BAE SYSTEMS

Advanced Technology Centre

2004 Command and Control Research and Technology Symposium

Quantitative Analysis of Situational Awareness (QUASA): Applying Signal Detection Theory to True/False Probes and Self-Ratings

Barry McGuinness
(Principal Scientist)

BAE SYSTEMS

Human Factors Department
Advanced Technology Centre
Sowerby Building
BAE SYSTEMS
FPC 267, PO Box 5, Filton
Bristol BS34 7QW
U.K.

Tel 44 117 302 8197
Fax 44 117 302 8007

Email barry.mcguinness@baesystems.com

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2004		**00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Quantitative Analysis of Situational Awareness (QUASA): Applying Signal Detection Theory to True/False Probes and Self-Ratings	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) **BAE Systems,FPC 267, PO Box 5, Filton,Bristol BS34 7QW,UK, ,**	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
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**85**	

Quantitative Analysis of Situational Awareness (QUASA): Applying Signal Detection Theory to True/False Probes and Self-Ratings

Barry McGuinness

BAE SYSTEMS

Human Factors Department
Advanced Technology Centre
Sowerby Building
BAE SYSTEMS
FPC 267, PO Box 5, Filton
Bristol BS34 7QW
U.K.

Abstract

This paper describes a technique for assessing the situational awareness of individuals such as warfighters participating in C2 experiments and exercises. Known as QUASA (quantitative analysis of situational awareness), the technique combines both objective queries (true/false probes) and subjective self-ratings of confidence for each probe response. The data so obtained are then analyzed and interpreted using the mathematical framework of Signal Detection Theory (SDT). The rationale behind the technique is described, followed by an example of its implementation and the results obtained. Further refinements of the technique based on recent research in experimental psychology are also considered.

INTRODUCTION

The key role of a warfighter is to influence the outcome of complex real-world situations by making appropriate decisions and taking effective action. This includes not only responding to problematic situations as they arise but also proactively anticipating and preventing them occurring in the first place. Situational awareness (SA) refers to all the situation-specific information and inferences represented in a person's mind which he or she uses to make such decisions. Put simply, SA is "knowing what is going on so you can figure out what to do" (Adam, 1993). It is also "what you need to know not to be surprised" (Jeannot *et al.*, 2003). It consists of whatever answers one currently has in mind to such questions as: *What is happening? Why is it happening? What will happen next? What does it mean in terms of my objectives? What can I do about it?*

Aspects of SA

The situations of most concern to the warfighter's SA are those which can vary rapidly and which he or she is responsible for managing through decisive action—the local battlefield situation, for instance. To be of value, the awareness of a situation of concern needs to be suitably comprehensive (covering all relevant factors), up-to-date, and expertly interpreted, capturing the real meaning and full implications of the situation. Beyond the main situations of interest, however, the individual should also have some awareness of contextual factors—related situations, historical background, cultural factors, and so on—which may have a bearing on situational understanding.

Another key factor to be taken into account is oneself, i.e. self-awareness of one's own intent and capabilities. The same situation can be interpreted and evaluated by an individual in different ways depending upon what the individual is trying to do and how equipped he or she is to do it. Self-awareness can also include metacognition of one's own SA, which may be assessed as being "complete" versus "incomplete", or "up to speed" versus "lagging behind", or "certain" versus "uncertain". For example, a commander's realisation that he has an incomplete and uncertain understanding of the battlefield situation could be a significant factor in his subsequent decision-making.

Thus SA is a dynamic, multifaceted phenomenon that not only includes perception, understanding and evaluation of the key situations of concern and their relevant contexts but also addresses the perception, understanding and evaluation of oneself. In fact, the role of SA in decision-making appears to be a function of two things: (a) **actual SA**, the level of SA acquired relative to the SA required to perform effectively; (b) **perceived SA**, the individual's own metacognitive perception of (a).

This paper presents a method for assessing SA that integrates techniques to measure both actual SA and perceived SA within the mathematical framework of Signal Detection Theory.

ASSESSING SITUATIONAL AWARENESS

Situational awareness is now a key concept in human factors research and practice, where the effects of ever-increasing technological and situational complexity on the human agent are a central concern. As a consequence, valid and meaningful measures of SA are required to help us assess the impact of new system designs, training methods, and so on. But how can we evaluate someone's SA? On what scientific basis can we say that one person has better SA than another, or that a new information system has led to decision-makers having better SA than before? The assessment of SA is far from straightforward. This is partly due to the multifaceted nature of SA itself, and partly due to the fundamental difficulty of observing and evaluating what is happening in another person's mind.

Over the last decade or so numerous techniques for assessing situational awareness have emerged. These essentially boil down to three approaches which can be distinguished according to the kinds of evidence they seek:

(1) **Inferential** techniques seek implicit evidence of SA from observable correlates. The individual's performance, behaviour and even physiology can be monitored as indirect evidence for the presence or absence of appropriate SA. For example, in their SALIANT technique Muñiz et al. (1998) infer a team's SA from observed

behaviours. The slightly different SABARS (Situational Awareness Behaviorally Anchored Rating Scale) technique requires expert observers to rate individuals on a number of observable behaviours related to SA processes (Strater et al., 2001). While this is an unobtrusive approach, the limitation is that a performance error or omission of an SA-related behaviour does not necessarily indicate a lack of SA, nor does good performance necessarily imply good SA (Baxter & Bass, 1998). Some sort of measurement intervention is often necessary to get a better understanding of individuals' SA.

(2) **Self-rating** techniques seek subjective evidence of SA by eliciting the individuals' own self-perceptions. This approach is embodied in such tools as the Participant Situation Awareness Questionnaire (PSAQ) used by Matthews et al. (2000), the Situational Awareness Rating Tool (SART) of Taylor (1990), and the Crew Awareness Rating Scale (CARS) described by McGuinness & Foy (2000). Such tools differ in the number of dimensions rated and the types of scales used. Self-ratings of SA can be obtained immediately after an exercise or experimental run or they can also be obtained one or more times mid-run with relatively minor disruption.

(3) **Probe** techniques or **query** techniques seek direct evidence of the content of individuals' SA. This approach involves eliciting a set of information from the individual on his or her perception and understanding of the situation, and then comparing this against the real thing, the ground truth. Probes are of two basic types. With *supply* probes, the individual is asked to supply specific information about the situation. For example, a pilot may be asked "What is your current rate of ascent?" With the more labour-intensive *selection* probes, the correct information is presented to the individual along with one or more incorrect alternatives; the individual must select the correct one. This technique is embodied as a multiple-choice format, for example, in the well-known Situation Awareness Global Assessment Technique (SAGAT) of Endsley (1995, 2000). This approach is the most intrusive and disruptive of the three, but generates the most direct evidence of the state of someone's SA.

QUANTITATIVE ANALYSIS OF SITUATIONAL AWARENESS (QUASA)

As suggested above, SA seems to have a twofold effect on decision-making (in the form of actual SA and perceived SA), yet it is rare for more than one approach to be used in a study. Probe techniques specifically address actual SA and self-ratings techniques specifically address perceived SA, while inferential techniques look for signs of SA without really making the distinction. Yet to measure one aspect without the other runs the risk of making only a partial assessment of SA insofar as it impacts on decision-making. While it has long been realised that SA is a multifaceted concept, this is not often reflected in the development of measurement techniques.

There is, however, an established line of research in psychology in which objective and subjective metrics are combined to analyse the degree of 'calibration' in people's perception, memory and knowledge (see below). The extension of this to situational awareness measurement is embodied in the QUASA technique which combines probe assessment with self-ratings. What follows is a description of the technique and its rationale.

SA requirements analysis

Procedurally, the technique actually begins long before any data gathering session (in the same manner as other probe methods such as SAGAT) with an analysis of subjects' SA requirements. This is a form of Cognitive Task Analysis which captures, in essence, both the explicit cues and the implicit inferences that are relevant to the decision-maker's thinking. These elements of information and understanding, which are needed by the individual to maintain at least adequate performance, serve also as the basis from which SA probes can be generated (for example, see Endsley, 1995). It is absolutely vital that the queries used are valid and relevant probes of the individual's SA, otherwise we are merely measuring their ability to respond to arbitrary and irrelevant questions.

True/false probes

In theory, the extent of a person's knowledge of a situation is indicated by his or her success in being able to judge the truth or falsity of propositions related to it (Ebel and Frisbie, 1991). With this in mind, the SA query method in the QUASA technique consists of true/false probes, whereby the subject is periodically presented with a set of descriptions of the situation and asked to indicate in each case whether the statement is true or false. For example:

Probe statement	Assessment
A column of enemy tanks is now leaving the city.	[ ] ***True*** [✓] ***False***

This differs from the multiple-choice format used in SAGAT in that, rather than presenting a question plus several possible answers, one of which is correct, the subject is presented a single statement which may or may not be true followed by two options: 'True' or 'False'. As in SAGAT, however, the true/false probe items must of course be carefully derived for their relevance to the subject's SA requirements within the task.

Simultaneous self-ratings

The self-rating method that is also used in QUASA consists of simultaneous self-ratings of confidence *for each and every true/false probe response*. For example:

Probe statement	Assessment	Confidence
A column of enemy tanks is now leaving the city.	[ ] ***True*** [✓] ***False***	[ ] ***Very High*** [ ] ***High*** [ ] ***Moderate*** [✓] ***Low*** [ ] ***Very Low***

This combination of true/false discriminations and ratings of confidence for those discriminations is an example of the *calibration* technique described below. It is also amenable to analysis by the methods of Signal Detection Theory, which we will also briefly summarise for the uninitiated.

CALIBRATION OF SA

The principle of calibration concerns, in essence, the extent to which people are able to judge the correctness of their own observations or decisions. In other words, it assesses the degree of correspondence between self-perceptions of accuracy and actual accuracy as a proportion of correct responses (Koriat & Goldsmith, 1996). A well-calibrated judge is one who is highly confident about those responses that are in fact correct, and unconfident about those responses that are in fact incorrect. In a poorly calibrated judge, there is no systematic relationship between real and perceived accuracy.

To assess calibration, an individual is typically presented a test item and asked to provide the correct answer; immediately afterward, they are prompted to give an indication of their confidence in the answer just given. (Alternatively, calibration may be studied predictively: the individual gives a confidence rating prior to answering a question.) A confidence rating may be elicited by way of a binary rating (e.g., *high confidence* versus *low confidence*), a multi-categorical ordinal rating (e.g., *very high, high, moderate, low, very low*), or a continuous scale (e.g., 0%–100%). After multiple test items and their respective confidence ratings, calibration analysis then quantifies the relationship between the accuracy of a person's judgements and their confidence in the accuracy those same judgements. A simple measure of calibration is the *bias score*, which is the average confidence rating across all test items minus the proportion of the same items that were judged correctly. A positive bias score implies overconfidence; a negative bias score implies underconfidence.

In terms of situational awareness, a well-calibrated individual is one who has a high level of actual SA *and* correctly perceives this to be the case in his or her perceived SA. This can be assessed by correlating SA probe responses with confidence ratings in those responses. It is of course possible for an individual to be poorly calibrated with respect to SA. In the worst case, the individual is excessively overconfident—he has low actual SA but does not realise this and instead has high confidence in it. In this case, his decision-making is likely to be most error-prone.

A graph of individuals' confidence plotted against accuracy (proportion correct) is called a calibration curve (Keren, 1991). In Fig 1, for example, results are averaged for two groups of individuals, older and younger car drivers. In this example from Lee *et al*. (1997), accuracy refers to subjects' SA whilst being presented with safety-related electronic messages by an Advanced Traveller Information System (ATIS). Situational awareness was measured using a two-alternative forced-choice version of the SAGAT technique (that is, subjects were asked a direct question such as "What is your current speed?" and then given two choices from which to select the correct answer). After each query, subjects rated their confidence in their answer on a continuous scale (50%-100%[1]). Results showed that older drivers had slightly lower accuracy than younger drivers and yet were also distinctly over-confident about their accuracy, while younger drivers were relatively better calibrated.

[1] The confidence rating is regarded as equivalent to a subjective estimate of the probability that the SA query was answered correctly. The chance level of probability for a two-alternative forced choice paradigm is 50%, therefore this is taken as the minimum of the range for subjective probability estimates.

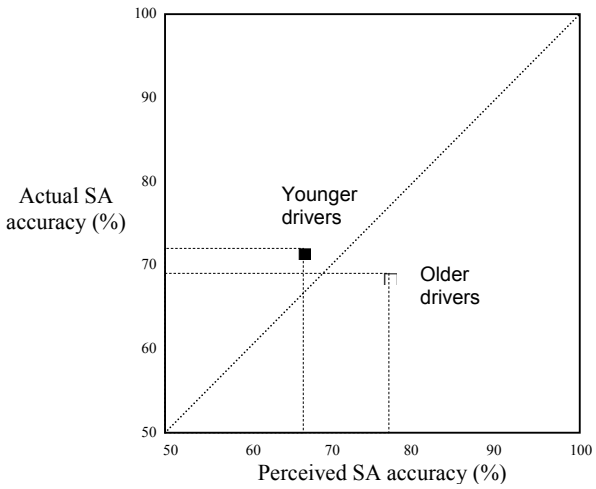


Figure 1: Calibration curve for actual and perceived SA of older and younger drivers in the study by Lee *et al.* (1997).

A similar method for assessing calibration is built into the QUASA technique. In this case, SA accuracy is quantified as the proportion of probes answered correctly (i.e., hits plus correct rejections) while perceived accuracy is obtained from self-ratings of confidence in individual probe responses. So far a five-point rating scale of confidence has been used, but this is subject to review (this is discussed below).

SIGNAL DETECTION THEORY

Signal Detection Theory (SDT) originated as a model of perceptual judgement, describing and analysing how people perform in tasks in which they must detect a certain type of stimulus. A "decision-making theory of visual detection" was first proposed by Tanner & Swets (1954), who showed how not only sensory processes but also decision processes can play a role in perceptual tasks. The theory posits that individuals do not merely react to stimuli in an automatic fashion—when confronted with uncertainty, they also set about deciding whether what they perceive signifies one thing rather than another. That is, they make a judgement.

The SDT framework is applicable to any situation in which an observer has to determine, on the basis of potentially incomplete or ambiguous evidence, which of two possible states of the world is actually the case. SDT describes the problem as one of discriminating between two types of stimulus, a **signal** and **noise**. The task is to detect specific signals but not confuse them with non-signals and other irrelevant stimuli (noise). The correct acceptance of a stimulus as a signal is referred to as a **hit**. However, two kinds of error are possible: a **false alarm** is the incorrect acceptance of a non-signal, while a **miss** is the incorrect rejection of a true signal.

The SDT model posits two stages in the signal detection process. First, perceptual evidence is detected and presumably aggregated in the brain as some unspecified kind of neural activation (termed the *internal response*). Second, an external response is generated according to the degree of evidence. Because elements of noise (both external or internal to the observer) can also trigger the internal response to some degree, and given the risk of making two kinds of error, the

individual may adopt a cognitive strategy for responding to uncertain stimuli (for example, to err on the side of caution by always "rejecting" when uncertain).

In analysing observers' response data in a perceptual task, SDT provides two quantitative measures of signal detection performance, both of which can be determined as parametric or nonparametric statistics:

(1) **Sensitivity** (denoted d' [parametric] or A' [nonparametric]) is the individual's actual ability to discriminate true signals from non-signals.

(2) Response **criterion** or **bias** (denoted β [parametric] or C [nonparametric]) specifies the setting of the individual's accept/reject criterion; in the case of bias, it quantifies the individual's response strategy for dealing with ambiguous stimuli. A conservative bias leans toward rejecting, a liberal bias leans toward accepting. Whether a conservative, liberal or neutral bias is best is entirely dependent upon the situation and the goals of the observer.

In the decades since its origination, the SDT framework has been adopted for the study of such diverse real-world tasks as military target detection, motorists' detection of vehicles and signs, and diagnostic tasks in such fields as medicine, forensics, information retrieval, weather forecasting, survey research, aptitude testing, and polygraph lie detection. Another recent extension of SDT has been to the analysis of recognition memory. In short, the framework extends far beyond the simple detection of sensory signals and applies to a host of varied tasks all of which can be characterised as instances of the same fundamental detection problem (Birdsall, 1955; Swets *et al*., 1961; Egan, 1975).

SDT and SA probes

How can the SDT approach be used to assess SA? In probe-based techniques, SA has typically been assessed as the proportion of queries that are responded to correctly. While this seems an obvious statistic to use in terms of face validity, on its own it fails to provide a full picture of the subject's awareness because it confounds sensitivity and response bias (Swets & Pickett, 1982). Does a low percentage correct reflect poor sensitivity or a highly conservative response strategy (responding selectively only when absolutely certain)? Discriminating between subjects' sensitivity on the one hand and response strategy on the other could be invaluable for understanding patterns in people's situation assessments.

