatment.
2. The allocation of resources to tasks, hypotheses or channels. As an example in the first case, (allocation to tasks), the physician may decide that it is important to allocate most of her attentional resources to the task of stabilizing the patient's blood pressure, but to continue to allocate some cognitive effort to diagnose the trauma. Allocating resources to considering more than one hypothesis is of interest any time there remains some residual uncertainty, regarding the correctness of the favored hypothesis. As an example of allocating resources to hypotheses, the battle commander may decide that 70% of the combat resources should be allocated based on an assumption that the enemy will attack from the north, and reserve 30% to prepare for a less likely, but still possible, attack from the west. Allocating resources to channels, following a diagnosis, will be required if further information is sought from certain channels, to resolve uncertainty, or if new information channels are sought. As an example of this case, the lost pilot may pull out a different map to study, because the previous one provided no evidence matching with the visual view of the far domain; or the pilot may ask air traffic control to try to locate him on the radar picture.
3. The assessment of confidence, or degree of belief in the diagnosis, or set of possible hypotheses regarding the current situation. As we have noted above, one aspect of this confidence is the basis for preparation for alternative situations or hypotheses to the most favored one. If the missile defense officer is absolutely certain that one contact is hostile, and the remainder are not, then full resources can be deployed toward defense against the one. However, lack of absolute confidence would warrant consideration of the possibility that others might also be hostile. The key source of input to confidence in a hypothesis, is the information value of the set of information sources or cues that support the hypothesis in question (Barnett and Wickens, 1988; Wickens 1992). Analytically, we describe information value (IV) as the product of the validity of a cue (or its diagnosticity in discriminating one hypothesis from the other), and the reliability of the cue (degree of credibility assigned to its perceived value). If both validity and reliability can be scaled between 0 and 1 (as is typical of the correlation coefficient employed in testing and assessment), then the IV of a cue can also be scaled from 0 to 1. If either validity or reliability is 0, then IV will be 0, no matter how high the other term is. Furthermore, only if V and R are both 1, is the information value of a cue = 1.0. This is a special case, because if a single cue IV = 1, then, formally, it is the only cue that needs to be attended for diagnosis.. All other confirming cues are redundant, and any disconfirming cues (i.e., favoring the alternative hypothesis) must by definition be wrong.

Two aspects of the representation of information integration in Figure 1 present a particular challenge to human information processing and, by extension, to engineering psychology. First, the multiplicity of sources, (e.g., four sources in Figure 1), characteristic of many such processes, challenges the human's selective and divided attention capabilities (Wickens and Carswell, 1995), as well as their memory capabilities to integrate information over

time, if all information is not available simultaneously. Such memory will sometimes be biased in favor of the first arriving cue (primacy), sometimes in favor of the last (recency), and sometimes in favor of both, with cues arriving in the middle of a sequence generally having less weight (Hogarth and Einhorn, 1992). Second, as shown in the figure, all sources may be perturbed by unreliability, due to the inherent uncertainties of the world, failures in the sensors upon which the cue is derived, failures in any automation devices, designed to process the information, or failures in the operator's own perceptual or cognitive abilities. Here the critical issue is the extent to which the operator's trust or belief in what a cue indicates is calibrated with the actual value of the information provided by the cue.

Both the multiplicity of channels, and the unreliability of a single channel represent challenges in their own right, and to some extent this report will address, broadly, issues related to each alone. However, the most significant challenges arise when multiple channels vary in their information value, because here is where it is possible to define an "optimal behavior" to which operators should conform: that is, an allocation of resources that is based upon the correctly calibrated levels of belief. For example, if the commander is calibrated in his belief of 90% certainty that attack will be from the north, then it is reasonable to allocate 90% of the resources to that defense. If a particular cue is known to have $IV = 0$ (because it is either totally undiagnostic or unreliable), then it is rational to allocate no resources to its processing.

In the following report, we will consider ways of successfully inducing such calibration, primarily through the critical role of displays, although we recognize the important role of training in this process. However, as engineering psychologists, in order to understand the influences of display remediation, we must first establish the magnitude of miscalibration problems, and the circumstances that either amplify or diminish such problems.

2. RELEVANT LITERATURE

A more precise way of representing the relevant aspects of the process in Figure 1, which will provide a framework for the literature we review, is presented in Figure 2. The figure depicts three generic sources of information, represented as cues. As noted in Figure 1, the cues can vary in their information value. In Figure 2, two of the cues bear upon hypothesis 1; a third cue is relevant to hypothesis 2. The operator can allocate resources or effort to any of the three cues and/or to the two hypotheses. Allocating resources to cues typically involves visual or auditory attention. Allocating attention to hypotheses involves cognitive resources. Allocation to cues and to hypotheses may often be mutually facilitating, as when the operator allocates more resources to processing cue 3, he is inherently allocating more resources to the hypothesis (2) supported by cue 3. As a concrete example, one might consider hypothesis 1 being an enemy attack from the north, and hypothesis 2 being an enemy attack from the south. Three channels are available: airborne observations and sensor data (C_1 and C_2) indicate a north attack. The report of a local indicates that the south attack is likely.

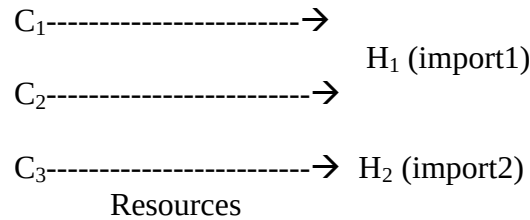


Figure 2: Representation of three cues, supporting two hypotheses.

The allocation of effort or processing across cues and hypotheses should be dictated by two factors: trust and importance. Cues that are more trustworthy (higher information value) should received relatively more processing, and weight in forming a diagnosis. Hypotheses that have relatively more evidence in their favor (trust), should also be anticipated to be more likely to be true than contradictory hypotheses supported by less evidence, and proportionately greater preparation should be made for conditions triggered by the likely hypothesis. However, this allocation should be modulated by importance. That is, if an unlikely hypothesis still has important consequences if it IS true, then some preparation should be made for its occurrence, and progressively more so, given the degree of importance. This preparation may involve allocation of physical resources to deal with the less likely hypothesis (for example, leave some combat resources in reserve to prepare for an attack from an unlikely corridor), or allocation of cognitive resources to prepare for its occurrence (for example, prepare an alternative strategy of response, should the unlikely attack occur).

Within the general framework of Figure 2, we can identify two ways of describing the calibration of the human cognitive response, with specification of the optimal response. We describe type T, or Trust calibration as a preparation for an alternative that is commensurate with the best estimate of the actual reliability, believability or trust in a source of information or a hypothesis. This calibration may apply to a hypothesis supported by a single source of information, as well as by many. We describe type A or Attention calibration, as uniquely applicable when there are multiple sources of information, which may be associated with a single hypothesis (C_1 and C_2), or with multiple hypotheses (those supporting H_1 vs. that supporting H_2). When either type T or type A calibration is violated, we can speak of a type A or type T **bias**. A type T or overtrust bias would characterize the operator who responds to an alarm that is inherently unreliable, or spends a long time analyzing the testimony of a witness who is a notorious liar. A type T undertrust bias is one in which the operator ignores a generally reliable alarm because it has produced a single false alarm in the past. A type A bias involves an allocation of attention between sources that is not appropriate given the relative information value of those sources, or the cost of ignoring a source (i.e., its importance). In the following, we review the sources of literature that have examined these sources of biases, first in isolation, and then in combination. Then we address the particular characteristics of display manipulations that have been shown to either enhance or mitigate their effect.