In the case of true/false SA probes, a hit can be defined as the correct acceptance of a true description of the situation (see Table 1). SA sensitivity can thus be interpreted as an individual's ability to correctly discriminate between valid and invalid descriptions of the situation. That is, an individual with good SA should make fewer misses and false alarms when responding to true/false SA probes. We would no doubt expect experts, for instance, to show greater sensitivity to SA probes than novices. We might also predict that sensitivity would improve in a setting of enhanced information sharing, collaboration and shared awareness, as the individuals concerned would be aware of a more complete picture of the situation.

Decision criterion or bias in responses to true/false SA probes can be interpreted as an individual's leaning towards either more readily accepting or more readily rejecting situational descriptions when he or she is uncertain as to their validity. It remains to be seen through future research whether, say, a conservative

bias reflects a stable disposition of the individual (e.g., a marked scepticism towards unconfirmed information), or a more dynamic situation-specific strategy (reflecting, for example, a recent occurrence of misinformation), or is in fact peculiar to the probe technique and is unrelated to SA itself.

		Subject's response	
		Accept	Reject
Probe type	True	**HIT**	**MISS** (error)
	False	**FALSE ALARM** (error)	**CORRECT REJECTION**

Table 1: Contingency table showing the four possible outcomes of a true/false probe response, depending on type of probe (True or False) and the response made (Accept or Reject).

EXAMPLE OF QUASA IN USE

To date we have developed this technique through a number of iterations by conducting SA assessments in a number of military trials of varying size, the most significant being the second Limited Objective Experiment (LOE 2), a multinational experiment held in February 2003.

LOE 2 experiment

Led by the U.S. Joint Forces Command, LOE 2 involved five nations plus representatives from NATO collaborating via the Collaborative Federated Battle Lab Network, a secure online environment designed to facilitate allied experimentation. The experiment focused on the 'operational net assessment' (ONA) process by which distributed and collocated teams can collaborate online in the development of a coalition knowledge base, in this case for an emerging (fictional) crisis situation. This multinational task was used to test collaboration and information sharing across different security domains.

SA probes

Part of the LOE2 analysis activity focused on human factors aspects of situational awareness and shared awareness (led by my colleague Andrew Leggatt). To this end, 58 players located in five countries were asked to answer SA probe statements at 2-hourly intervals. The probes were descriptive statements of elements of the situation of interest compiled both from baseline knowledge in the ONA database and new information that was added during the experiment. As an example:

The Commander of the [...] Air Force has recently resigned over corruption charges.

Equal numbers of true and false probe statements were carefully formulated, with the probe construction process going through several iterations. It was important to avoid skewing the subjects' responses by, for example, inadvertently cueing the presence of false statements. Simply negating a true statement to construct a false statement is often transparent to the reader. Each probe therefore went through a process of checks prior to its use in the experiment. First, the probes were shown to independent evaluators (who had no involvement in the experiment or knowledge of the scenario), who were asked to judge the likelihood of each statement being true or false based purely on the given wording. When inadvertent cues were found, the wording of a probe was altered to make it more neutral. Second, the probes were assessed by a German human factors analyst for intelligibility. Probes were then altered if necessary to ensure that the non-native English speakers would be able to understand them clearly. Finally, each probe was evaluated by an intelligence specialist for its operational significance. Only those statements that passed all three tests were used in the final probe set.

Five probes were presented to all subjects during each bi-hourly interruption. Each statement was followed by a prompt to select *True* or *False* and then a second prompt to select a confidence level from five options: *Very low, Low, Medium, High, Very high*. The subjects were instructed to complete the questions in silence, without consulting other participants or the database. They were also asked not to discuss the questions after presentation. When the answer to a probe was not known, subjects were instructed to make a guess at the true/false response and then indicate they were guessing by marking 'Very low' on the confidence scale.

Results

By the end of the two-week experiment, 45 subjects had answered at least 100 true/false probes each. Hit rates and false alarm rates were found both for each individual subject and for each of the five national teams, and then used to generate measures of sensitivity (d′) and response bias (β). Going by *correct responses* alone, we found that all teams achieved a similar SA accuracy score of approximately 0.71 (±0.02). Team A was found to have the highest *hit rate* (0.81), however, the other four nations averaging 0.68 (Fig 2). Did this mean that team A had the highest situational awareness? To address this, we now turn to the sensitivity scores which are, in principle, an index of actual SA.

In terms of team-level *sensitivity* there was little difference between the nations, with team averages for d′ ranging from 1.1 to 1.3. In other words, one nation was as good as any other at discriminating true versus false statements (Fig 3). A wide range of sensitivity scores was found across individuals, though, with a few exhibiting very good discrimination between true and false statements (d′ > 2.0) and a few showing little such ability (d′ < 0.5).

What accounts for team A's higher hit rate but average sensitivity? The answer lies in an examination of response bias scores. When averaged across all subjects, response bias was essentially neutral (β = 1.1, s.d. = 0.4). The overall spread, however, ranged from 0.4 (distinctly liberal) to 2.6 (distinctly conservative). When we look at the response biases of the national teams, one team alone, team A, shows a liberal bias (β = 0.7). That is, team A had a consistent tendency to over-accept statements as true—and so had the highest false alarm rate (0.45)—while other nations were on average either neutral or leaning to conservative bias in their probe

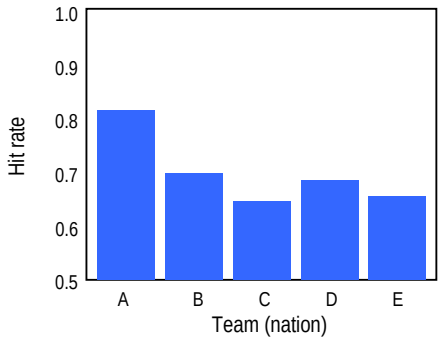


Figure 2: Mean SA probe hit rates per team in LOE 2.

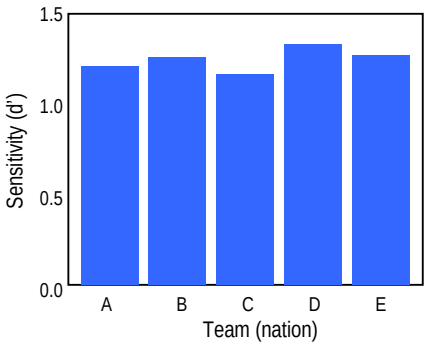


Figure 3: Mean SA probe sensitivity scores per team in LOE 2.

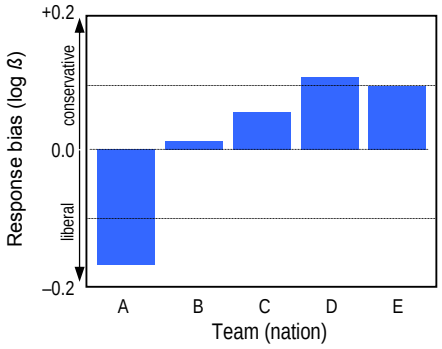


Figure 4: Mean SA probe response bias (log ß) per team in LOE 2.

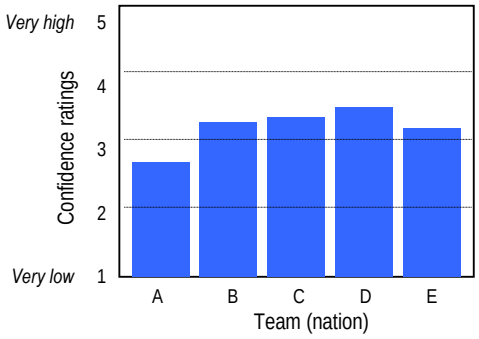


Figure 5: Mean SA probe confidence ratings per team in LOE 2.

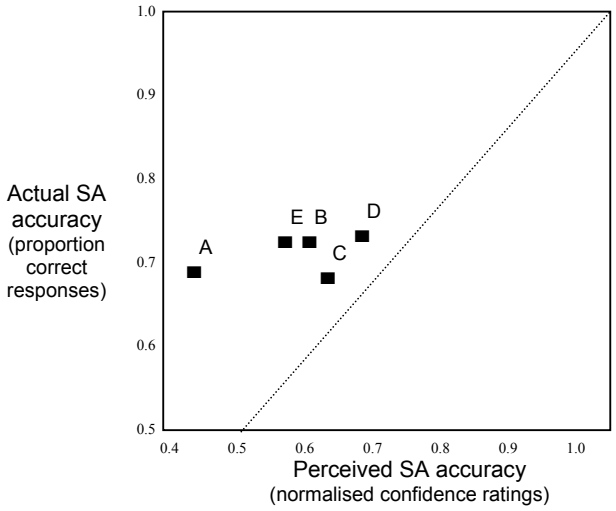


Figure 6: SA probe calibration curve for average responses per team in LOE 2.

responses (Fig 4). Because of this tendency, team A's seemingly superior hit rate was merely an effect of erring on the side of accepting more uncertain statements.

It was also found that team A had the lowest overall confidence in probe responses (Fig 5). To assess SA calibration, average confidence ratings were

normalised and probe accuracy scores (proportion of hits plus correct rejections) were subtracted from the result to provide a calibration bias statistic. The results are illustrated in the calibration curve shown in Fig 6. As we can see, all teams were leaning into the under-confident portion of the graph; however, team A is by far the least well calibrated (bias = –0.26), while two teams, C and D, are quite well calibrated (bias = –0.08 and –0.07 respectively).

In summary, the QUASA technique of combining true/false SA probes with simultaneous self-ratings of confidence for each probe was applied in the LOE 2 experiment and yielded a set of useful and potentially insightful quantitative results. The SA statistics were also able to shed light on information from post-experiment debriefings and the analysis of performance data. Note that at this time the methods of SDT were applied only to the actual true/false probe data. A next step will be to apply this also to the analysis of confidence data, as explained below.

TYPE 2 SDT ANALYSIS

In a generic signal detection task, the observer must judge which of two possible states (signal present or not present) is actually the case in a given observation. Clark *et al.* (1959) further distinguished between two such kinds of discrimination task, labelled *Type 1* and *Type 2* (Fig 7).

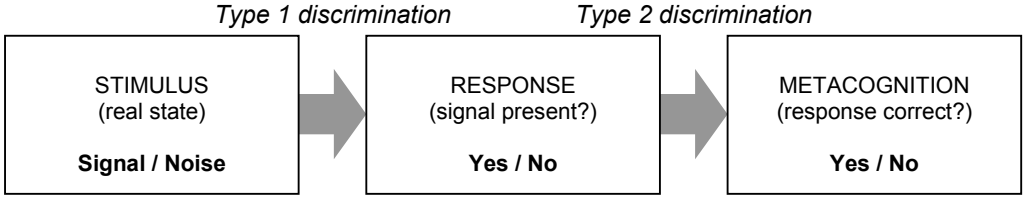


Figure 7: The relationship between Type 1 and Type 2 discrimination tasks.

In a Type 1 task, the observer must discriminate between two possible *stimulus* states. For example, a radiologist observes a patient's X-ray and decides whether or not the patient has a tumour. The essential feature of a Type 1 task is that the situation to be discriminated exists independently of the observer. A Type 2 task, in contrast, is a metacognitive judgement about a Type 1 discrimination. In a Type 2 task, the observer discriminates between two possible situations that are defined by the observation just made: whether it was actually *correct* or *incorrect*. The only fundamental difference between the two types of discrimination lies in the nature of the 'signal' to be detected: an external stimulus in the Type 1 case, and the correctness of one's response to that in the Type 2 case. Otherwise, the methods of Signal Detection Theory still apply as shown in Table 2. If, for example, the observer judges that a particular Type 1 response he or she has just made is correct, and in fact it *is* correct, then we have a Type 2 hit.

Surprisingly little use has been made of Type 2 judgements in the context of applied Signal Detection Theory, however (Macmillan & Creelman, 1991). A notable exception is the study by Kunimoto *et al.* (2001) in which confidence ratings were subjected to SDT analysis to evaluate the extent of implicit learning in an artificial grammar learning task. More recently, Galvin *et al.* (2003) have presented a general theory of Type 2 decisions combining the paradigms of both SDT and calibration theory, showing how to derive probability functions underlying Type 2 decisions from

		Type 2 decision about Type 1	
		"Correct"	"Incorrect"
State of Type 1 decision	Correct	**HIT**	**MISS** (error)
	Incorrect	**FALSE ALARM** (error)	**CORRECT REJECTION**

Table 2: Contingency table showing the four possible outcomes of a Type 2 (metacognitive) judgement about the state of a Type 1 response (Correct or Incorrect).

those for the Type 1 task used. The theory predicts that the probability distributions for Type 1 and Type 2 decisions are not equivalent, even if both are based on the same evidence (internal response state).

The theory presented by Galvin *et al*. (2003) assumes that the Type 2 task is a binary decision task (or at least can be treated as such). The binary decision options are effectively either "My response was correct" (*C*) or "My response was incorrect" (*I*). In other words, the individual decides whether signal event *C* or non-signal event *I* has occurred. Galvin *et al.* (2003) point out that "a meaningful Type 2 decision can be made only if the Type 1 decision that precedes it is [also] binary, because the observer must be either right or wrong for the event *C* or the event *I* to have occurred" (p. 847). Their Type 2 Signal Detection Theory therefore applies to a binary rating of confidence or perceived correctness pertaining to a binary discrimination of a stimulus, such as sensory signal detection, recognition memory, and clinical diagnosis. Interestingly, in a recent comparison of subjective measures of awareness by Tunney & Shanks (2003), it has been found that participants are better able to place their levels of confidence on a binary confidence rating scale (e.g., high versus low, as used by Kunimoto *et al*., 2001) rather than a continuous scale (e.g., 0%-100%). A binary Type 2 decision appears to yield more sensitive results than a continuous scale.

Inspired by these recent developments, we now intend to incorporate Type 2 (as well as Type 1) SDT analysis in the QUASA technique. At present we are preparing to run a validation study of this form of the technique, with certain augmentations to further analyse different dimensions of team/shared SA.

CONCLUSIONS

A sophisticated technique for quantitatively analysing individuals' situational awareness has been presented. The impetus for the technique is twofold: first, the intuition that combining subjective ratings simultaneously with individual SA probes gives a much fuller picture of SA; second, the finding that applying SDT analysis to probe results yields sensitivity and bias statistics which give insights into subjects' SA that would be not available using percentage correct alone. Recent developments in SDT and calibration theory point to a further refinement of the technique, the application of SDT analysis to the self-ratings of confidence themselves as instances of metacognitive awareness.

REFERENCES

BAXTER, G.D. & BASS, E.J. (1998). Human error revisited: Some lessons for situation awareness. In *Proceedings of the 4th Symposium on Human Interaction with Complex Systems.* Los Alamitos: IEEE Computer Society Press, 81-87.

BIRDSALL, T.G. (1955). The theory of signal detectability. In: H. Quastler (Ed.), *Information Theory in Psychology* (pp. 391-402). Glencoe, Illinois: Free Press.

CLARKE, F.R., BIRDSALL, T.G. & TANNER, W.P. (1959). Two types of ROC curves and definitions of parameters. *Journal of the Acoustical Society of America, 31*, 629-630.

EBEL, R.L. & FRISBIE, D.A. (1991). *Essentials of Educational Measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

EGAN, J.P. (1975). *Signal Detection Theory and ROC Analysis*. New York: Academic Press.

ENDSLEY, M.R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors, 37* (1), 65-84.

ENDSLEY, M.R. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. In: M.R. Endsley & D.J. Garland (Eds.), *Situation Awareness Analysis and Measurement*. Mahwah, New Jersey: LEA.

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.

KEREN, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217-273.

KORIAT, A. & GOLDSMITH, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490-517.

KUNIMOTO, C., MILLER, J. & PASHLER, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness & Cognition, 10,* 294-340.

LEE, J.D., STONE, S., GORE, B.F., COLTON, C., MACAULEY, J., KINGHORN, R., CAMPBELL, J.L., FINCH, M. & JAMIESON, G. (1997). *Advanced Traveller Information Systems and Commercial Vehicle Operations Componments of the Intelligent Transportation Systems: Design Alternatives for In-Vehicle Information Displays*. U.S. Federal Highway Administration technical report FHWA-RD-96-147. McLean, Virginia.

MACMILLAN, N.A. & CREELMAN, C.D. (1991). *Detection Theory: A User's Guide.* Cambridge: Cambridge University Press.

MATTHEWS, M.D., PLEBAN, R.J., ENDSLEY, M.R. & STRATER, L.D. (2000). Measures of infantry situation awareness in a virtual MOUT environment. In: *Proceedings of the 1st Human Performance, Situation Awareness and Automation Conference,* Savannah, Georgia, 15-19 Oct 2000. SA Technologies, Inc.