2.1 Type T Calibration and Bias

Evidence for biases in the degree of trust, or faith in a single source of information is relatively robust. It should be noted initially that there is a historical literature that suggests to some extent that people are calibrated as to the reliability or expectancy of information. For

example, the classic Hick-Hyman law of reaction time (Hyman, 1953), describes the human tendency to respond more rapidly to stimuli that are expected, and more slowly to those that are surprising, reflecting the modulation of preparation, based upon external probabilities. Indeed the tendency to perceive what is expected, and misperceive that which is surprising is another manifestation of this same general calibration of beliefs to real world probabilities.

Yet numerous departures from such calibration have also been observed. One such departure is the “sluggish beta” in signal detection (Green and Swets, 1988; Wickens, 1992), in which observers do not adjust their decision criterion as much as is warranted, by changes in the probability (expectancy) of a signal or by changes in the payoff matrix (importance in Figure 2). Such a failure is particularly pronounced in the case of varying probabilities. This bias appears to take on the form of “overpreparation” for very rare events, and seems to be consistent with some studies of probability estimation, in which the probability of very rare events is overestimated (see Wickens, 1992 for a review). (However, we will note some clear exceptions to this trend in the literature below.)

A third major source of data regarding trust, and expectancy has been in the area of human response to automation (Parasuraman and Riley, 1997; Wickens, Mavor, Parasuraman, and McGee, 1998). Here the issue is the extent to which humans trust automation more than is warranted, on the basis of the reliability of the automation (Muir, 1988; Lee and Moray, 1992). Parasuraman and his colleagues have used the term complacency to describe the circumstances of overtrust in which operators, believing automation to be relatively failsafe, fail to monitor it adequately, and hence, fail to intervene in a timely and appropriate fashion when it fails (Parasuraman et al., 1993). In the context of Figure 2, we describe the operator as “overpreparing” for one hypothesis (that the system will operate normally), and hence underpreparing for the alternative hypothesis. Such a tendency has been manifest for example in extremely long response times (30-50 seconds) for pilots to intervene when autopilots failed in simulated flights (Beringer, 1996), and seems to be more pronounced as automation monitors are removed progressively farther from the control loop (Kaber, Onal, and Endsley, 1998).

There are, in addition, examples of undertrust of automation. Often these occur at times when the automation has failed once, in a salient fashion (Lee and Moray, 1992), or is confusing, complex, and poorly displayed, so that it appears to do things that are inappropriate (Sarter and Woods, 1995). One powerful example of such undertrust is in human response to alarms (Sorkin, 1988). Alarms can be considered a kind of automation that monitors a continuous variable, and when some threshold value is exceeded notifies the operator with a salient (usually auditory) signal. When this threshold is set at too sensitive a level, “alarm false alarms” will sound, like the fabled “the boy who cried wolf.” Such examples of unreliability often lead operators to an unwarranted level of undertrust, in which they will simply ignore alarms that may, in fact, be true (Sorkin, 1998; Wickens, Gordon, and Liu, 1998). Solutions to this problem, inherent in display techniques, will be described in Section 3. Thus, we see examples of both under and overtrust in the area of automation effects. We will return to the issue of automation trust when we address the integration of T and A calibration effects.

A fourth example of T calibration bias may be found in the examination of the **overconfidence bias** in human judgment. Such a bias, well documented in decision research (Brenner et al., 1996; Fischhoff and MacGregor, 1982; Fischhoff, Slovic, and Lichtenstein, 1977;

Bjork, 1996; Henry and Snizek, 1993) describes circumstances when human's tend to assess the accuracy of their own actual or predicted performance as higher than is warranted. This is true whether people are assessing the accuracy of their own predictions about future events (Fischhoff and MacGregor, 1982), the safety of their own behavior (Svenson, 1981), the accuracy of their perception of "eyewitness events" (Wells et al., 1979), the accuracy of their long term memory for facts (Fischhoff et al., 1977), the viability of their learning strategies (Bjork, 1996), or of their decision and judgment processes (Brenner et al., 1996). In an interesting combination of self-overconfidence, and automation undertrust, Liu, Fuld, and Wickens (1993) found that people tended to trust their own abilities as monitors better than those of an automated system when in fact the two were equivalent.

The prevalence of overconfidence findings does not mean of course that these are always observed, and indeed there are certainly circumstances in which humans are underconfident in their own abilities, particularly when pitted against the guidance of automated systems (Conejo and Wickens, 1997; Lee and Moray, 1992; Mosier et al., 1998). We address some of these in section 2.3 below.

2.2 Type A Calibration Bias

The type A calibration bias can occur whenever it is explicitly possible to measure the allocation of attention to a variety of information sources, and the allocation policy can be compared against some optimal prescription (Senders, 1964). Of course, there is solid baseline evidence that people can and do allocate attention to tasks, in, proportion to importance weights (Navon and Gopher, 1979; Wickens and Gopher, 1977; Sperling and Doshier, 1986; Tsang and Wickens, 1984; see Gopher, 1993 for a review). A review of studies from sampling theory (Moray, 1986) correspondingly finds that people are able to modulate their visual sampling of different locations in the environments, proportional to both the information content and the importance of the source (e.g., Senders, 1964; Wickens and Seidler, 1997). This information sampling modulation seems to nicely characterize visual fixations on aircraft instruments (Bellenkes, Kramer, and Wickens, 1997). Gronlund et al. (1998) have provided an important analysis of how air traffic controllers allocate attention across aircraft on the display, as a function of their perceived importance, and the features of the aircraft (e.g., altitude, conflict potential) that underlie perceived importance. Finally, a long line of research pioneered by Posner and his colleagues (Posner and Snyder, 1975; see Posner 1978; Pashler, 1998; Egeth and Yantis, 1998 for summaries), has found ample support for this attentional modulation in the "cost benefit analysis" of target cueing. When cues, directing the subject toward a particular target, are only partially reliable (accurate say on 80% of the trials), subjects will respond faster (than an uncued control) when the cue is valid (a benefit) but substantially slower and sometimes erroneously, when the cue is invalid.

Thus, although there is good support for the fact that people can modulate attention to multiple sources according to external prescriptions, the literature has also revealed certain systematic departures from such optimal behavior. For example, in studies of instrument scanning, Senders and his colleagues (Senders, 1964; Carbonnell, Ward, and Senders, 1968), note that people tend to oversample channels with low bandwidths (low information content) and undersample those with high bandwidths. This “flattening” of actual sampling behavior compared to optimal (Figure 3), is reminiscent of the sluggish beta phenomenon discussed above.

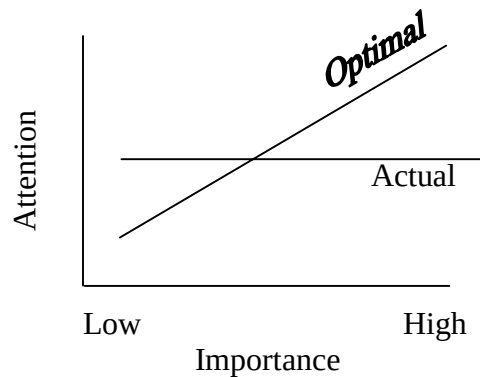


Figure 3

A second example, reflecting the same failure to modulate behavior as responsively as optimal models prescribe, may be found in information integration tasks (e.g., Dawes, Faust, and Meehl, 1989). Here it is found that when people must integrate several channels of varying information value, in order to form a belief, they fail to adequately modulate the amount of weight, placed on channels of different value, effectively placing more weight on those of low value, and less on those of high value. Because this behavior is one that suggests that operators treat all channels **as if** they contained the same value, it is sometimes described by the label of the “as if” heuristic (Cavenaugh et al., 1973; Schum, 1975; see Wickens, 1992 for a summary).

One way of describing the A biases characteristic of information sampling and information integration, is in terms of a “cognitive laziness” that inflicts the integrator of information. When optimal strategies call for quantitative modulation of the impact of varying sources, this imposes a need on cognitive resources that are otherwise conserved, or used for other concurrent tasks. This will be particularly true when the nature of the differential importance or value of the source is not inherently obvious (Wickens and Seidler, 1997). Indeed, when these differences can be made more explicit in a display, the cognitive leveling of the **as if** heuristic appears to be diminished somewhat, and integration becomes more optimal (Barnett and Wickens, 1986).

Type A biases, however, appear to be somewhat more pronounced when they are encouraged by **salient** display features, which signal, inappropriately, that certain cues are more important than others (Wallsten and Barton, 1982). In the context of Figure 2, for example, if C_1 were presented in large print in the middle of a display, while C_2 were in small print at the bottom, C_1 would probably receive more attention, and hence more weighting in the information integration process. Furthermore, if the hypothesis (H_1) supported by C_1 and C_2 was supported

by this displayed information, while the other hypothesis (H_2) was supported by a subtle environmental event, a salience bias would favor the belief in H_1 .

The previous example suggests that displays, which themselves can be thought of as forms of automation (even if they faithfully reflect environmental events), may act as inherent drivers of attention, away from environmental events. This A bias was clearly illustrated in an experiment on helmet mounted display cueing, carried out by Yeh, Wickens, and Seagull (1998). The investigators found that display cueing, to guide operators attention to simulated targets (tanks and mines) did so at the expense of directing attention to a much higher priority (but less expected) nuclear device.

The paradigm used by Yeh et al. (1998) had the properties that all channels of information were reliable. Hence, the issue of trust was not explicitly examined, since subjects knew that the high priority weapons were never cued, but nevertheless could occur in conjunction with a cued tank. However, there are a number of other studies that have combined multi source viewing, (examination of A biases) with varying levels of or reliability of one or more of those sources (examination of T biases) to assess how they work together. We turn to this issue in the following section.

2.3 A and T Biases in Conjunction

The “as if” heuristic, described above, characterizes situations of multiple channels, with varying information value, and hence, situations in which both T and A biases can be manifest. Of even greater interest here, are the circumstances in which multiple channels of different value can also be discriminated on the basis of some salient source characteristic, that leads one to be overprocessed at the expense of the other (i.e., failure to appropriately allocate attention).

There are three generic categories of source characteristics that might affect the allocation of resources (A bias) in a way that does not reflect the true information value of the source (T bias). First, as we discussed above, the perceived source of the information may be from automation, rather than human observations. As we have noted, some findings suggest that humans may overprocess automation guidance (Conejo and Wicknes, 1998; Mosier et al., 1998). Second, information may be associated with a perceived age. Given the fact that, by definition, information changes over time (something that never changes provides zero information), then sources that are sampled more recently tend to provide more accurate information. Hence, a primacy or anchoring bias to overweight initially encountered (and therefore older) sources (Barnett and Wickens, 1988; Hogarth and Einhorn, 1992) will characterize an A bias.

The third source characteristic, which is the focus of the current review is related to physical properties of the display, which we address in the following section.

3. DISPLAY EFFECTS

The preceding discussion has suggested a variety of circumstances that lead to departures from optimality either in terms of A calibration, T calibration, or both. In this section we consider the relevance of data that suggest that properties of a display can influence these calibrations, and hence, if carefully employed by the designer, can reduce the biases. We consider two forms of display variables: **implicit** variables are those inherent properties of a

display that lead to more or less processing. For example, we have already noted that salient display features, such as size, intensity, or centrality in the viewing area, lead to more extensive allocation of attention. Furthermore, any 3D display will, inherently direct attention to one region of space, and thereby away from others (behind the viewpoint); hence, implicitly signaling the greater importance of the center of the forward view. **Explicit** variables are those such as highlighting, or use of numerical indicators to indicate information value. Designers may include such explicit variables in a display, or they may harness and capitalize upon the implicit ones, in achieving the sort of calibration desired, for optimal deployment of attention and preparation.

In our review of the literature, a total of 21 studies were identified that had the following characteristics: (a) all were relatively applied studies, appearing in technical reports or applied journals or proceedings, (b) multiple sources of information were presented to the subject; (c) there was variation across these sources in the information value of the source and/or the importance of the source, (c) subjects made either an integrated judgment, based upon the information from the source, or made separate processing decisions on each source (i.e., multitask performance). The 21 studies were then further assigned to one or more of the following three categories, employed to extract generalizations regarding the display of semi-reliable information.

1. CALIBRATION. The results of studies in this category could address the extent to which participants allocated resources or weights proportional to the reliability or importance of the information source: i.e., were calibrated. In addition, secondary variables (besides source reliability) were identified that might modulate the degree of calibration, (i.e., change the shape of the function represented schematically in Figure 3).

2. ATTENTION GUIDANCE. Studies in this category provided multiple sources of different reliability (or importance) information. The defining aspect of this category was the imposition of attention guidance (i.e., cueing) display formats, designed to direct attention to particular sources, at the expense of others. Such guidance is appropriate when the source is more important or reliable, and will produce a benefit. However, the guidance maybe inappropriate if the cue incorrectly guides attention to a lower priority or less reliable source (i.e., an incorrect instantiation of a less than fully reliable cue). Hence, it may produce a cost relative to either the reliable cue, or to an uncued display. Our interest is in how the display cueing characteristics modulate this cost benefit tradeoff.

3. DISPLAY INDUCING CALIBRATION. Here our interest is in those studies in which display properties have been employed to try to appropriately calibrate the allocation of resources across information sources. Many of these studies are also entered in category 1 (calibration), but under this third category, the interest is in the influence of a second independent variable, explicitly manipulated by the experimenter, on moderating the degree of calibration. When the meaning of this second variable is explained to the subject (e.g., "the circle surrounding the data point indicates a 95% confidence estimate of the true threat location"; or "an amber light represents a 30-70% confidence in the presence of a signal"), then we refer to an explicit manipulation of reliability display. When the meaning of the variable is not explained to the subject, but its effect on calibration is examined, then we refer to this as an implicit manipulation. Our interest is in the general degree of success of either implicit or implicit

manipulations on calibration, AND the extent to which other task variables cause these calibration manipulations to be more or less successful.