MCGUINNESS, B. & FOY, L. (2000) A subjective measure of SA: The Crew Awareness Rating Scale (CARS). In: *Proceedings of the 1st Human Performance, Situation Awareness and Automation Conference,* Savannah, Georgia, 15-19 Oct 2000. SA Technologies, Inc.

MUÑIZ, E.J., STOUT, R.J., BOWERS, C.A. & SALAS, E. (1998). A methodology for assessing team situation awareness: Situation awareness linked indicators adapted to novel tasks (SALIANT). In *Proceedings of the RTO Meeting: Collaborative Crew Performance in Complex Operational Systems*. Quebec, Canada: Canada Communication Group.

STRATER, L.D., ENDSLEY, M.R., PLEBAN, R.J. & MATTHEWS, M.D. (2001). Measures of platoon leader situation awareness in virtual decision making exercises. U.S. Army Research Institute for the Behavioral and Social Sciences research report ARI 1742, Alexandria, Virginia.

SWETS, J.A. & PICKETT, R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press.

SWETS, J.A., TANNER, W.P. & BIRDSALL, T.G. (1961). Decision processes in perception. *Pyschological Review, 68*, 301-340.

TANNER, W.P. & SWETS, J.A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*, 401-409.

TAYLOR, R.M. (1990) Situation Awareness Rating Technique (SART): the development of a tool for aircrew systems design. Paper 3 in: *Situational Awareness in Aerospace Operations*, AGARD-CP-478. Neuilly-sur-Seine, France: NATO-AGARD. pp. 3/1-3/17.

TUNNY, R.J. & SHANKS, D.R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cogntiion, 31* (7), 1060-1071.

Advanced Technology Centre BAE SYSTEMS

Quantitative Analysis of Situational Awareness

(QUASA)

Applying Signal Detection Theory to True/False Probes and Self-Ratings

Barry McGuinness
Principal Scientist
Human Factors Dept
Advanced Technology Centre
BAE Systems
Bristol, UK

BAE SYSTEMS

Overview

1. Situational Awareness (SA)
2. Assessing SA
3. QUASA Approach
4. Signal Detection Theory
5. Calibration of SA
6. Example: LOE 2 data
7. Further Developments

BAE SYSTEMS

Situational Awareness

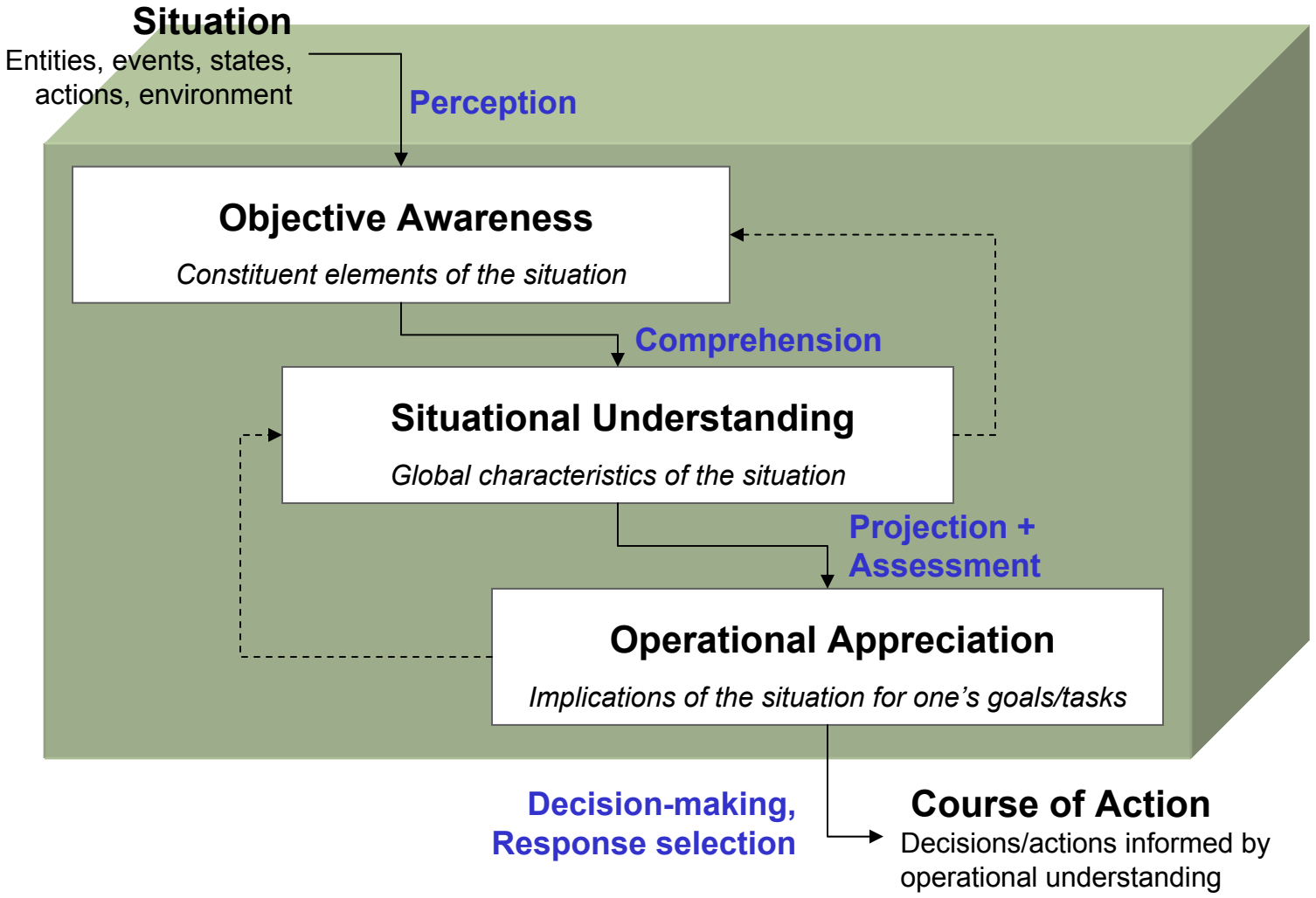
"Knowing what's going on so you can figure out what to do."

"What you need to know not to be surprised."

Who is where? What are they doing?
What's going on? Why?
What will happen next?
What does it mean for my task?

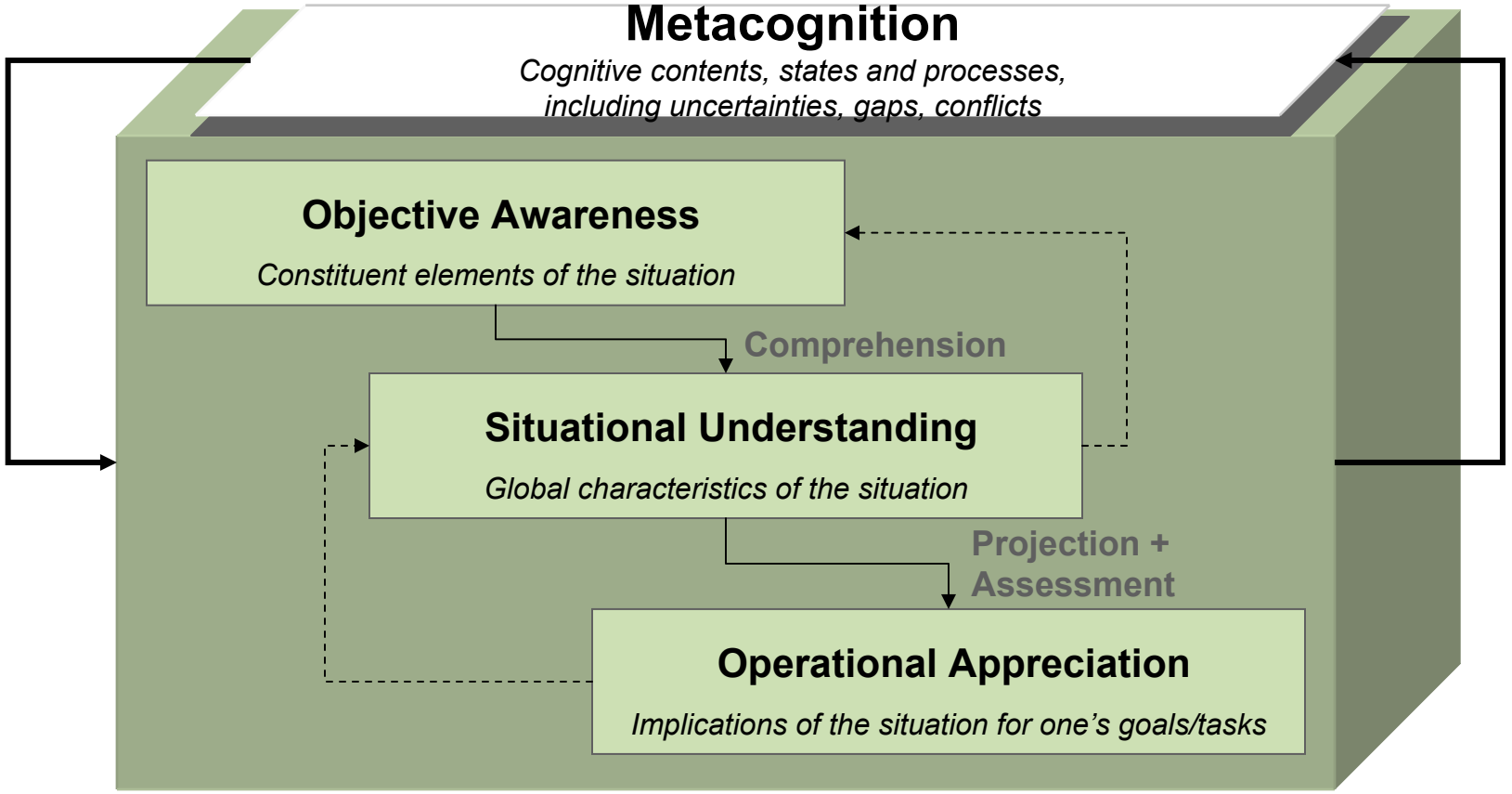
BAE SYSTEMS

Situational Awareness



BAE SYSTEMS

Situational Awareness



BAE SYSTEMS

Situational Awareness

COGNITION	METACOGNITION
• Fighting in the city has mostly ceased	• This is certain. Current info, very reliable.
• Column of red tanks is leaving south of the city	• Not sure about this. Reports may not be from reliable source. Need to check.
• Enemy is beginning retreat	• Confidence in this -- 50-60% Need to look for evidence.
"Actual SA"	"Perceived SA"

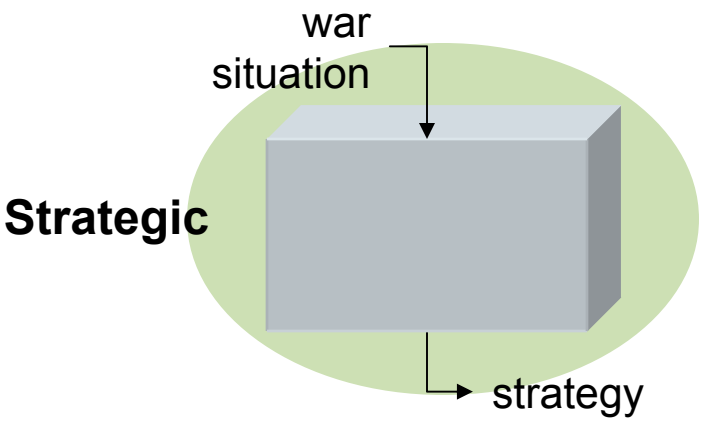
BAE SYSTEMS

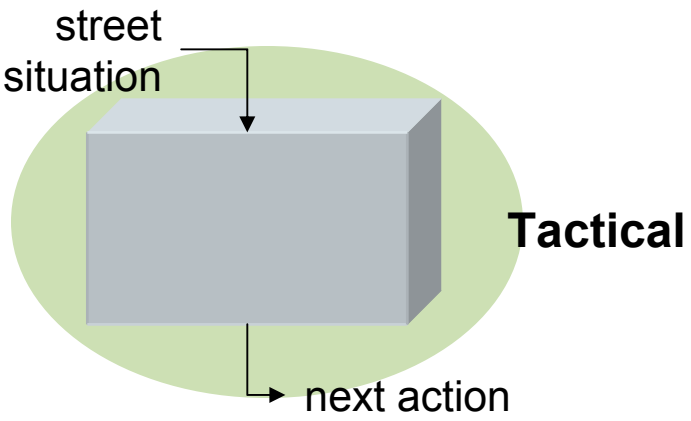
Situational Awareness

- Hence, SA is not just about having positive knowledge of actual events
- It's also about
  - **Being aware of what is <u>not</u> the case**
  - **Being aware of what we <u>don't know</u> and may need to find out**
  - **Being aware of what <u>others</u> are aware of and unaware of**
- So, SA is a complex, multi-faceted phenomenon

BAE SYSTEMS

Situational Awareness





BAE SYSTEMS

Assessment of Situational Awareness

- OBJECTIVE INDICES / CORRELATES
  - Performance
  - Behaviours
  - Physiology

SALIENT

SABARS

- SELF-RATINGS
  - Unidimensional
  - Multidimensional

SARS

PSAQ

SART

CARS

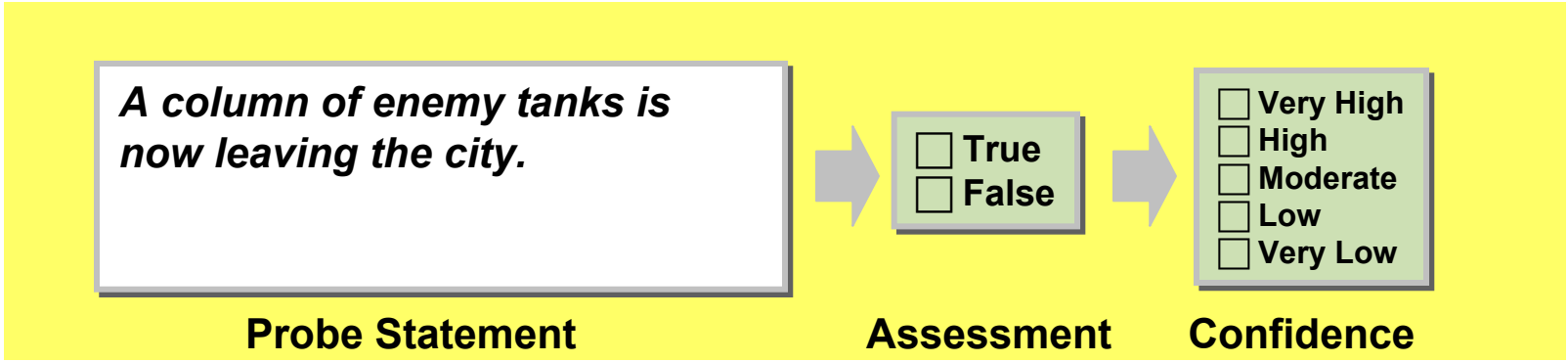
- DIRECT PROBES / QUERIES
  - Situation reports
  - Multi-choice questions
  - True / false probes

SAGAT

QUASA

QUASA

- Quantitative Analysis of SA
  - Combination of direct probes and simultaneous self-ratings
  - True/false probes
  - Responses analysed using Signal detection Theory
  - Extension of CALIBRATION theory to SA
- Probes and ratings
  - True/false probe = a statement about the situation [a 'report'] which may or may not be true.
  - Self-rating = indication of confidence in one's probe response



QUASA

SA Requirements Analysis

- A form of Cognitive Task Analysis with SMEs to capture SA contents
  - Generic for the role/task
  - Specific to the scenario

Probe construction

- Formulate equal numbers of true & false probes
- Ensure that probes are
  - relevant to the subject's task
  - plausible as potentially 'true' descriptions when in fact false
- Process of checks & iterations:
  - independent 'blind' assessment of true/false likelihood
  - assessment of intelligibility
  - assessment of plausability w.r.t. the scenario
  - assessment of relevance to the subject's task

BAE SYSTEMS

QUASA in use

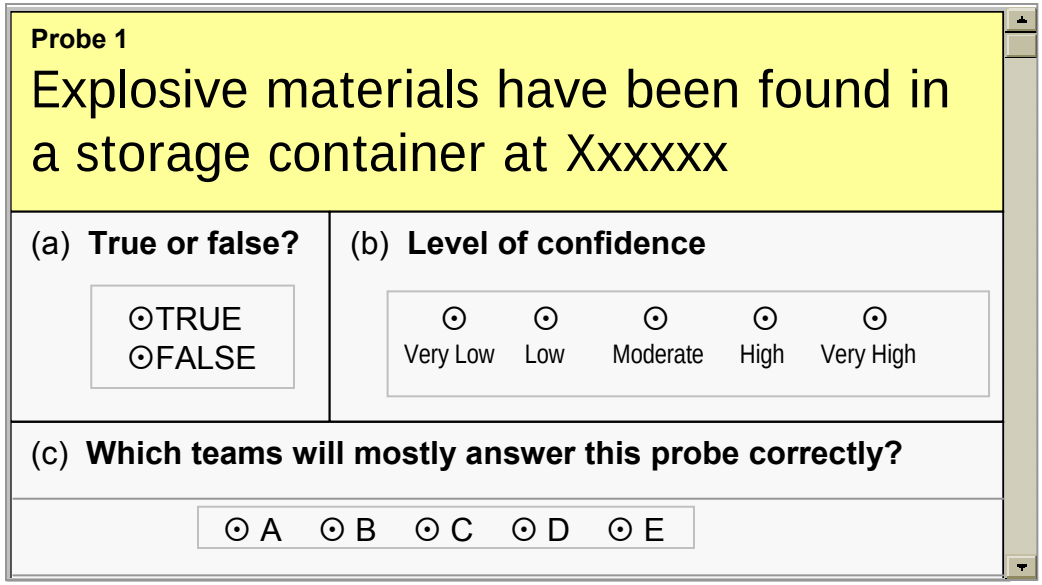
MN LOE 2 experiment

- 5 nations + NATO
  - US lead (JF COM)
- Collaborative planning
  - distributed teams
  - network
  - information sharing agreements
  - ONA process
- 46 subjects in 2 roles
  - Analysts vs Planners
- 2 conditions (methods of online collaboration), each lasting 1 week
- 50 T/F probes per subject per condition
  - 5 at a time every few hours

BAE SYSTEMS

QUASA in use

LOE 2 SA data collection

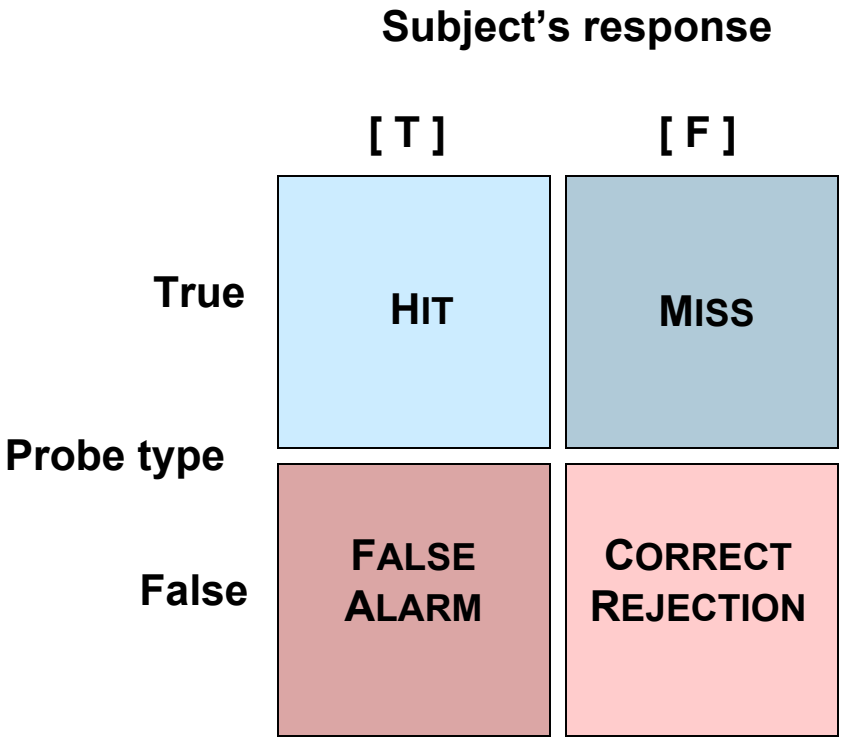


- True / false probe
- Subjective confidence level
- Perception of other teams' SA

BAE SYSTEMS

Analysis of probes data

Contingency table



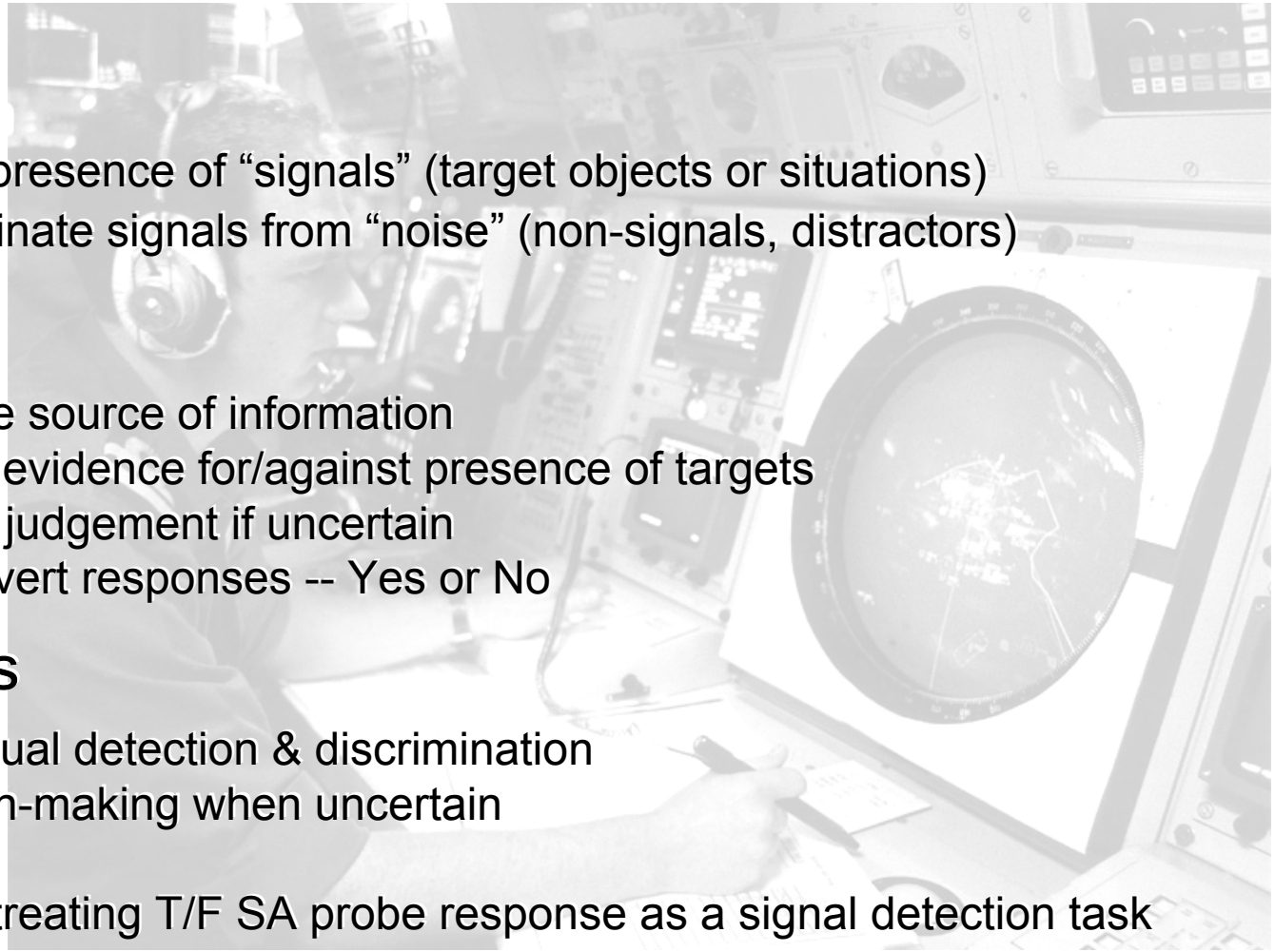
Enemy forces have captured bridge Charlie. [T] [F]

BAE SYSTEMS

Signal Detection Theory

BAE SYSTEMS

Signal Detection Theory



Goal

- Detect presence of “signals” (target objects or situations)
- Discriminate signals from “noise” (non-signals, distractors)

Task

- Observe source of information
- Assess evidence for/against presence of targets
- Make a judgement if uncertain
- Make overt responses -- Yes or No

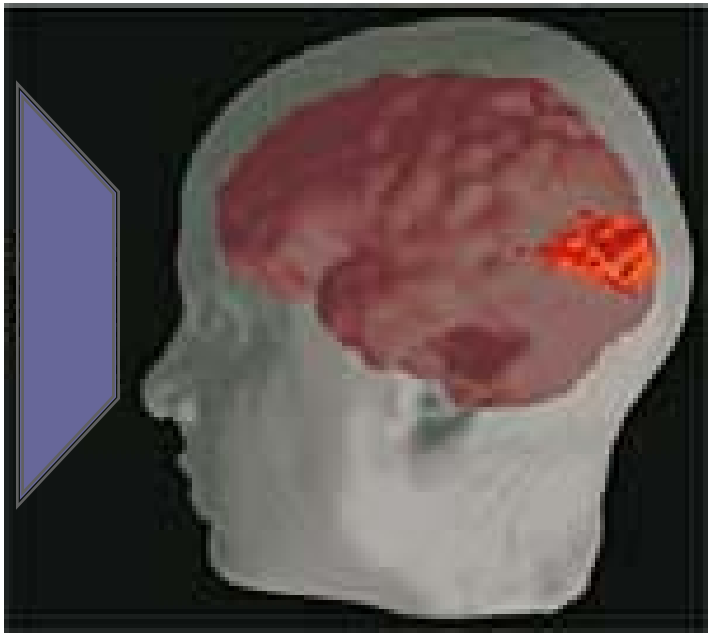
Processes

- Perceptual detection & discrimination
- Decision-making when uncertain

… We’re treating T/F SA probe response as a signal detection task

BAE SYSTEMS

Signal Detection Theory



Non-signals
(Noise)