In the following table, the studies are listed as they appear in one or more column. Below are listed the general conclusions drawn from each category. Finally the appendix describes the general characteristics of each of the studies.

Table

CATEGORIES OF STUDIES			
	CALIBRATION	ATTENTION GUIDANCE	DISPLAY INDUCING CALIBRATION
1			Andre & Cutler, 1998
2	Banbury et al., 1998		
3	Barnett & Wickens, 1986		Barnett & Wickens, 1986
4	Barnett & Wickens, 1988		Barnett & Wickens, 1988
5		Conejo & Wickens, 1997	
6		Donner et al., 1991	
7		Entin, 1998	
8		Fisher et al., 1989	
9		Fisher & Tan, 1989	
10	Gempler & Wickens, 1998	Gempler & Wickens, 1998	Gempler & Wickens, 1998
11	Kantowitz et al., 1997		
12	Kershtholt et al., 1996		
13			Kirschenbaum & Arruda, 1994
14	Laios, 1978		Laios, 1978
15	Montgomery & Sorkin, 1996		Montgomery & Sorkin, 1996
16		Ockerman & Pritchett	
17	Schipper & Doherty, 1983		Schipper & Doherty, 1983
18		Sorkin et al., 1988	Sorkin et al., 1988
19		Wickens et al., 1999	
20		Yeh, Wickens & Seagull, 1998	
21		Yeh & Wickens, 1999	
GENERAL CONCLUSIONS			
	Unreliability hurts performance on integration and divided attention tasks (except one study found good calibration to unreliability ³)	Cueing is better than no cueing (except when on a well-formatted display it's equivalent ⁶) and when invalid it's worse	Displaying uncertainty helps performance more than no display of uncertainty
	As reliability decreases, performance decreases (except low levels of unreliability may be tolerated and one study found no difference between medium and high uncertainty ¹⁴)	Overtrust valid cueing (especially at expense of uncued, less expected targets ^{20,21})	Displaying <i>predictor</i> uncertainty shows benefits for divided attention ¹⁴ and no effect on integration tasks ¹⁰
	Calibration is better for equal levels of (un)reliability than unequal levels ¹⁵ .	Invalid cueing is worse than no cueing except when less than 50% (un)reliable ⁹	Display format affects calibration (except one study found a difference only with difficult tasks ¹³ and another study found no difference at all ³)

	Calibration hurt by time stress ^{3,17} or task difficulty ¹⁷	High validity is better than low validity cueing (except when target is present but not highlighted they're about the same ⁸)	Explicit display of uncertainty helps integration performance, especially when display is highly unreliable (except for <i>predictor</i> uncertainty as noted above ¹⁰)
	As reliability decreases, trust decreases, but some undertrust for highly reliable system (especially with familiar settings ¹¹)	Highlighting slows performance as more items are highlighted ⁸	Implicit display of reliability helps performance on integration tasks
	Calibration/trust may be restored with moderate/low levels of unreliability ^{10,11}	May overrely on cueing for some display formats: pictures ¹⁶ , immersed ¹⁹ , head-up/conformal and helmet mounted displays ^{20,21}	Explicit reliability display more effective than implicit display for integration performance ¹
		Four levels of cueing (i.e., likelihood alarms) better than binary (more than 4 levels not tested) ¹⁸	Explicit reliability display helps divided attention performance, especially when task is difficult ¹³
		Integration tasks helped by valid cueing, hurt by invalid cueing	

4. REFERENCES

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

Banbury, S., Selcon, S., Endsley, M., Gorton, T., and Tatlock, K. (1998). Being certain about uncertainty: how the representation of system reliability affects pilot decision making. *Proceedings of the Human Factors Society 42nd Annual Meeting*. Santa Monica, CA: Human Factors Society.

Barnett, B.J., and Wickens, C.D. (1986). *Non-optimality in diagnosis: An investigation within a dynamic system*. University of Illinois Cognitive Psychophysiology Laboratory (Tech. Rep. CPL-86-8). Champaign, IL: Department of Psychology.

Barnett, B.J., and Wickens, C.D. (1988). Display proximity in multicue information integration: The benefit of boxes. *Human Factors*, 30, 15-24.

Bellenkes, A.H., Wickens, C.D., and Kramer, A.F. (1997). Visual scanning and pilot expertise: The role of attentional flexibility and mental model development. *Aviation, Space, and Environmental Medicine*, 68(7), 569-579.

Beringer, D.B. (1996). Automation in general aviation: Responses of pilots to autopilot and pitch trim malfunctions. *Proceedings of the 40th Annual Meeting of the Human Factors & Ergonomics Society* (pp.86-90). Santa Monica, CA: Human Factors Society.

Bjork, R.A. (1996). Assessing our own competence: Heuristics and illusions. *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application. Focus on Application Address: Session on Consciousness and Behavior*. Beit Oren, Israel, July.

Brenner, L.A., Koehler, D.J., Liberman, V., and Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior & Human Decision Processes*, 65(3), 212-219.

Carbonnell, J.R., Ward, J.L., and Senders, J.W. (1968). A queuing model of visual sampling: Experimental validation. *IEEE Trans. on Man-Machine Systems*, MMS-9:82-87.

Cavanaugh, Spooner, and Samet (1973).

Conejo, R., and Wickens, C.D. (1997). *The effects of highlighting validity and feature type on air-to-ground target acquisition performance*. University of Illinois Institute of Aviation Technical Report (ARL-97-11/NAWC-ONR-97-1). Savoy, IL: Aviation Res. Lab.

Conejo, R.A., and Wickens, C.D. (1998). The effects of highlighting validity and feature type on air-to-ground target acquisition performance. *Proceedings of the 16th Biennial Applied*

Behavioral Sciences Symposium. USAF Academy, CO: Dept. of Behavioral Sciences & Leadership.

Dawes, R.M., Faust, D., and Meehl, P.E. (1989). Clinical versus statistical judgment. *Science*, 243, 1668-1673.

Donner, K.A., McKay, T., O'Brien, K.M., & Rudisill, M. (1991). Display format and highlighting validity effects on search performance using complex visual displays. *Proceedings of the Human Factors Society 35th Annual Meeting*. Santa Monica, CA: Human Factors Society

Egeth and Yantis, (1998). Visual attention. *Annual Review of Psychology*.

Entin, E.B. (1998). The effects of decision aid availability and accuracy on performance and confidence. *Proceedings Applied Behavioral Sciences Symposium*. U.S. Air Force Academy, Colorado Springs, CO.

Fischhoff, B., and MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting*, 1, 155-172.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception & Performance*, 3, 552-564.

Fisher, D.L., Coury, B.G., Tengs, T.O., and Duffy, S.A. (1989). Minimizing the time to search visual displays: The role of highlighting. *Human Factors*, 31(2), 167-182.

Fisher, D.L., and Tan, K.C. (1989). Visual displays: The highlighting paradox. *Human Factors*, 31, 17-30.