Signals

low

high

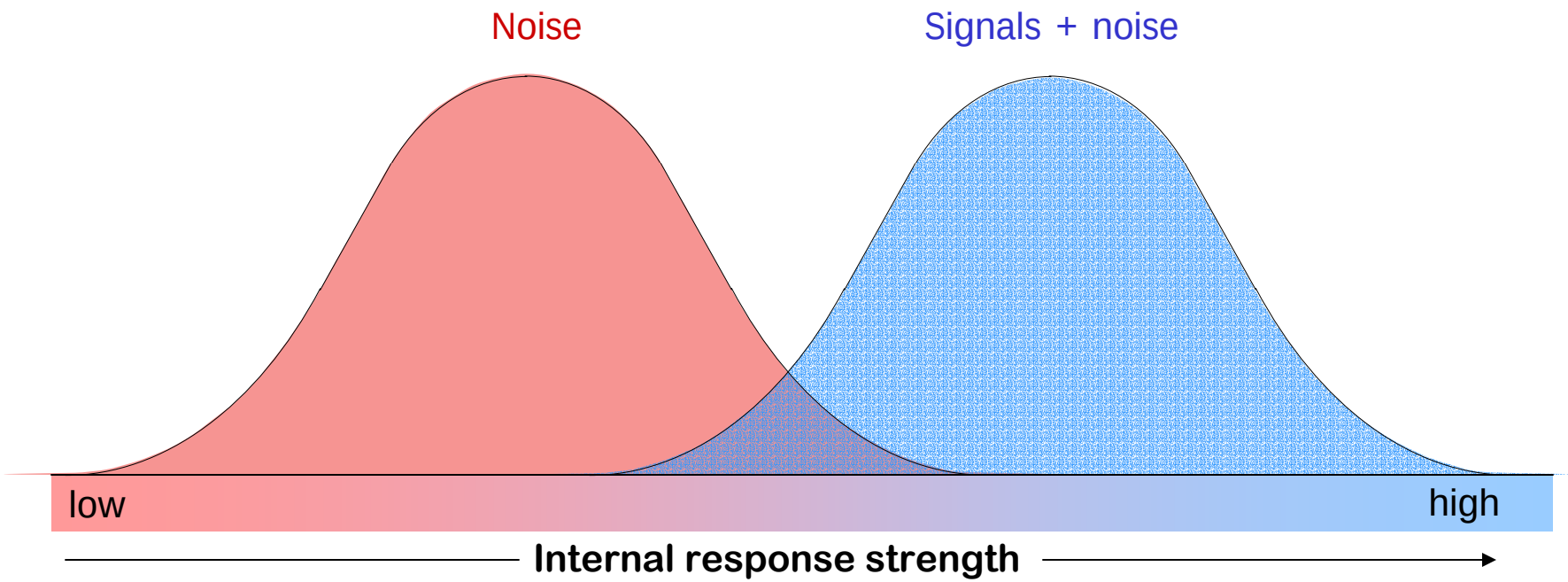
Internal response strength

REJECT

ACCEPT

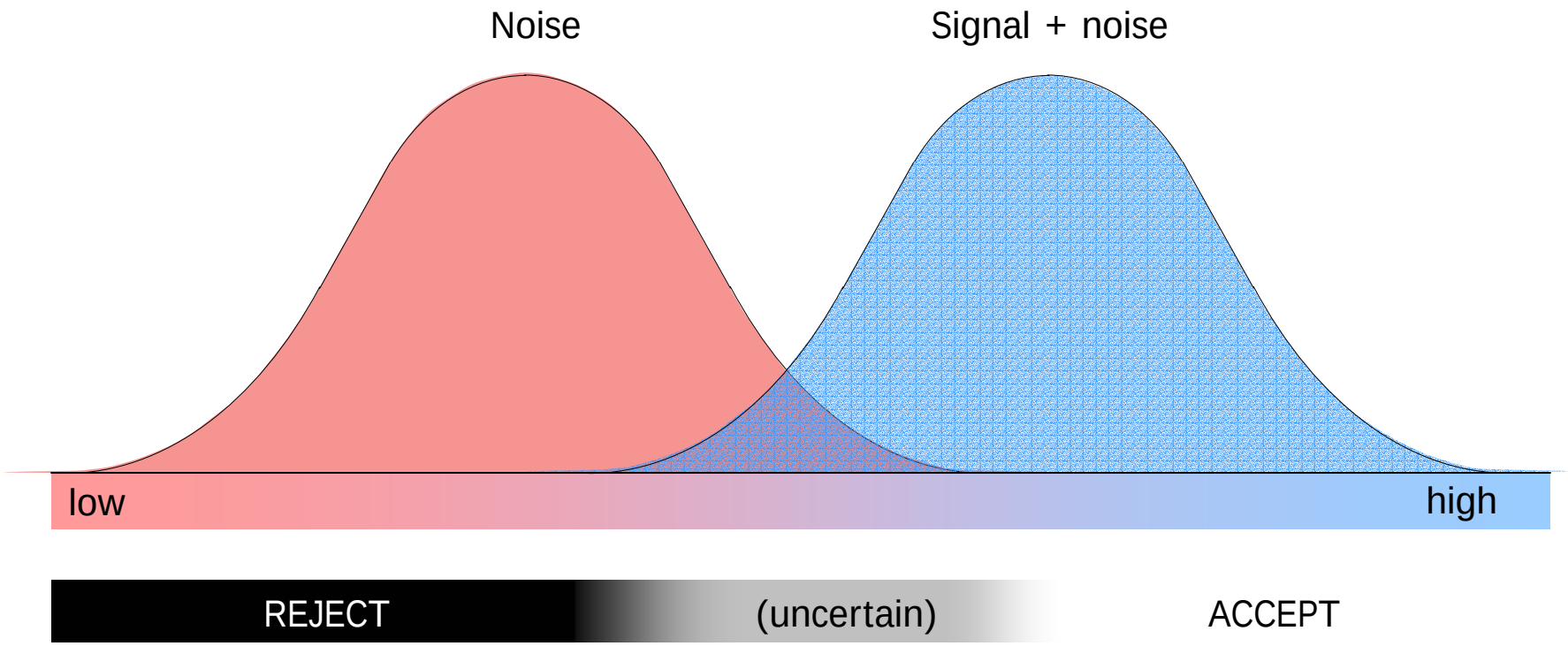
BAE SYSTEMS

Signal Detection Theory



BAE SYSTEMS

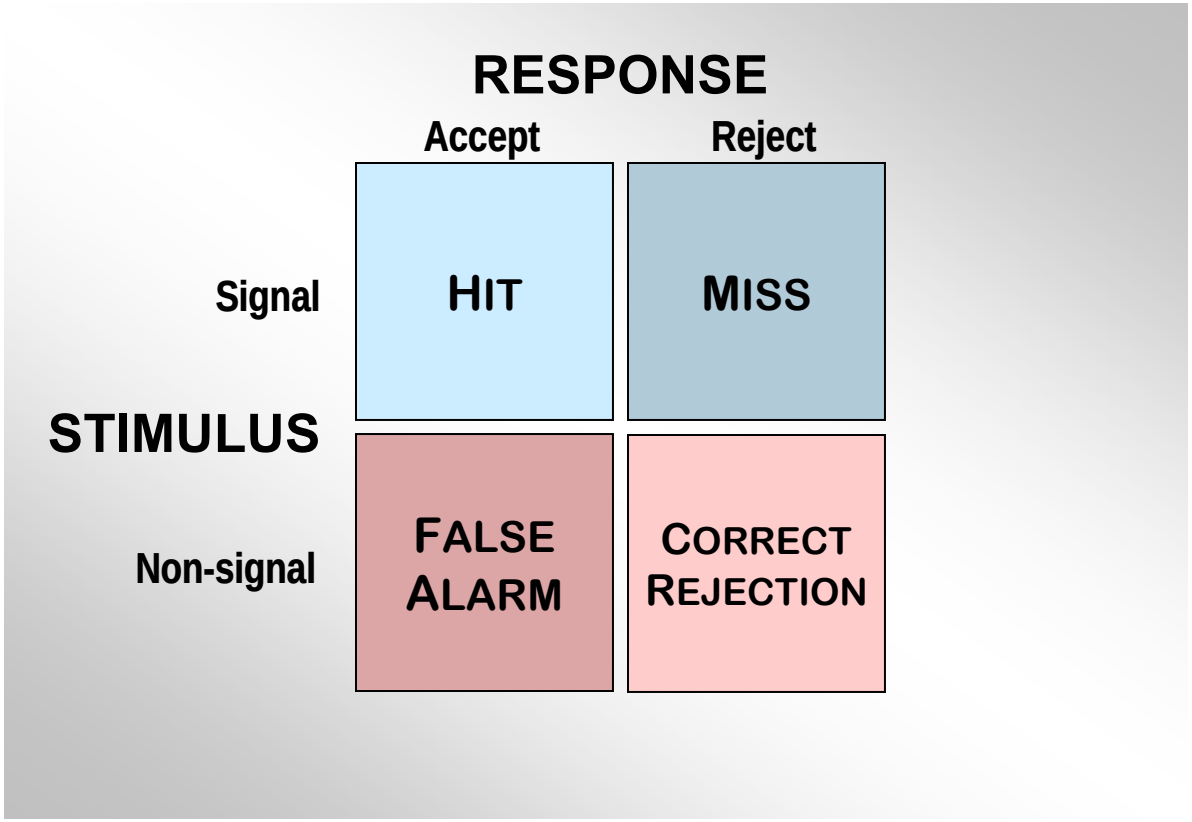
Signal Detection Theory



BAE SYSTEMS

Signal Detection Theory

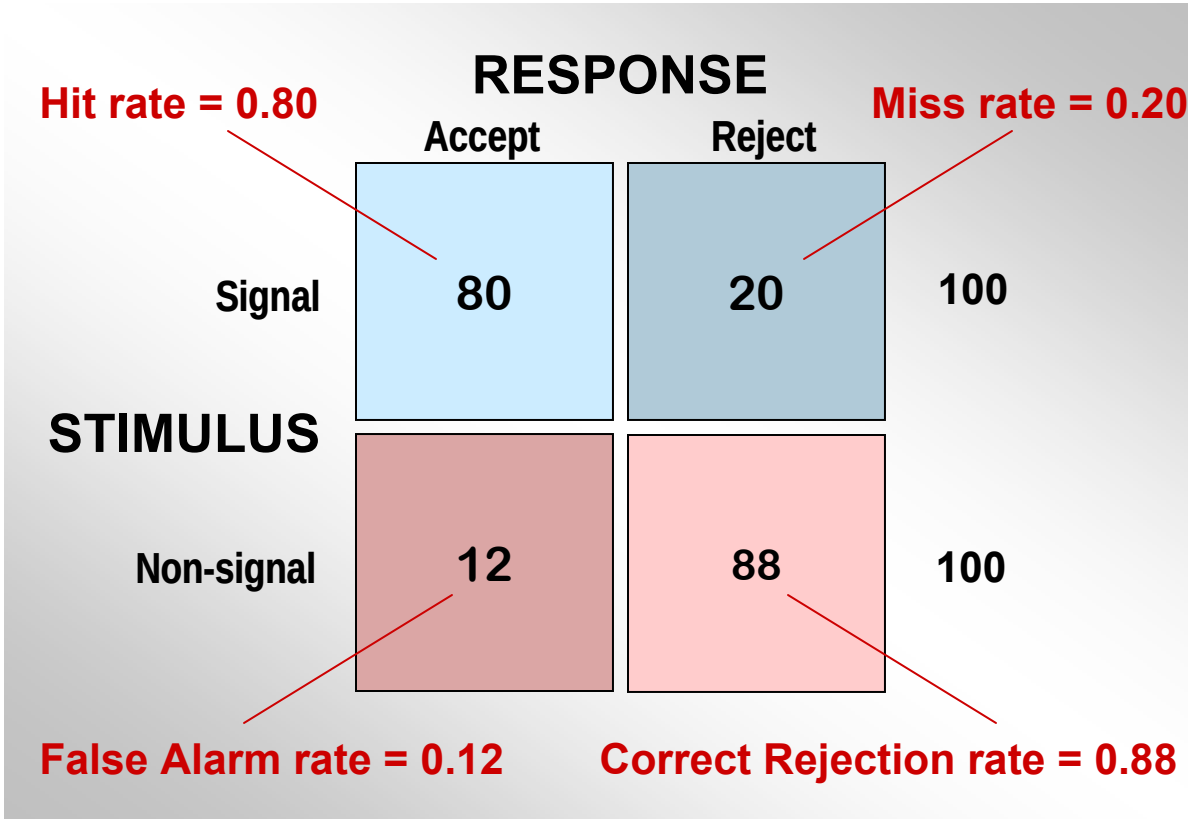
- Contingency table -- 4 possible outcomes



BAE SYSTEMS

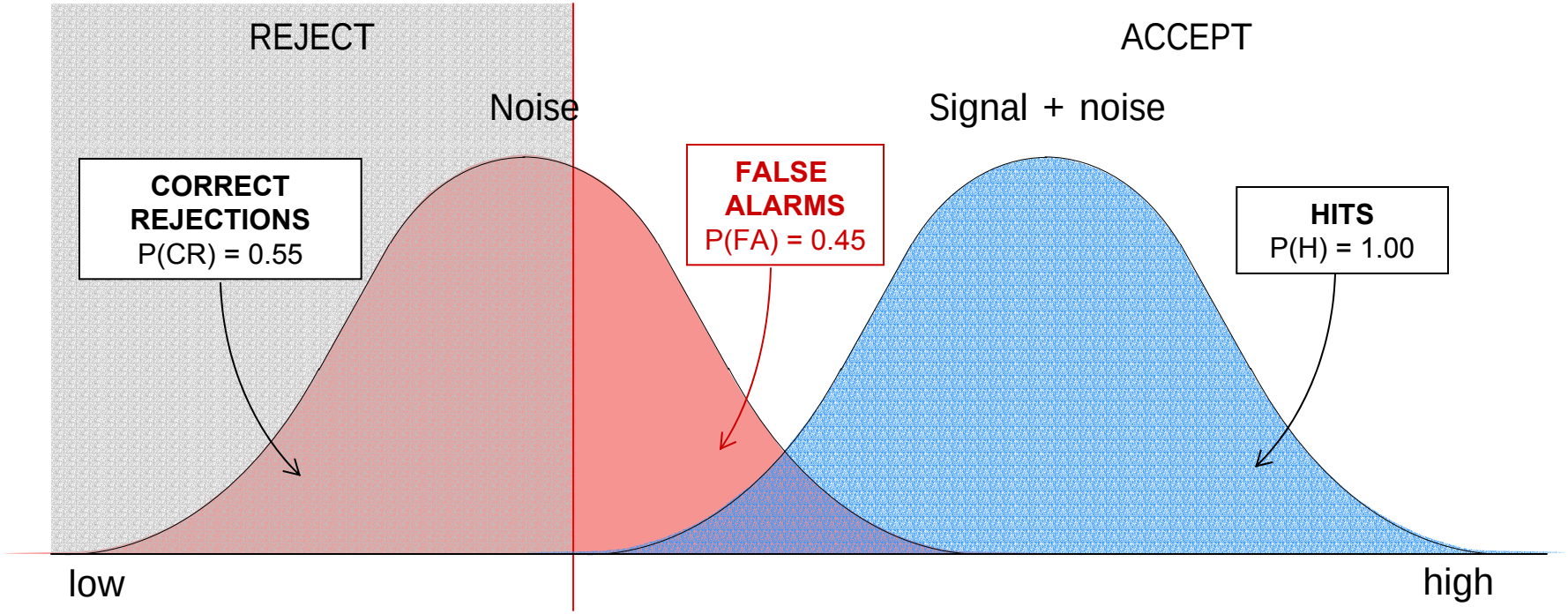
Signal Detection Theory

- Contingency table -- 4 possible outcomes



BAE SYSTEMS

Signal Detection Theory



Low criterion (liberal, inclusive)

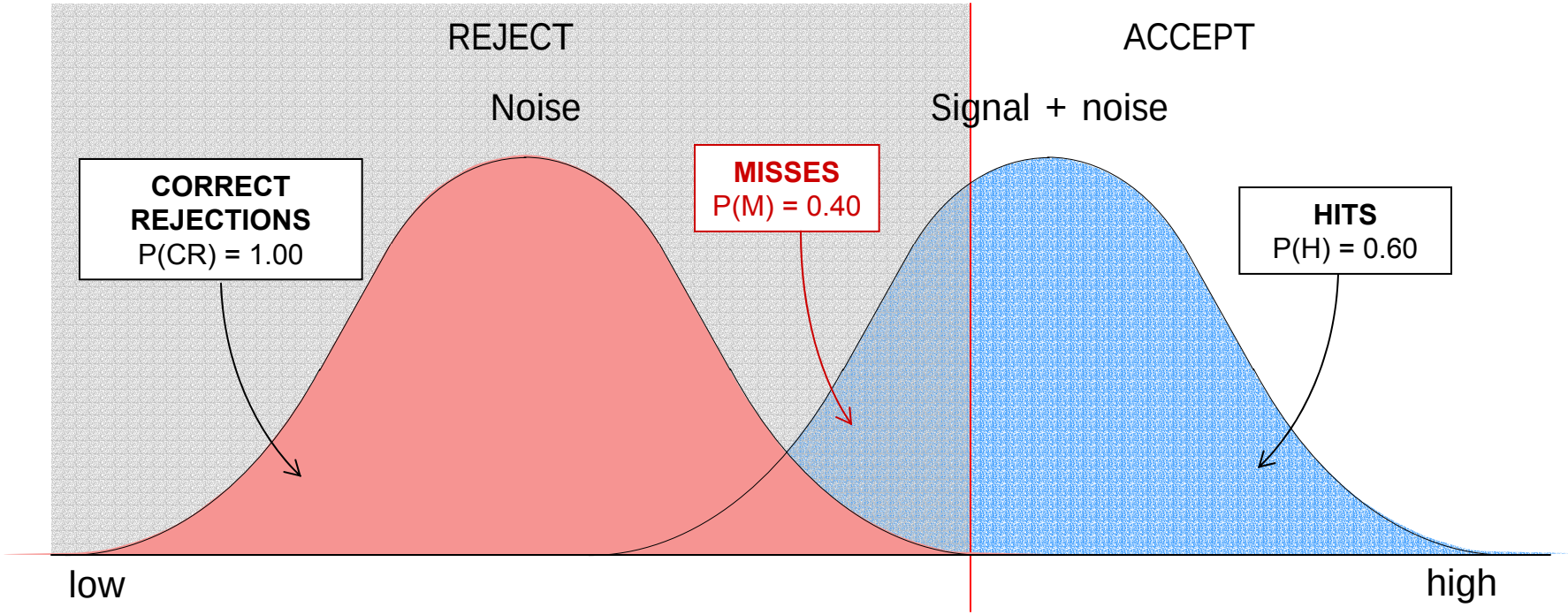
Letting no true signal slip through the net

Maximum hits, no misses

Prone to false alarms

BAE SYSTEMS

Signal Detection Theory



High criterion (conservative, exclusive)

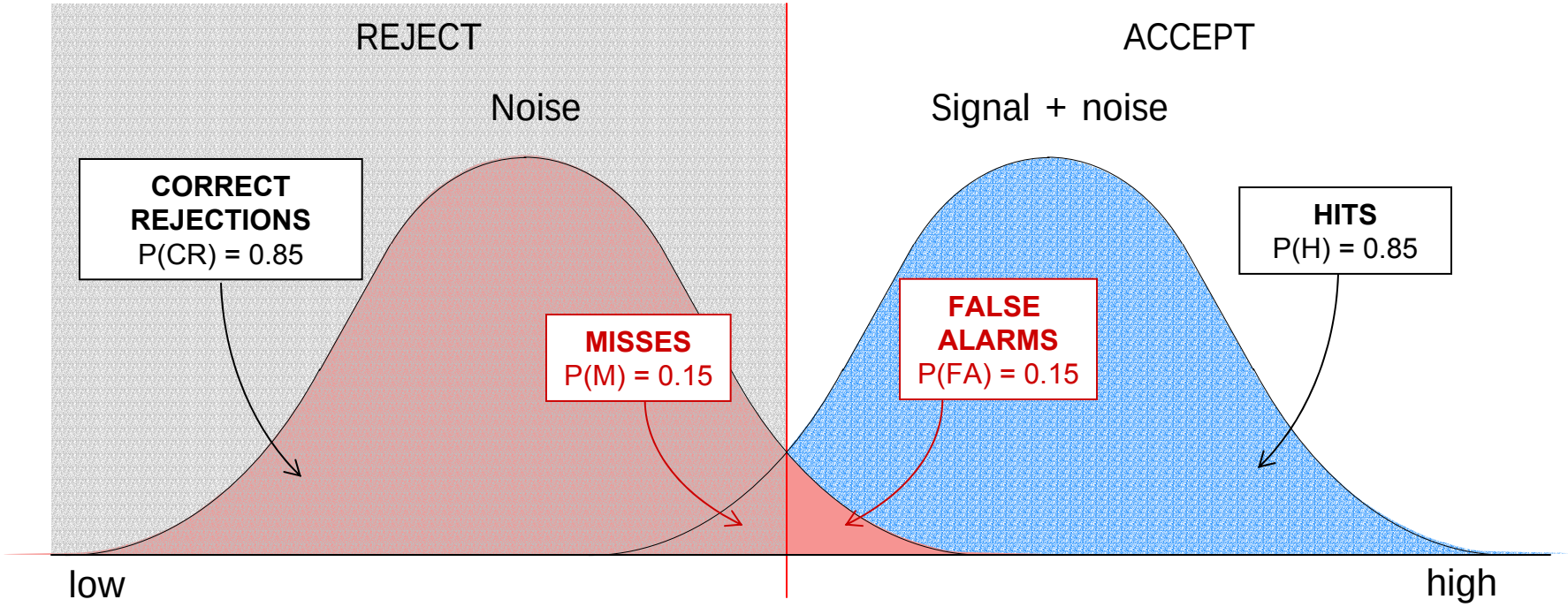
Accepting nothing but definite true signals

Maximum correct rejections, no false alarms

Prone to misses

BAE SYSTEMS

Signal Detection Theory



Central criterion (neutral, balanced)

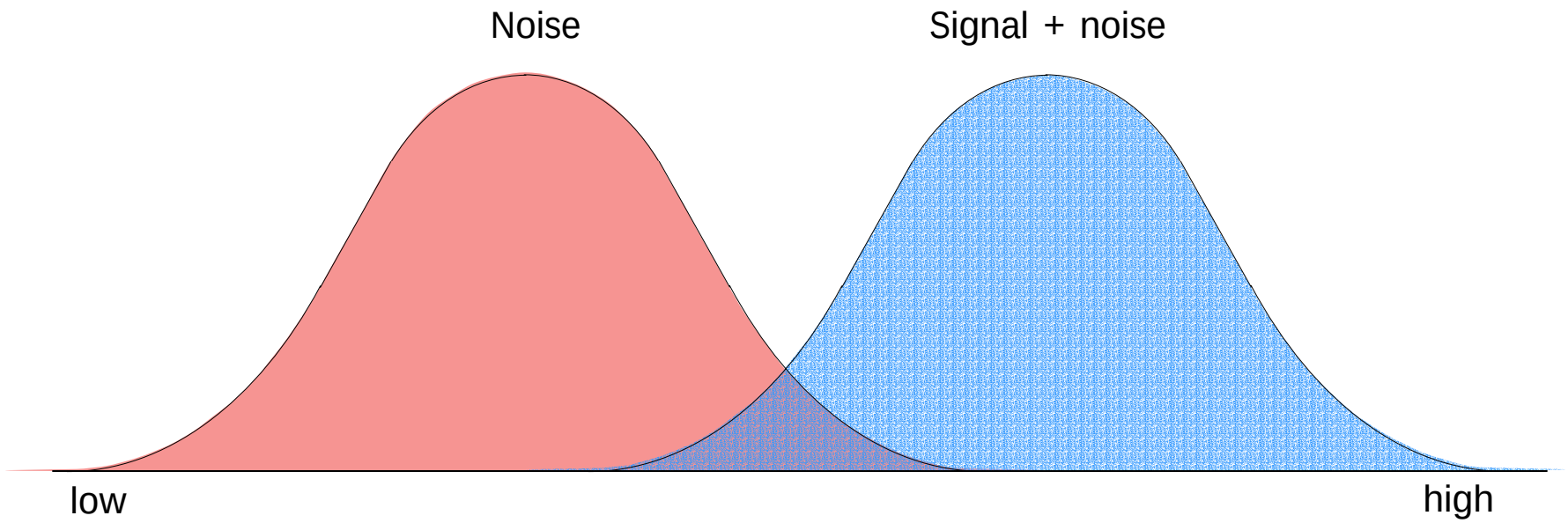
Threshold set at the mid-point of uncertainty

Equal numbers of misses and false alarms

Prone to equal numbers of misses and false alarms

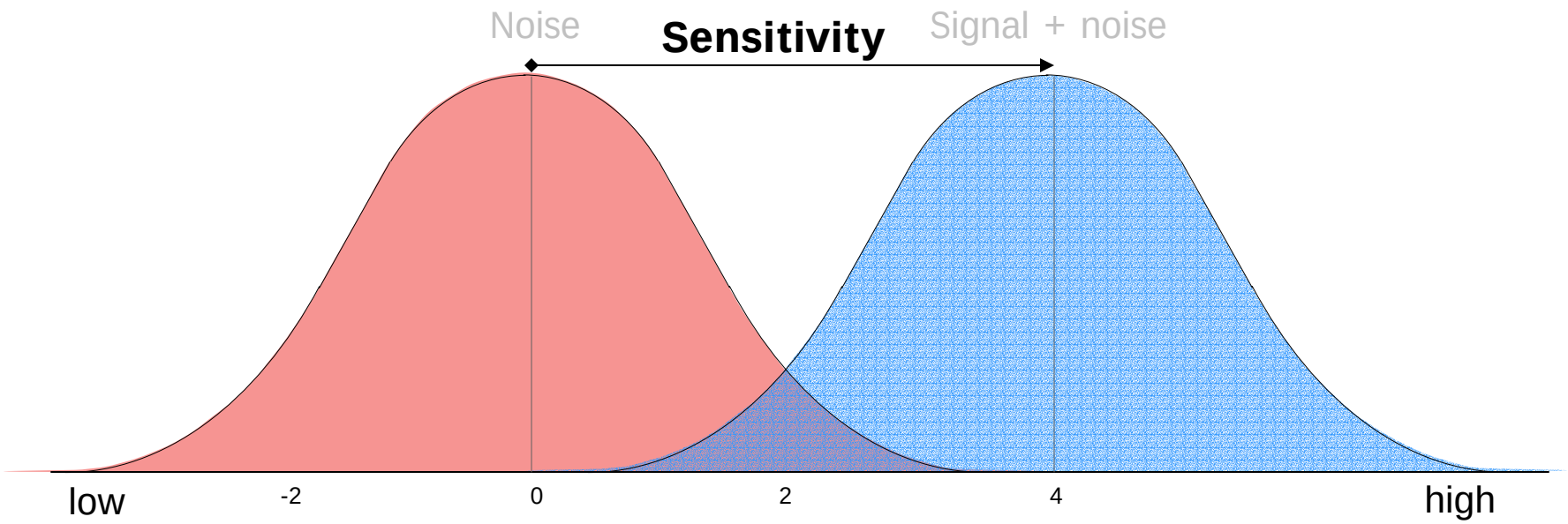
BAE SYSTEMS

Signal Detection Theory



BAE SYSTEMS

Signal Detection Theory



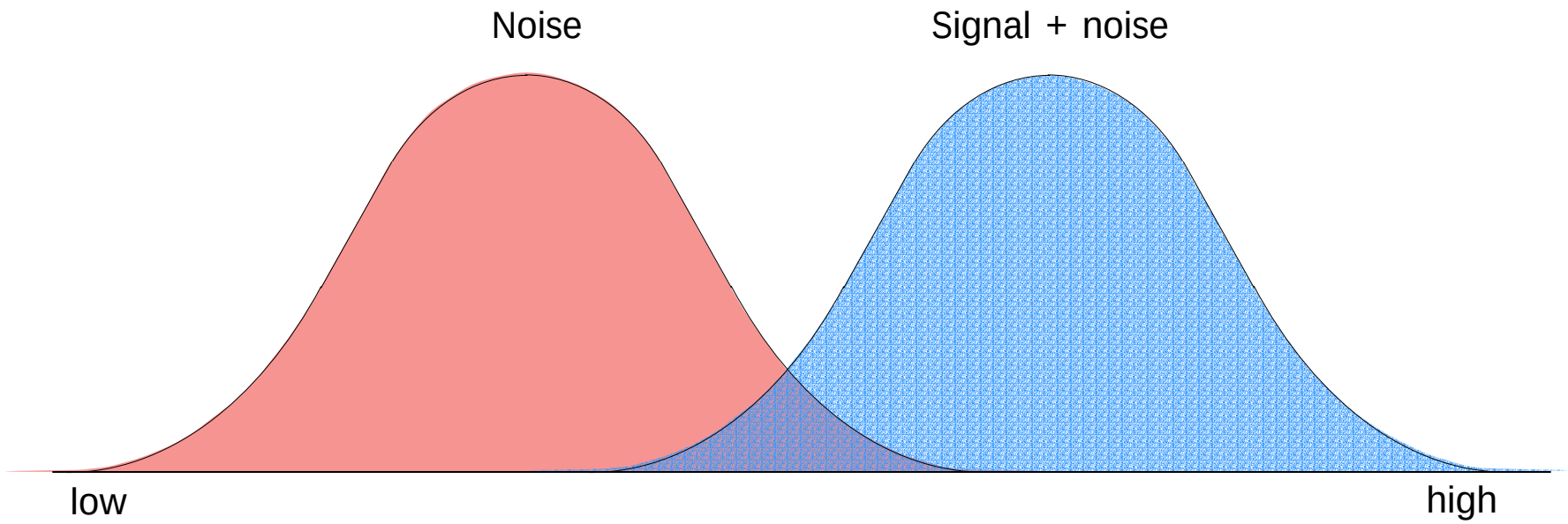
Sensitivity
Difference between noise and signal distributions, relative to their spread (variance)

d' = Z(H)-Z(FA)

d' = 4.00

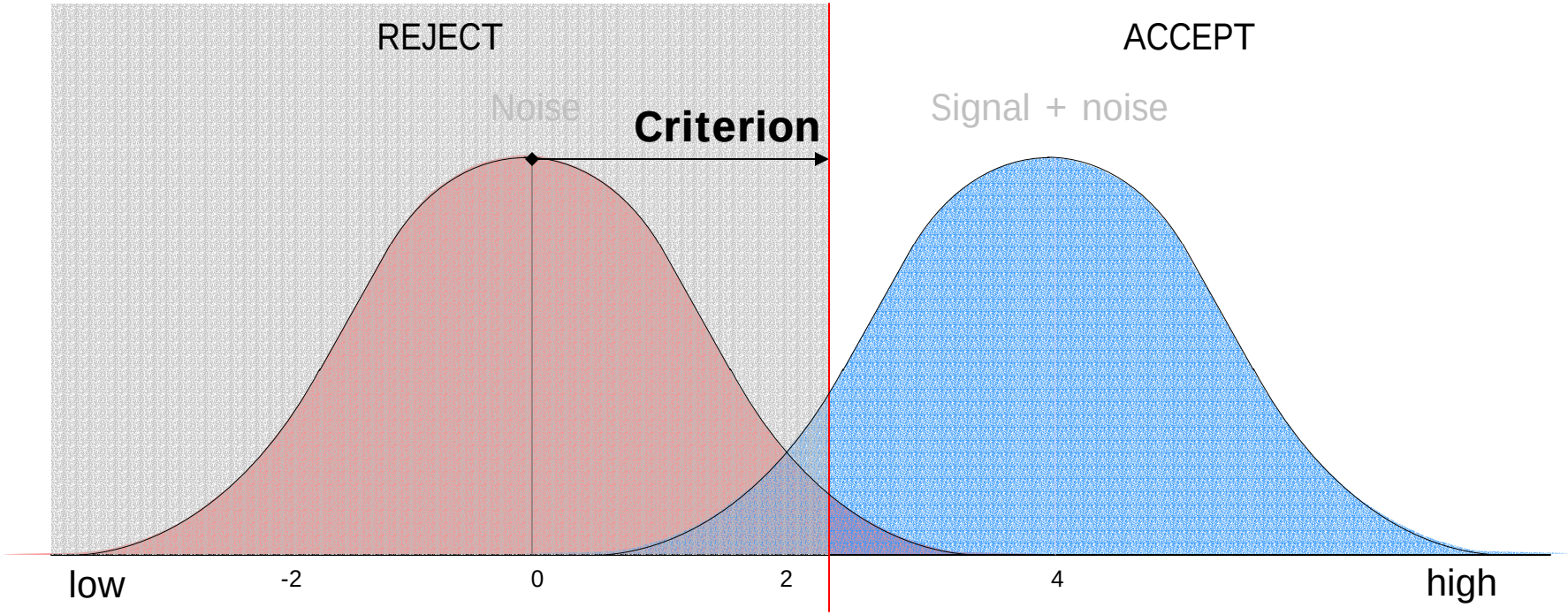
BAE SYSTEMS

Signal Detection Theory



BAE SYSTEMS

Signal Detection Theory



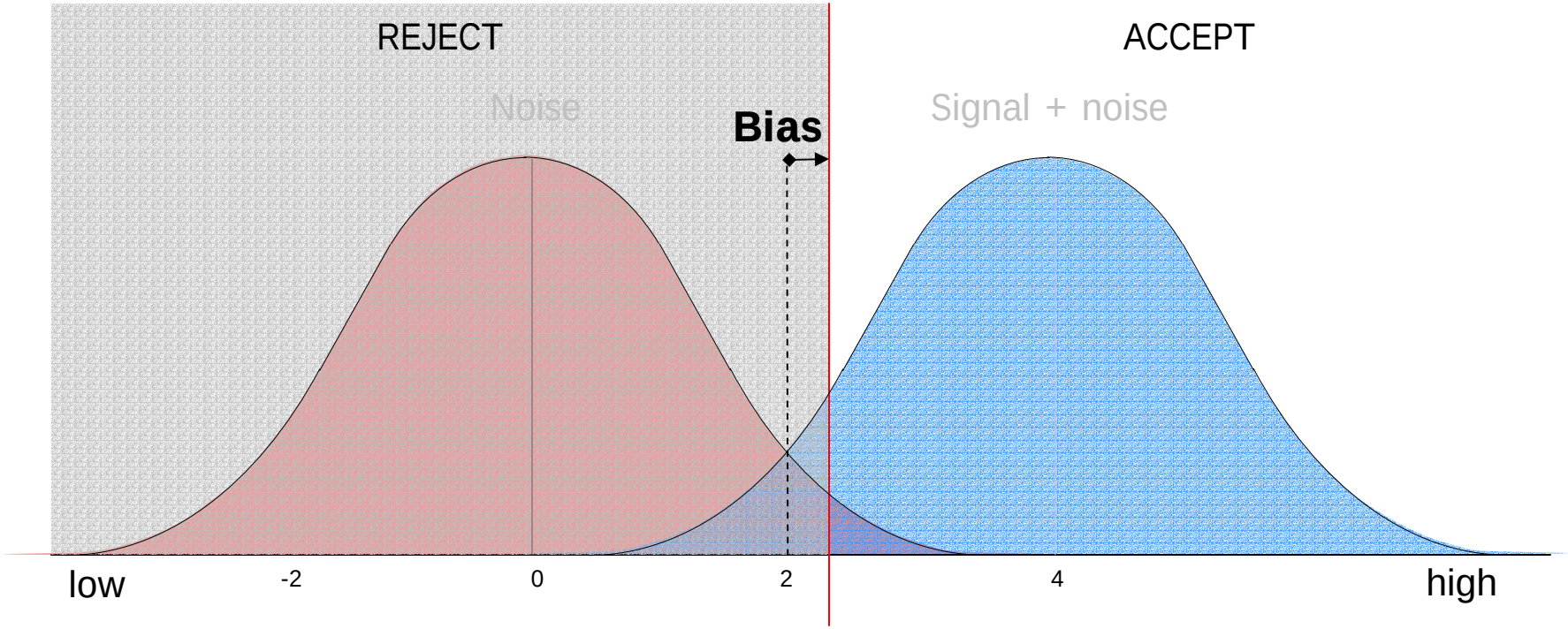
Criterion
Threshold for "accept" response, measured by distance from middle of noise distribution

k = –Z(FA)

k = 2.16

BAE SYSTEMS

Signal Detection Theory

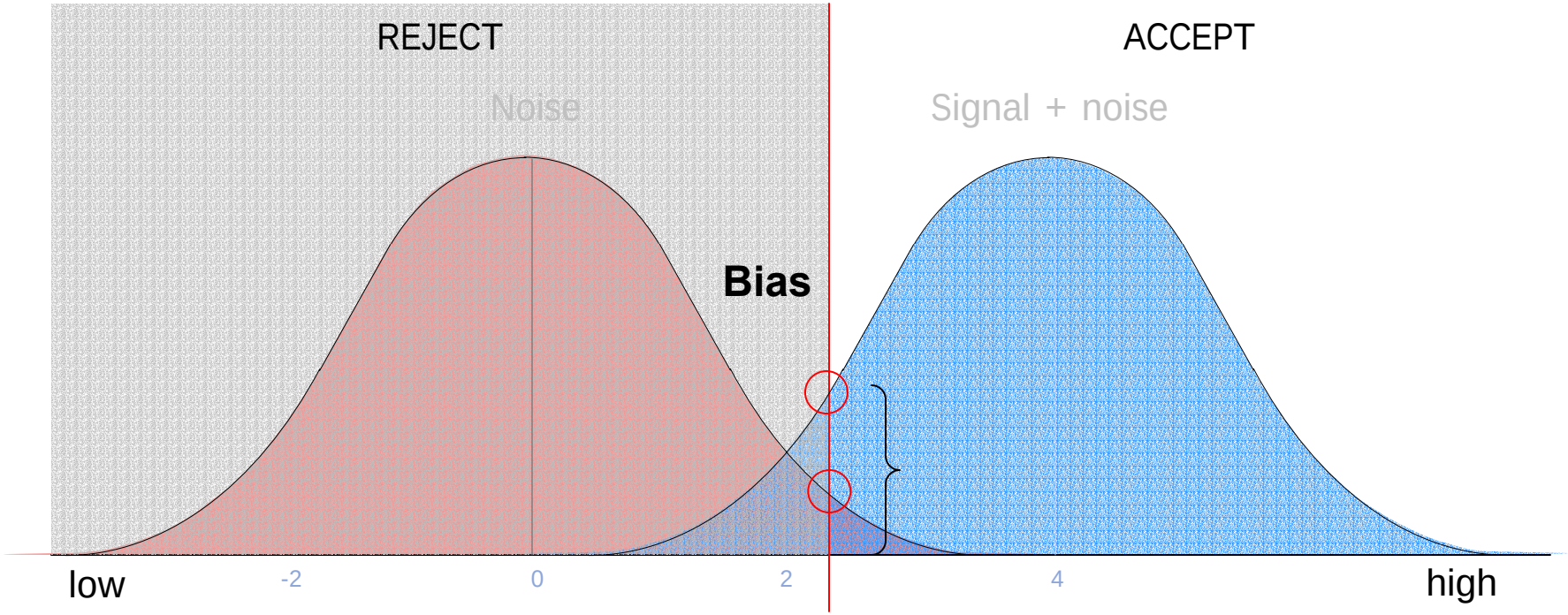


Bias (1)
Distance of actual criterion from neutral or central criterion

C = k **–** d'/2
C = 2.16 – 2.00
= 0.16

BAE SYSTEMS

Signal Detection Theory



Bias (2) and (3)
Likelihood ratio of probability densities of the two distributions at the criterion