Gempler, K.S., and Wickens, C.D. (1998). *Display of predictor reliability on a cockpit display of traffic information*. ARL Technical Report ARL-98-6/ROCKWELL-98-1.

Gopher, D. (1993). The skill of attentional control: Acquisition and execution of attention strategies. In D. E. Meyer & S. Kornblum (Eds.), *Attention & Performance XIV* (pp. 299-322). Cambridge, MA: MIT Press.

Green, D. M., and Swets, J. A. (1988). *Signal detection theory and psychophysics*. New York: Wiley.

Gronlund et al. (1998)

Henry, R.A., and Snizek, J.A. (1993). Situational factors affecting judgments of future performance. *Organizational Behavior and Human Decision Processes*, 54, 104-132.

Hogarth, R. M., and Einhorn, H. J. (1992). Order effects of belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1-55.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, 45, 423-432.

Kaber, D.B., Onal, E., and Endsley, M.R. (1998). Level of automation effects on telerobot performance and human operator situation awareness and subjective workload (pp. 165-170). In M.W. Scerbo & M. Mouloua (eds.), *Automation technology and human performance: Current research and trends*. Mahwah, NJ: Lawrence Erlbaum.

Kantowitz, B.H., Hanowski, R.J., and Kantowitz, S.C. (1997). Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *Human Factors*, 39(2), 164-176.

Kerstholt J. H., Passenier P. O., Houttuin K., and Schuffel H. (1996). The Effect Of A Priori Probability And Complexity On Decision Making In A Supervisory Control Task. *Human Factors*, 38(1), 65-78.

Laios, L. (1978). Predictive aids for discrete decision tasks with input uncertainty. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-8(1), 19-29.

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

Liu, Y., Fuld, R., and Wickens, C.D. (1993). Monitoring behavior in manual and automated scheduling systems. *Int. J. Man-Machine Studies*, 39, 1015-1029.

Montgomery, and Sorkin (1996). Observer sensitivity to element reliability in a multielement visual display. *Human Factors*, 38(3), 484-494.

Moray, N. (1986). Monitoring behavior and supervisory control. In: *Handbook of perception and performance, Vol II*, edited by K.R. Boff, L. Kaufman, and J.P. Thomas, New York: Wiley & Sons, pp. 40-1/40-51.

Mosier, K.L., Skitka, L.J., Heers, S., and Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology* 8(1):47-63.

Muir, B.M. (1988). Trust between humans and machines, and the design of decision aids. In: *Cognitive engineering in complex dynamic worlds*, edited by E. Hollnagel, G. Mancini, and D.D. Woods (eds.), London: Academic Press (pp. 71-83).

Navon, D., and Gopher, D. (1979). On the economy of the human processing system. *Psychological Review*, 86, 214-25.

Ockerman, J.J. & Pritchett, A.R. (unpublished)

Olmos, O., Wickens, C.D., and Chudy, A. (in press). Tactical displays for combat awareness. *International Journal of Aviation Psychology*.

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

Parasuraman, R., Molloy, R., and Singh, I.L. (1993). Performance consequences of automation-induced complacency. *Intl. J. of Avia. Psych.* 3(1):1-23.

Pashler, H.E. (1998) . *The psychology of attention*. Cambridge, MA: The MIT Press.

Posner, M.I. (1978). *Chronometric explorations of the mind*, Hillsdale, NJ: Erlbaum Associates.

Posner, M.I., and Snyder, C.R.R. (1992). Attention and cognitive control. In: *Information processing and cognition: The Loyola Symposium*, edited by R. Solso, Lawrence Erlbaum.

Sarter, N.B., and Woods, D.D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors* 37(1):5-19.

Schipper, L.M., and Doherty, M. (1983). *Decision making and information processing under various uncertainty conditions*. Air Force Human Resources Laboratory Report AFHRL-TR-83-19.

Schum, D. (1975). The weighing of testimony of judicial proceedings from sources having reduced credibility. *Human Factors*, 17, 172-203.

Senders, J.W. (1964). The human operator as a monitor and controller of multidegree of freedom systems. *IEEE Trans. On Human Factors in Electronics* HFE-5:2-6.

Sorkin, R.D. (1988). Why are people turning off our alarms? *Journal of the Acoustical Society of America*, 84:1107-1108.

Sorkin, R.D., Kantowitz, B.H., Kantowitz, S.C. (1988). Likelihood alarm displays. *Human Factors*, 30(4), 445-459.

Sperling, G., and Doshier, B.A. (1986). Strategy and optimization in human information processing. In: *Handbook of perception and performance, Vol. I*, edited by K. Boff, L. Kaufman, and J. Thomas, New York: Wiley, pp. 2-1/2-65.

Svenson, O. (1981). Are we less risky and more skillful than our fellow drivers? *Acta Psychologica*, 47, 143-148.

Tsang, P., and Wickens, C.D. (1984). The effects of task structures on time-sharing efficiency and resource allocation optimality. *Proceedings of the 20th Annual Conference on Manual Control and Mental Workload*. Moffett Field, CA: NASA Ames Research Center.

Wallsten, T.S., and Barton, C. (1982). Processing probabilistic multidimensional information for decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 8, 361-384.

Wells, G.L., Lindsay, R.C., and Ferguson, T. I. (1979). Accuracy, confidence, and juror perceptions in eyewitness testimony. *Journal of Applied Psychology*, 64, 440-448.

Wickens, C.D. (1992). *Engineering psychology and human performance* (2nd ed.). New York: Harper Collins.

Wickens, C.D., and Carswell, C.M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors*, 37(3), 473-494.

Wickens, C.D., and Gopher, D. (1977). Control theory measures of tracking as indices of attention allocation strategies. *Human Factors*, 19, 249-366.

Wickens, C.D., Gordon, S., and Liu, Y. (1998). *An introduction to human factors engineering*. New York: Addison Wesley Longman, Inc.

Wickens, C.D., Mavor, A.S., Parasuraman, R., and McGee, J.P. (Eds.). (1998). *The future of air traffic control: Human operators and automation*. Washington, DC: National Academy Press.

Wickens, C.D., Thomas, L., Merlo, J., and Hah, S. (1999, to be published in *Proceedings ARL Federated Laboratory Third Annual Symposium*). Immersion and battlefield visualization: Does it influence cognitive tunneling?

Wickens, C.D., and Seidler, K.S. (1997). Information access in a dual-task context: Testing a model of optimal strategy selection. *Journal of Experimental Psychology: Applied* 3(3), 196-215.

Yeh, M., and Wickens, C.D. (1999, to be published in *Proceedings ARL Federated Laboratory Third Annual Symposium*). Visual search and target cueing with augmented reality: A comparison of head mounted with hand-held displays.

Yeh, M., Wickens, C.D., and Seagull F.J. (1998). Effects of frame of reference and viewing condition on attentional issues with helmet-mounted displays. *Proceedings of the 2nd Annual FedLab Symposium: Advanced displays & interactive displays* (pp. 107-114). College Park, MD: U.S. Army Research Federated Laboratory Consortium.

Appendix

Study: Andre, A.D. & Cutler, H.A. (1998). Displaying uncertainty in advanced navigation systems. Proceedings of the Human Factors Society 42nd Annual Meeting. Santa Monica, CA: Human Factors Society.	Domain: Generic
Reliability: ownship 100%, meteor obstacle location was accurate, moderately uncertain (i.e., location radius off by 15 pixels) or highly uncertain (i.e., location radius off by 50 pixels). Reliability Display: -Text: yellow numeral (i.e., 0, 15 or 50) -Graphical-implicit: color coded uncertainty (i.e., green = 0; yellow = 15, red = 50) -Graphical-explicit: variable radius circles (i.e., 0, 15 or 50 pixels)	Bias: Type A and T Cueing: None.
Display: a PC and a 17 inch color monitor displayed ownship moving along straight line to goal, with meteor obstacle along path Task Demands: Integration. Travel to goal along shortest route, avoiding meteor. Other Variables/Manipulations: none.	
Results: All three uncertainty symbologies produced better performance vs no information displayed at all, with the graphical-explicit representation supported the best calibration by the subjects for the highest level of position uncertainty.	

Study: Banbury, S., Selcon, S., Endsley, M., Gorton, T., & Tatlock, K. (1998) Being certain about uncertainty: how the representation of system reliability affects pilot decision making. Proceedings of the Human Factors Society 42nd Annual Meeting. Santa Monica, CA: Human Factors Society.	Domain: Aviation
Reliability: 3, 6, 9, 21 & 39% confidence (thus, 97, 94, 91, 79 & 61% uncertainty) Reliability Display: Reliability framing (system reliability vs system uncertainty)	Bias: Type A and T Cueing: None.
Display: 6X6" plan view of tactical situation (i.e., an unknown aircraft heading toward ownship) on Macintosh PC with 12" monitor Task Demands: Integration. Determine shoot/no shoot given tactical situation. Other Variables/Manipulations: target identifier (none, friendly or enemy)	
Results: In general, framing had no effect on calibration (i.e., RT and decision to shoot). When target identifiers were provided, pilots were faster for "confidence" displays, compared to the same information framed as uncertainty. Reliability significantly affected performance (low uncertainty/high confidence resulted in faster and more shots taken than high uncertainty/low confidence). Also, displaying the secondary target as a potential friendly (although equal potential existed for secondary being enemy) significantly affected decision to shoot. Levels of uncertainty above 9% carried an unacceptable level of risk of fratricide (fast RT and fewer shots taken for friendlies). RTs were slowest around 6-9% uncertainty, suggesting subjects had difficulty resolving this level of ambiguity.	

Study: Barnett, B.J. & Wickens, C.D. (1986). Non-optimality in diagnosis: An investigation within a dynamic system. University of Illinois Cognitive Psychophysiology Laboratory (Tech. Rep. CPL-86-8). Champaign, IL: Department of Psychology. Reliability: 8 levels (1-8) of information worth (i.e., cue reliability * diagnosticity) Reliability Display: Cue location: center informative (i.e., higher information worth located centrally) vs left informative vs right informative Display: CRT display of aircraft system states Task Demands: Integration. Determine fly/no fly given system states. Other Variables/Manipulations: secondary task load Results: Overall, performance was close to optimal, with modest departures. Moderate time stress led to less than optimal calibration, with some favoritism for cues at the top-left. A trend was found for highly informative cues being underweighted, while less-informative sources were over weighted, another indication of less than optimal calibration. Number of cues did not affect calibration.	Domain: Aviation Bias: Type A and T Cueing: 5 or 8 information source cues
Study: Barnett, B.J. & Wickens, C.D. (1988). Display proximity in multicue information integration: The benefits of boxes. <u>Human Factors</u>, 30(1), 15-24. Reliability: unique value of information worth (i.e., reliability & diagnosticity) for each cue ranging from 2 to 25 Reliability Display: bar graph v rectangle vs integral rectangle displays of information worth Display: CRT display of aircraft system states Task Demands: Integration. Determine fly/no fly given system states. Other Variables/Manipulations: none. Results: Calibration (i.e., correlation with optimal weighting of cue information worth) was best for integral rectangles condition (>.90), followed by rectangles (>.88) and then bar graphs (>.85). Calibration generally improved as display proximity improved (i.e., simultaneous display outperformed sequential display).	Domain: Aviation Bias: Type A and T Cueing: 4 (of 8) information sources displayed sequentially in space and time v sequentially in time v simultaneous display w/time constraint v simultaneous display w/o time constraint

Study: Conejo, R. & Wickens, C.D. (1997). The effects of highlighting validity and feature type on air-to-ground target acquisition performance. (Technical Report ARL-97-11/NAWC-ONR-97-1). Savoy, IL: University of Illinois, Institute of Aviation, Aviation Research Laboratory. Reliability: overall highlighting 70% valid, target highlighting 40% (valid highlighting vs invalid highlighting vs invalid highlighting/target absent) Reliability Display: None. Display: Evans & Sutherland display of aerial bombing run Task Demands: Integration. Determine shoot/no shoot given world, map and cueing indications. Other Variables/Manipulations: lead-in feature type (natural vs cultural); target feature type (natural vs cultural) Results: Valid highlighting led to increased confidence but no corresponding increase in accuracy. Invalid highlighting produced accuracy costs, leading pilots down a garden path.	Domain: Aviation Bias: Type A and T Cueing: no highlighting, target highlighted red, target highlighted/lead-in blinking
Study: Donner, K.A., McKay, T., O'Brien, K.M., & Rudisill, M. (1991). Display format and highlighting validity effects on search performance using complex visual displays. Proceedings of the Human Factors Society 35th Annual Meeting. Santa Monica, CA: Human Factors Society. Reliability: valid and invalid Reliability Display: None. Display: IBM PC/XT with 13" color monitor; 2 complex text-based display types (Orbit Maneuver Execute, Relative Navigation); 2 display formats (current, reformatted) Task Demands: Integration. For each trial, subjects responded to a system status question after viewing display. Other Variables/Manipulations: none. Results: In general, performance times increased from valid to invalid highlighting, with valid highlighting significantly faster and invalid highlighting producing significant time costs. However, highlighting benefits & costs varied with display types. Search times on a poorly-formatted display benefit from highlighting (without cost), while the reformatted display search times were neither helped nor hurt by highlighting.	Domain: Space Bias: Type A and T Cueing: highlighting (none, brightness, color (blue), flashing or reverse video) present on 80% of displays

Study: Entin, E.B. (1998). The Effects of Decision Aid Availability and Accuracy on Performance and Confidence. Proceedings Applied Behavioral Sciences Symposium. U.S. Air Force Academy, Colorado Springs, CO.	Domain: Battlefield management
Reliability: high (.9 hits/.1 false alarms) vs low (.9 hits/.4 false alarms) Reliability Display: none, but objects were visually obscured to represent noise in system	Bias: Type A and T Cueing: An Aided Target Recognition (ATR) system marked targets with red squares
Display: Seven to 10 objects displayed, comprised of a 50:50 mix of targets and non-targets. Task Demands: Integration. Identify targets vs non-targets. Other Variables/Manipulations: confidence in target selection.	
Results: Aided performance was generally more accurate than unaided performance, however, subjects underrelied on the highly reliable ATR and thus, their performance was less than optimal. Interestingly, they tended to agree with the highly reliable ATR target selection (90% agreement) than its selection of non targets (85%); and they were significantly more confident in their own target (vs non-target) decisions.	