$ß = f_S(k)/f_N(k)$ $\quad \log ß = ½(Z^2(FA)–Z^2(H))$

$ß = \exp^{d'C}$ $\quad \log ß = d'C$

$ß = 1.38$ $\quad \log ß = 0.32$

BAE SYSTEMS

Signal Detection Theory

Basic findings

– Perceptual performance depends upon

STIMULUS DISCRIMINABILITY

- **Stimulus quality**
- **Actual signal-noise ratio**

OBSERVER SENSITIVITY

- **Ability to detect signals**
- **Ability to discriminate signals from noise (distractors)**

OBSERVER RESPONSE STRATEGY IN UNCERTAINTY (CRITERION / BIAS)

- **Perceived signal probability**
- **Motivation to maximise hits or minimise false alarms**

– SDT has established that individuals are not just mechanical information processors but also make conscious judgements in conditions of uncertainty

BAE SYSTEMS

Signal Detection Theory

- SDT in the real world
  - Early studies of radar observer performance
  - More recently:
    - **Recognition memory**
      - eyewitness memory
      - remember / know paradigm
    - **Diagnostic tasks**
      - medical tests
      - weather forecasting
      - psychometric tests
      - polygraph lie detectors
      - forensic tests
  - In principle, any situation that calls for judgement in uncertainty

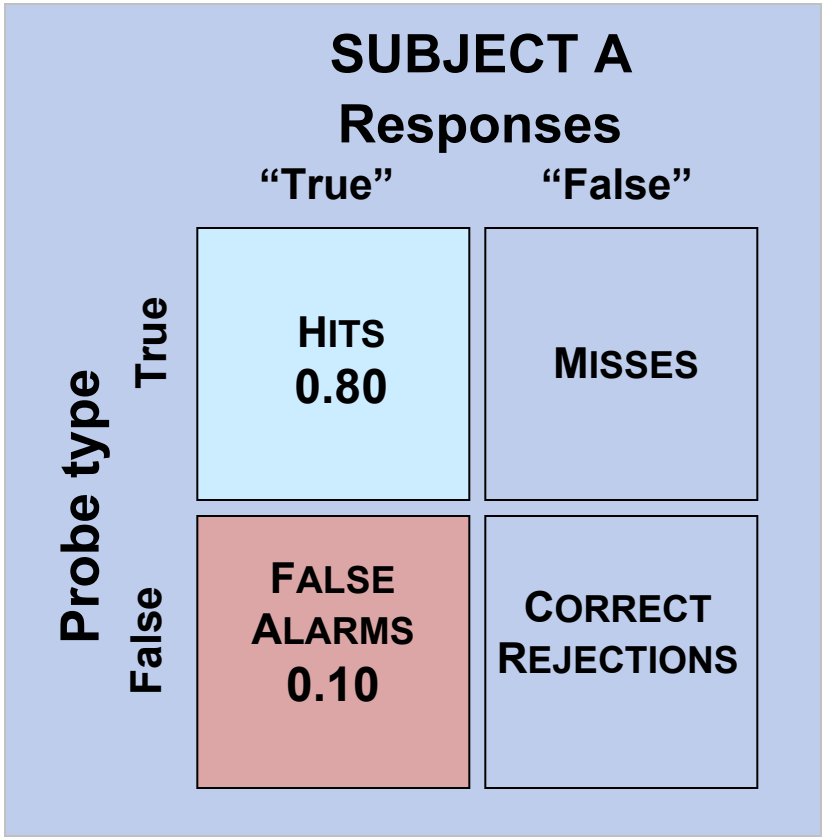
BAE SYSTEMS

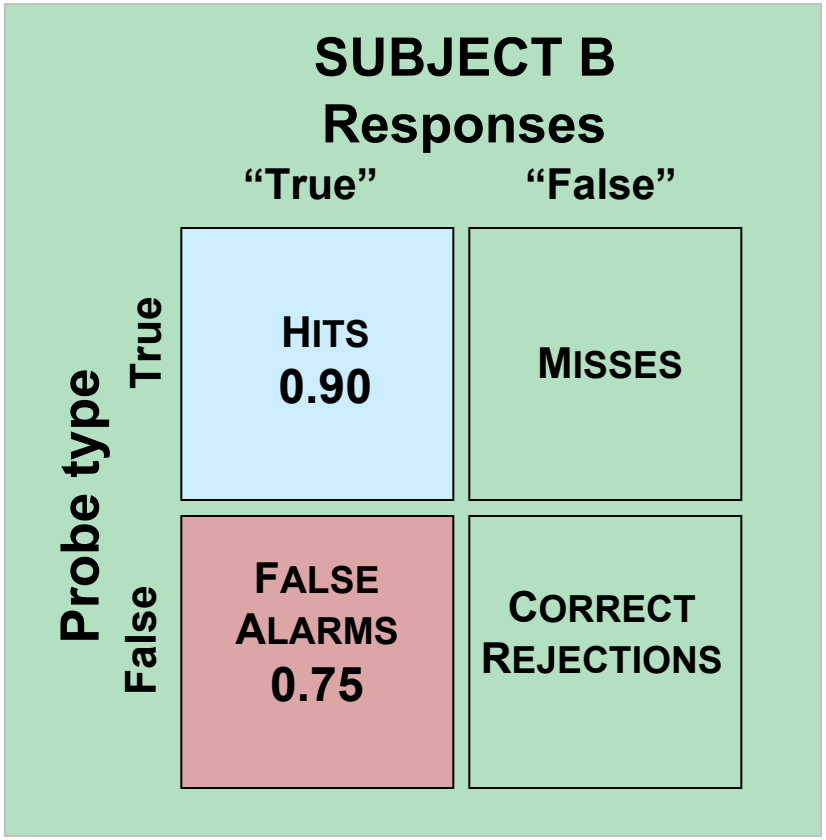
SDT and Situational Awareness

- **Assessing SA with T/F probes**
  - Why use them?
  - Output of T/F probes = contingency table
    - **HITS / MISSES**
    - **FALSE ALARMS / CORRECT REJECTIONS**
  - Traditionally, we have assessed SA using % correct responses to questions about the situation
  - This tells us little or nothing about
    - **What the subject knows is not the case**
    - **What the subject wrongly believes is the case**
  - SDT provides separate measures of SENSITIVITY and CRITERION / BIAS

Results

– Compare two subjects (LOE 2)