Study: Fisher, D.L., Coury, B.G., Tengs, T.O., & Duffy, S.A. (1989). Minimizing the time to search visual displays: The role of highlighting. <u>Human Factors</u>, 31(2), 167-182.	Domain: Generic
Reliability: trial reliability ranged from 0 to 100%, given target present on 75% of trials. Overall probability of target highlighting in a block of 16 trials considered low (25%) vs high (75%) validity Reliability Display: None.	Bias: Type A and T Cueing: 0, 1, 3, 6, or 12 out of 36 words highlighted (none, blocked or random yellow highlighting)
Display: 36 words in a 6 X 6 matrix, using a colored monitor and a DEC 350 microcomputer Task Demands: Integration. Identify target, given the probability that a target would be present, the level of highlighting validity, and the number of highlighted options before each block. Other Variables/Manipulations: none.	
Results: Overall RTs: 1. (highlighting) < (no highlighting) 2. as # of highlighted words increases, search time increases (except slope ~ 0 when target was absent & low validity highlighting) 3. (target present & highlighted) < (target present & not highlighted) < (highlighting + target absent) 4. (high validity highlighting) < (low validity highlighting) (target present & highlighted) or (target absent + highlighting) 5. (high validity highlighting) ~ (low validity highlighting) (target present & not highlighted)	

Study: Fisher, D.L. and Tan, K.C. (1989) Visual displays: The highlighting paradox. <u>Human Factors</u>, 31, 17-30. Reliability: When highlighting was displayed, always 50% reliable. Reliability Display: None. Display: Five white digits in horizontal array appeared on a DEC color monitor Task Demands: Integration. Identify target. Other Variables/Manipulations: none. Results: Target highlighting using color results in faster RT than highlighting using blinking or reverse video. Also, 50% reliable color highlighting is no faster than no highlighting. Subjects do not always attend to first to highlighting when it is 50% reliable. Highlighting may be worse than no highlighting at all when less than 50% reliable.	Domain: Generic Bias: Type A and T Cueing: highlighting a single digit (i.e., none, color (yellow), blinking or reverse video)
Study: Gempler, K.S., & Wickens, C.D. (1998). Display of predictor reliability on a cockpit display of traffic information. ARL Technical Report ARL-98-6/ROCKWELL-98-1. Reliability: 100% self; 83% intruder Reliability Display: CDTI wedge predictor (95% confidence interval of future position) v. single line predictor Display: Silicon Graphics workstation and 20" monitor display of Cockpit Display of Traffic Information for ownship and potential intruder Task Demands: Integration. Stay on glideslope, avoiding traffic conflicts. Other Variables/Manipulations: vertical & longitudinal traffic geometry Results: Showed costs for invalid predictors, which were amplified on descending trials. The wedge predictor did not affect pilots' trust calibration or perceived reliability, perhaps due to its additional clutter (v single line predictor) or possibly due to how pilots strategically used the display. Automation failures (i.e., invalid trials) had short-lived effects on time in conflict (i.e., performance costs primarily on trials where predictor fails), perhaps indicating good initial calibration with little recalibration required following failure.	Domain: Aviation Bias: Type A and T Cueing: Length of predictor line decreased as time to predicted conflict decreased. Actual conflict highlighted intruder in yellow
Study: Kantowitz, B.H., Hanowski, R.J., & Kantowitz, S.C. (1997). Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. <u>Human Factors</u>, 39(2), 164-176. Reliability: The navigation information display was 100%, 71% or 43 % accurate. Reliability Display: None. Display: A Battelle Route Guidance Simulator, which consists of two linked Intel 486 computers and two video displays ("real-time" traffic display vs route guidance). Task Type: Divided attention. Drive to location using display information. Other Variables/Manipulations: Level of trust; familiarity of the driving area Results: As information reliability decreased, performance (i.e., optimal route selection), subjective opinion and operator trust decreased. Operator trust was restored with subsequent accurate information; however, restoration was less likely with lower reliability levels. Familiarity with environs resulted in less effective use of information (i.e., poorer calibration), inaccurate information was more harmful for familiar settings because drivers relied on automation less when they thought they knew the area better.	Domain: Driving Bias: Type A and T Cueing: None

Study: Kerstholt J. H., Passenier P. O., Houttuin K., & Schuffel H. (1996) The Effect Of A Priori Probability And Complexity On Decision Making In A Supervisory Control Task. <u>Human Factors</u>, 38(1), 65-78.	Domain: Ship control
Reliability: cues were 100%; fault rates had low (12 failures) or high probability (21 failures) during 4 hours of experiment Reliability Display: None.	Bias: Type A and T Cueing: Faults in any of 4 independent subsystems were auditorily cued after 65 seconds
Display: Computer screen displaying subsystem status, one at a time upon request Task Demands: Divided attention. Monitor system, detect faults. Other Variables/Manipulations: complexity (number of system disturbances occurring at a time)	
Results: Showed a failure to calibrate response rates to the fault rates of the different sub-systems. Subjects may not have had enough time to realize that one subsystem was less reliable than the others. Inherent differences in the sub-systems may have further washed out any differences caused by the varying fault probability.	

Study: Kirschenbaum, S.S., & Arruda, J.E. (1994). Effects of graphic and verbal probability information on command decision making. <u>Human Factors</u>, 36(3), 406-418.	Domain: Submarine operations
Reliability: Ownship location 100% reliable; target location estimated using correct vs mismodeled algorithms and low vs high oceanic noise conditions Reliability Display: Ellipse (95% confidence in location) vs verbal indicator (poor, fair, good)	Bias: Type A and T Cueing: None
Display: A Macintosh IIFX displayed spatial information on ownship and target, in addition to abundant reliable system information regarding ship course, range, speed, range rate, and bearing Task Demands: Divided attention. Determine enemy status information (e.g., range, location) using displays. Other Variables/Manipulations: Subjective confidence in responses	
Results: Target range estimates were most accurate with the ellipse format (in other words, spatial format best supported spatial problem solving), although there was no significant difference in subjective confidence. Ellipse advantage held only under high noise/correct modeling (a difficult task), it did not help anytime noise was low (possibly due to simplistic nature of problem for the highly trained subjects). Trend analysis further suggests subjects more easily misled by distorted information (i.e., mismodeling/high noise) than those relying on verbal reliability. Highly robust effects considering additional cues!	

Study: Laios, L. (1978). Predictive aids for discrete decision tasks with input uncertainty. <u>IEEE Transactions on Systems, Man and Cybernetics</u>, SMC-8(1), 19-29.	Domain: Industrial control systems
Reliability: low (0), medium, high uncertainty arrival times followed rectangular distributions ($x T, s^2$), T=actual arrival time; $s=a(T-10n)$, $a=.4, 1.0$ for medium, high uncertainty, respectively, $n=10$ time units. With every update (i.e., 10 time units), the arrival estimates became more accurate Reliability Display: predictive display showing intervals for arrival times, with longer intervals showing more uncertainty (much like standard error bars on graphs)	Bias: Type A Cueing: None
Display: bar display on computer showing the expected arrival time of ingots out of 4 soaking pits Task Demands: Divided Attention. Subjects had to maintain a constant flow of ingots to each soaking pit (with different soaking times). Other Variables/Manipulations: none.	
Results: Performance under uncertainty conditions was significantly poorer than under no uncertainty, however, no significant difference between high and medium uncertainty. When information was accurate (i.e., no uncertainty), predictive display significantly helped performance. Under uncertainty, the predictive display benefit only helped when the uncertainty was displayed.	

Study: Montgomery and Sorkin (1996). Observer sensitivity to element reliability in a multielement visual display. Human Factors, 38(3), 484-494. Reliability: low, high, equal Reliability Display: luminance indicated reliability (white = high reliability; grey = low/equal reliability) Display: Nine vertical gauges displayed on 27cm color monitor Task Demands: Integration. Determine whether information in array of sources was due to noise or signal. Other Variables/Manipulations: stimulus duration Results: Without luminance cues, observers were better calibrated when all gauges had equal reliability than when reliability differed across gauges (high reliable gauges were rated only slightly higher than low-reliability ones). Calibration significantly improved when stimulus duration increased and luminance indicated reliability.	Domain: Generic Bias: Type A Cueing: 9 information sources (gauges)
Study: Ockerman, J.J. & Pritchett, A.R. (unpublished) Reliability: Ten checklist items (i.e., faults) were either not displayed or incorrectly displayed on computer Reliability Display: None. Display: A wearable computer with a small monitor was mounted on the head, displaying preflight checklist items. Checklist menus were voice-driven. Task Demands: Divided attention. Perform preflight on aircraft. Other Variables/Manipulations: Pilot acceptability of wearable computer. Results: Trend analysis indicated pilots heavily relied on the computer checklist, which resulted in benefits and costs. All pilots wearing the computer picked up faults listed on the computer that several control pilots forgot. The computer also led to overreliance, as pilots missed faults not specifically written out, and the text + picture format led to the least number of faults detected.	Domain: Aviation Bias: Type A and T Cueing: preflight checklist items cued via text alone vs text and picture vs control (memory)
Study: Schipper, L.M. & Doherty, M. (1983). Decision making and information processing under various uncertainty conditions. Air Force Human Resources Laboratory Report AFHRL-TR-83-19. Reliability: high (7), medium (4) or low (2) probability of occurrence Reliability Display: none Display: CRT display of event reliability; histogram bars (HB), list format (LF; 7, 4, 2) or geometric numeric (GN; equal distances between equally different probabilities) Task Demands: Integration. Determine event likelihood given information. Other Variables/Manipulations: display duration (3, 6, 9 seconds); evaluating decision-making strategies Results: The more information sources available (i.e., cues) the greater the tendency to average their probability of occurrence (i.e., treat them as equally weighted). Each reliability format produced about the same level of error (i.e., inferred likelihood of occurrence), with the highest error associated with HB, lowest error with GN. Strong individual preferences existed for different formats. Longer display durations and fewer cues supported more accurate performance. Other experiments suggested subjects were better calibrated to symmetrical information displays, subjects perceived scattered information arrays as more reliable than dense arrays, and outliers were discounted given a sufficiently large distance between the outlier and information cluster.	Domain: Generic Bias: Type A and T Cueing: 3, 5 or 7 cues

Study: Sorkin, R.D., Kantowitz, B.H., Kantowitz, S.C. (1988). Likelihood alarm displays. <u>Human Factors</u>, 30(4), 445-459. Reliability: 100% Reliability Display: None. Display: IBM PC displayed an array of 4 three digit numbers at the bottom of a color monitor Task Demands: Integration. Respond to alarm while performing primary tracking task. Other Variables/Manipulations: primary tracking task; secondary monitoring task Results: Use of alarms yielded faster RT & higher accuracy (vs no alarms). Likelihood alarms yielded more accurate responses than binary alarms (though similar RT) when tracking was difficult. Alarm format (i.e., visual or auditory) did not appear to affect performance.	Domain: Generic Bias: Type A Cueing: Likelihood (e.g., no alarm, possible, likely and urgent) vs binary alarms (all or none); Visual (i.e., color) vs auditory cueing
Study: Wickens, C.D., Thomas, L., Merlo, J. & Hah, S. (1999, to be published in Proceedings ARL Federated Laboratory Third Annual Symposium). Immersion and battlefield visualization: Does it influence cognitive tunneling? Reliability: 100% Reliability Display: None. Display: A 3D exocentric (tethered) display vs a 3D immersed display of battlefield Task Demands: Divided attention. Look for enemy & friendlies in battlefield. Other Variables/Manipulations: Subjective confidence in responses. Results: The immersed display induced “cognitive tunneling” in which subjects were overly influenced by information in the initially presented forward view and failed to adequately pan behind their position. The confidence data revealed that immersed display subjects did not lower their confidence in the accuracy of their answers in a way commensurate with the loss of accuracy.	Domain: Battlefield Management Bias: Type A Cueing: None.
Study: Yeh, M. & Wickens, C.D. (1999, to be published in Proceedings ARL Federated Laboratory Third Annual Symposium). Visual search and target cueing with augmented reality: A comparison of head mounted with hand-held displays. Reliability: 100% Reliability Display: None. Display: Subjects were placed in a Cave Automatic Virtual Environment (CAVE) and wore a helmet mounted display (HMD) or used a hand-held display (small 3 ½ inch portable TV). Task Demands: Divided attention. Report friendly & enemy in environment. Other Variables/Manipulations: Cueing and information symbology is presented in either a heads up fashion using an HMD or in a head down manner using a hand held device (HHD). Secondary task performance was also measured. Results: Cueing presented on HMD guided subjects attention to simulated targets (tanks and mines) but did so at the expense of directing attention away from a much higher priority (but less expected) nuclear device. The HHD seemed to mediate this effect of attention tunneling in that the higher priority target was detected more times on average when the HHD was used.	Domain: Battlefield Management Bias: Type A Cueing: An arrow cue was present or not present. Highest priority target not cued.

Study: Yeh, M., Wickens, C.D., & Seagull, F.J. (1998). Effects of frame of reference and viewing condition on attentional issues with helmet-mounted displays (Technical Report ARL-98-1/ARMY-FED-LAB-98-1). Savoy, IL: University of Illinois, Institute of Aviation, Aviation Research Laboratory

Reliability: 100%
Reliability Display: None

Domain: Battlefield Management

Bias: Type A
Cueing: The cueing is manipulated by either being present or not present. Highest priority target not cued.

Display: Subjects were placed in a Cave Automatic Virtual Environment (CAVE) and wore shutter glasses that simulated the wearing of a helmet-mounted display (HMD).

Task Demands: Divided attention. Report friendly & enemy in environment.

Other Variables/Manipulations: Cueing and information symbology is presented world or screen referenced, additionally the symbology is presented in either a monocular or binocular fashion. Secondary task performance was also measured.

Results:

Found that display cueing guided operators attention to simulated targets (tanks and mines) at the expense of directing attention away from a much higher priority (but less expected) nuclear device.