BAE SYSTEMS

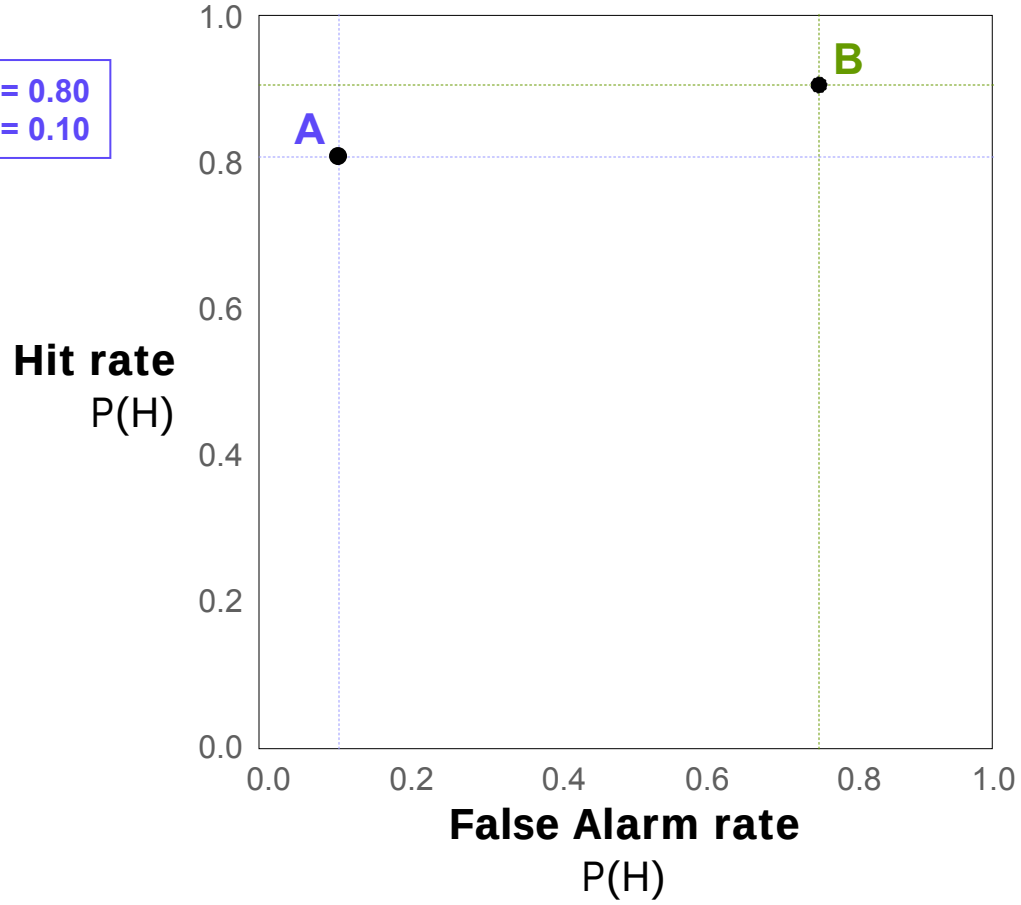
Reciever Operating Characteristic

A

Hit rate = 0.80
FA rate = 0.10

B

Hit rate = 0.90
FA rate = 0.75



BAE SYSTEMS

ROC – Criterion / Bias

A

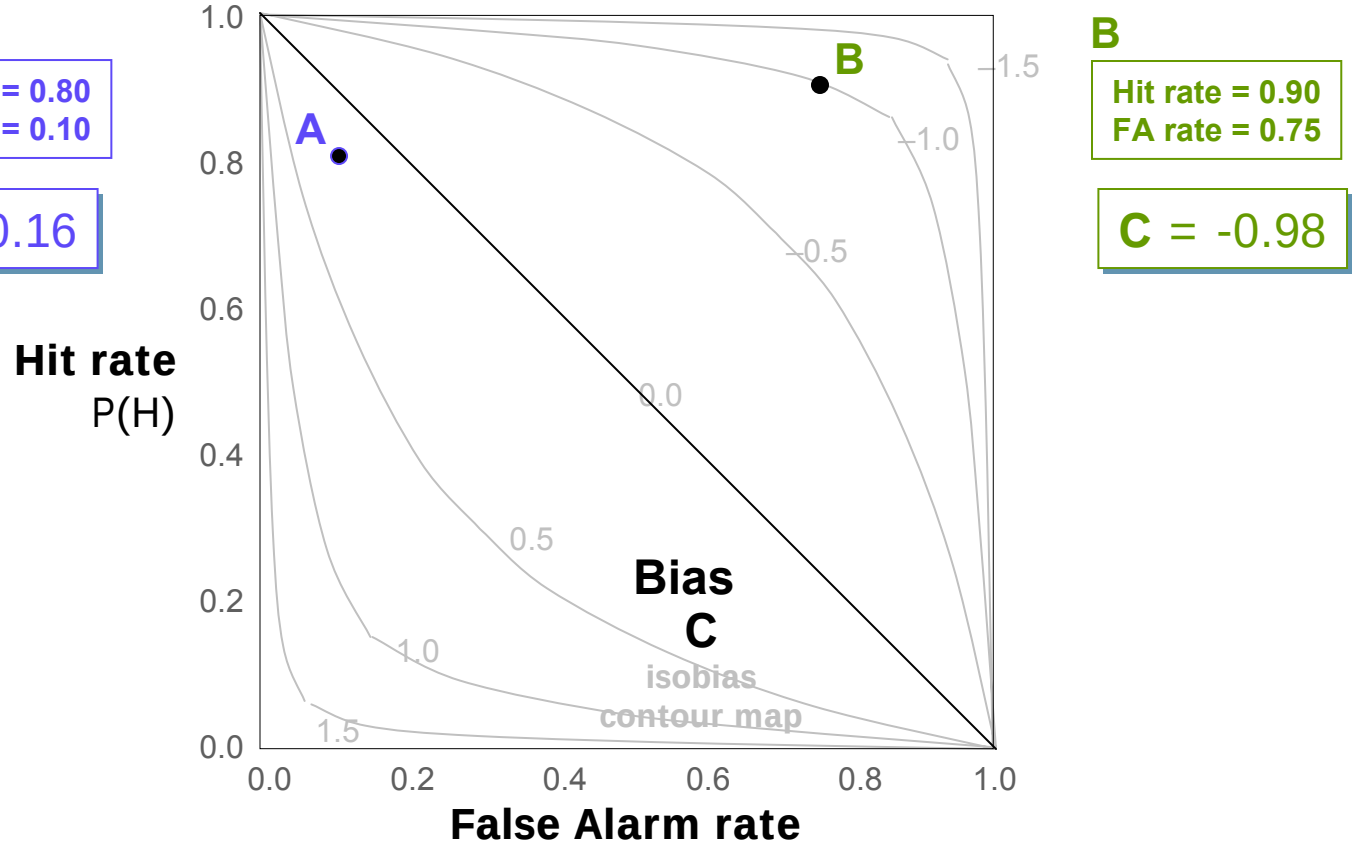
Hit rate = 0.80
FA rate = 0.10

C = 0.16

B

Hit rate = 0.90
FA rate = 0.75

C = -0.98



Subject A criterion = close to neutral

Subject B criterion = strong liberal bias

BAE SYSTEMS

ROC – Sensitivity

A

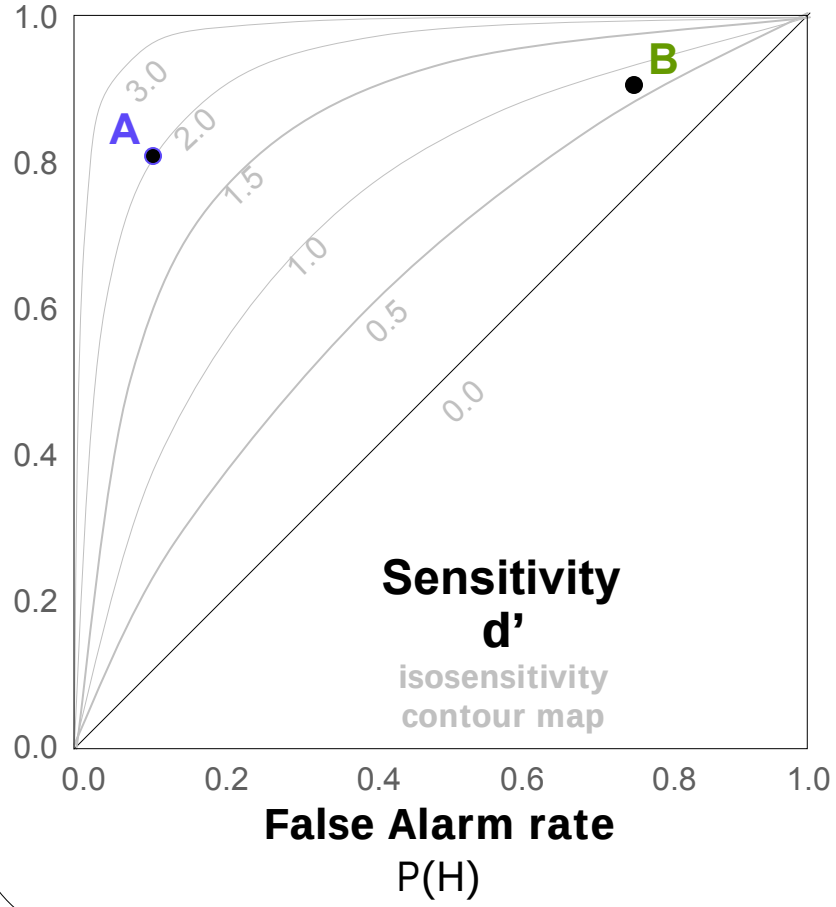
Hit rate = 0.80
FA rate = 0.10

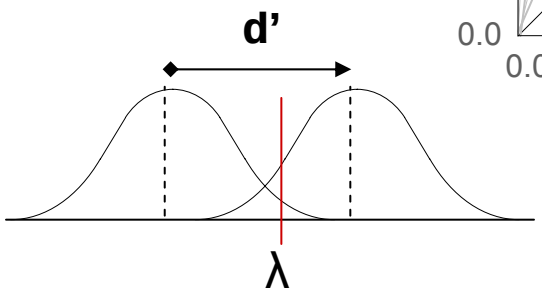
d' = 2.00

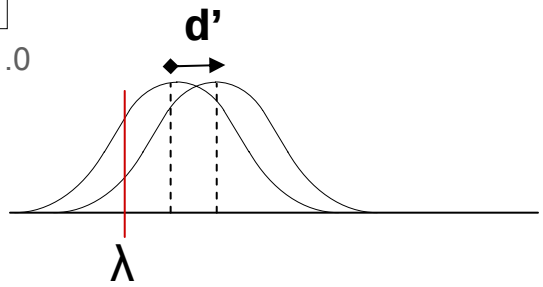
B

Hit rate = 0.90
FA rate = 0.75

d' = 0.60



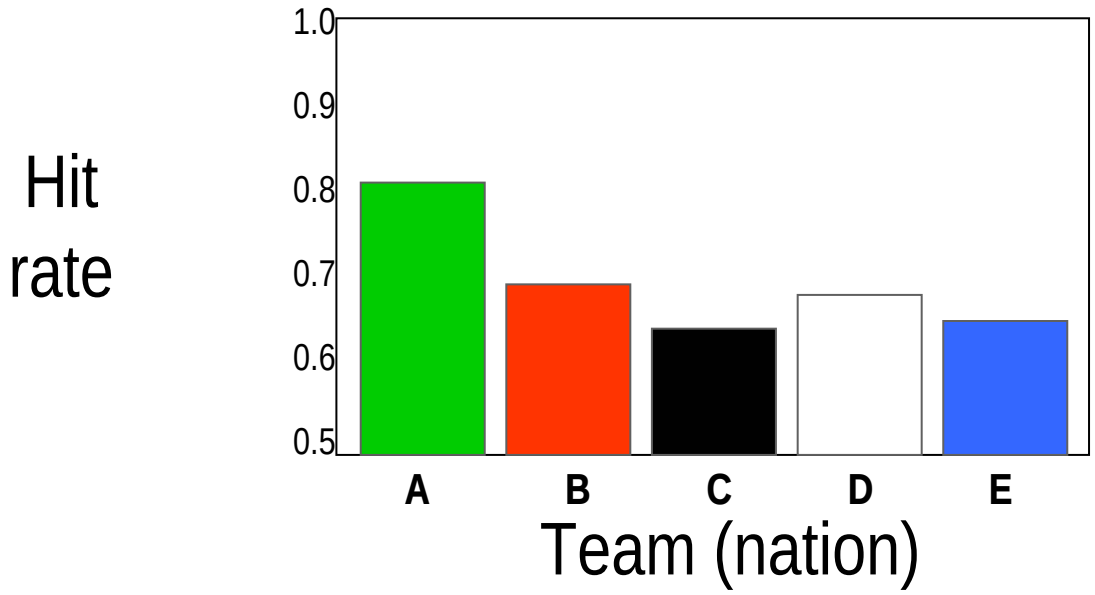




BAE SYSTEMS

QUASA data - LOE 2

SA probe hit rates

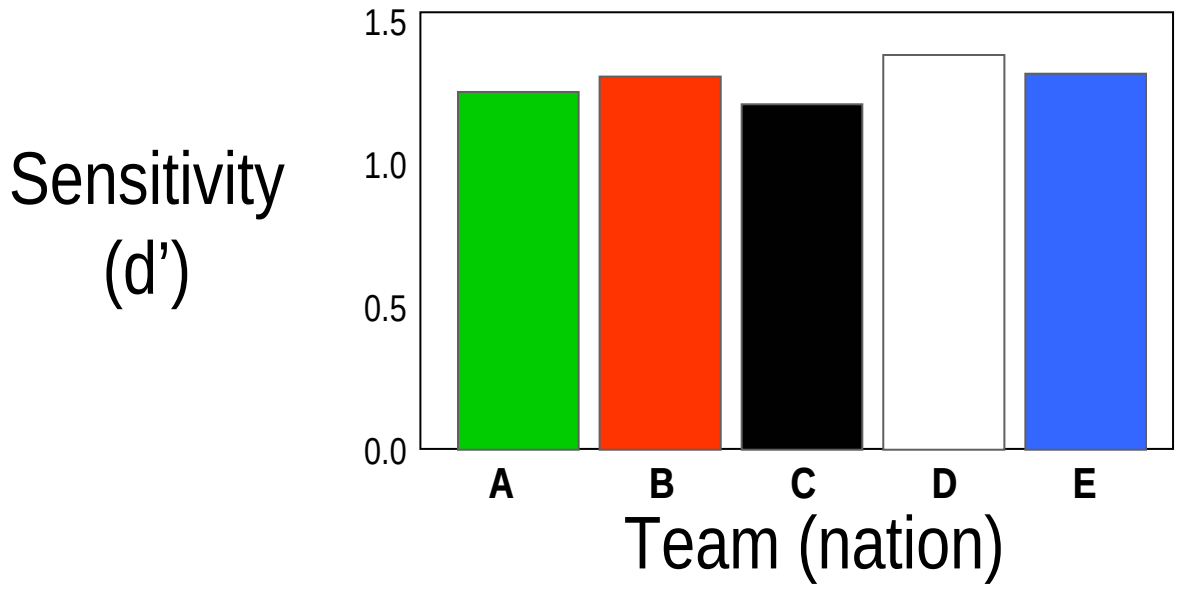


Team A has highest hit rate …

BAE SYSTEMS

QUASA data - LOE 2

SA probe sensitivity

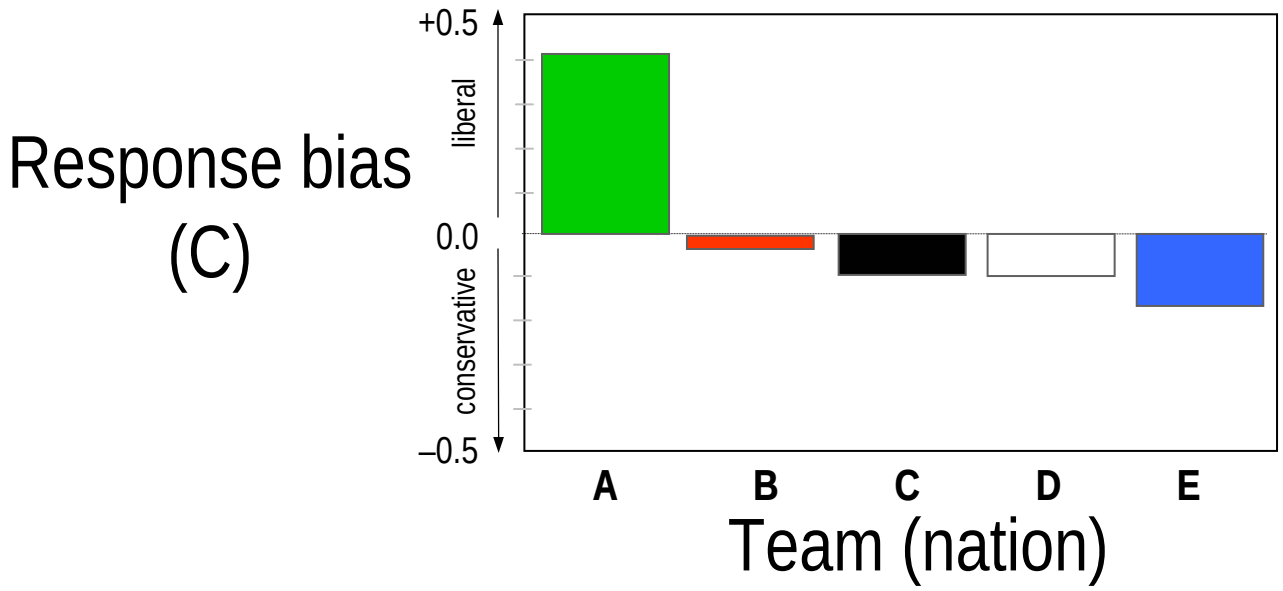


But team A is no more accurate overall at discriminating true from false probes

BAE SYSTEMS

QUASA data - LOE 2

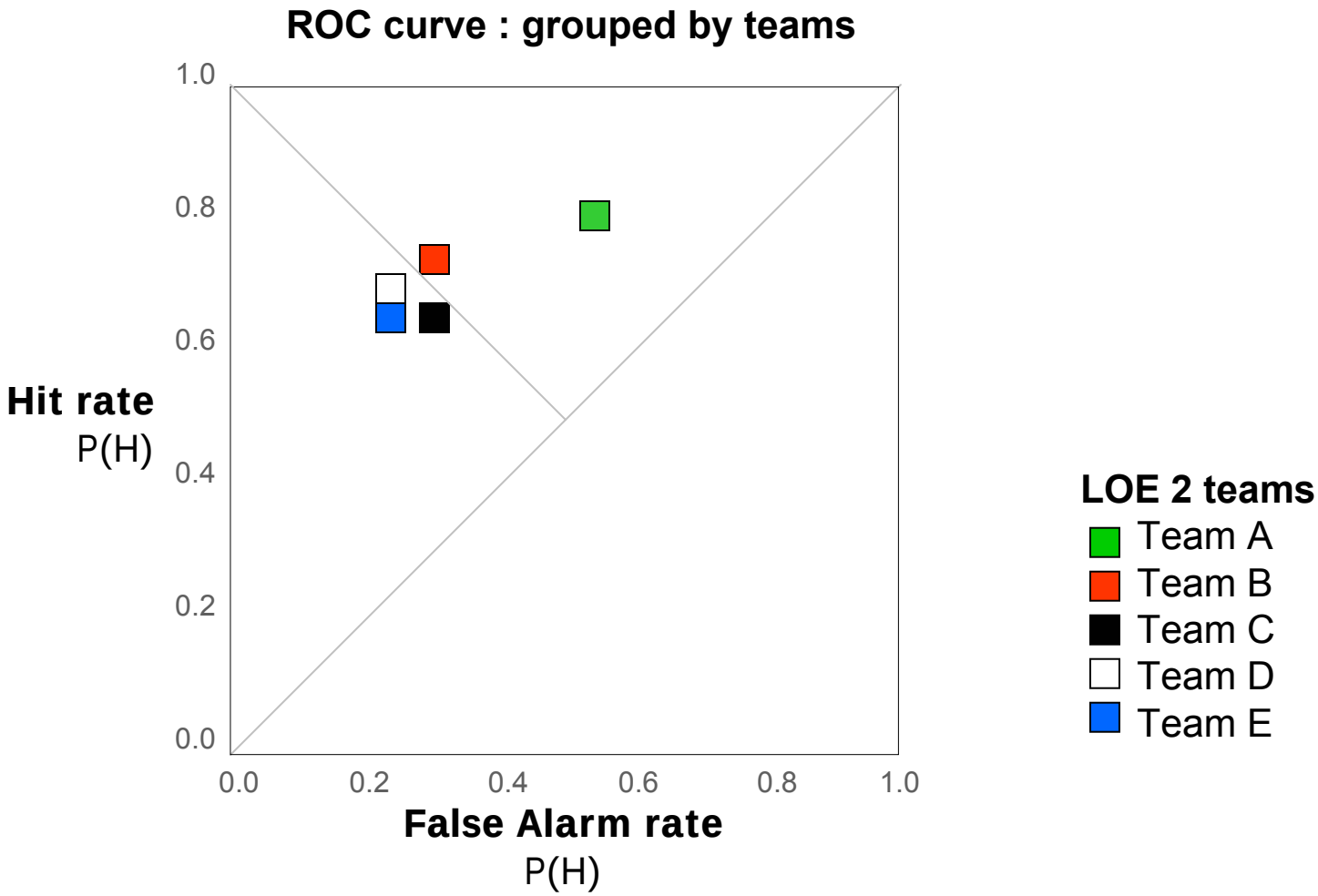
SA probe response bias



Team A is very liberal when uncertain (inclined to accept probes as true) -- hence the high hit rate

BAE SYSTEMS

QUASA data - LOE 2



BAE SYSTEMS

QUASA data - LOE 2

Summary so far

- Team A has highest hit rate on SA probes
- But SDT analysis shows all teams are only moderately accurate
- Team A's hit rate due to very liberal response bias when uncertain
- Other teams are neutral or slightly conservative

BAE SYSTEMS

Calibration

Concept

- Overconfidence / underconfidence
- The extent to which people are able to judge the correctness of their own observations or decisions

Method

- Obtain a judgement, then obtain self-rating of confidence in that judgement
  - binary ratings | continuous scales | ordinal ratings
- A well-calibrated person gives low ratings on incorrect / chance-level judgements (i.e. when uncertain) and high ratings on correct judgements (when certain)
- Calibration analysis quantifies this relationship in some way

BAE SYSTEMS

Calibration

Findings

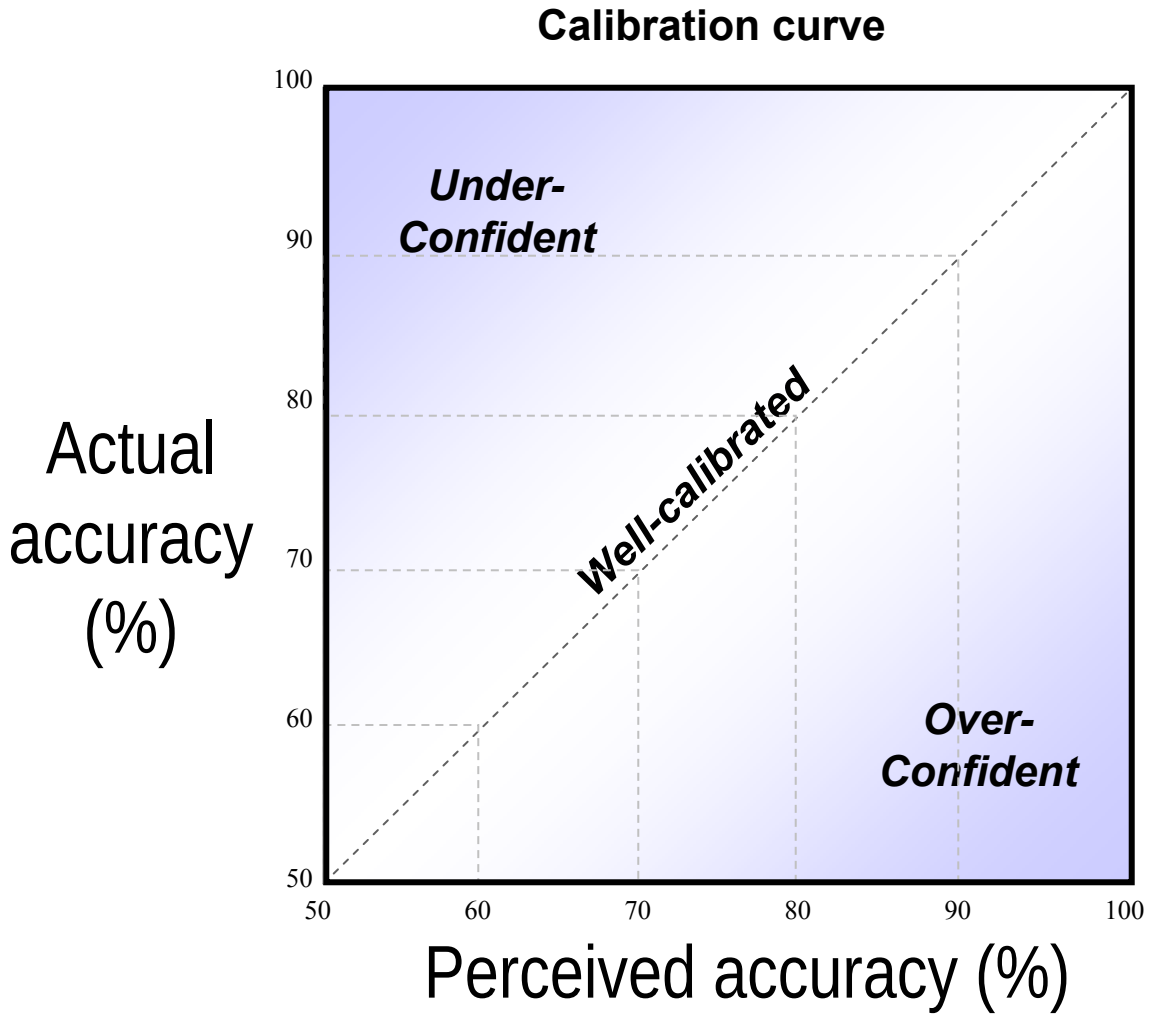
- Overconfidence common for cognitive tasks
- Underconfidence common for sensory tasks
- (May be an artefact of experimental methods)

Applications

- Eyewitness reports
  - Juries and police tend to be persuaded by highly confident witness reports, but these don't always correlkate with actual accuracy.
- Intelligence analysis
  - Don't want overconfident intelligence reports based on dubious data
- Situational awareness
  - Accidents attributed to over onfidence in poor/inaccurate SA

BAE SYSTEMS

Calibration



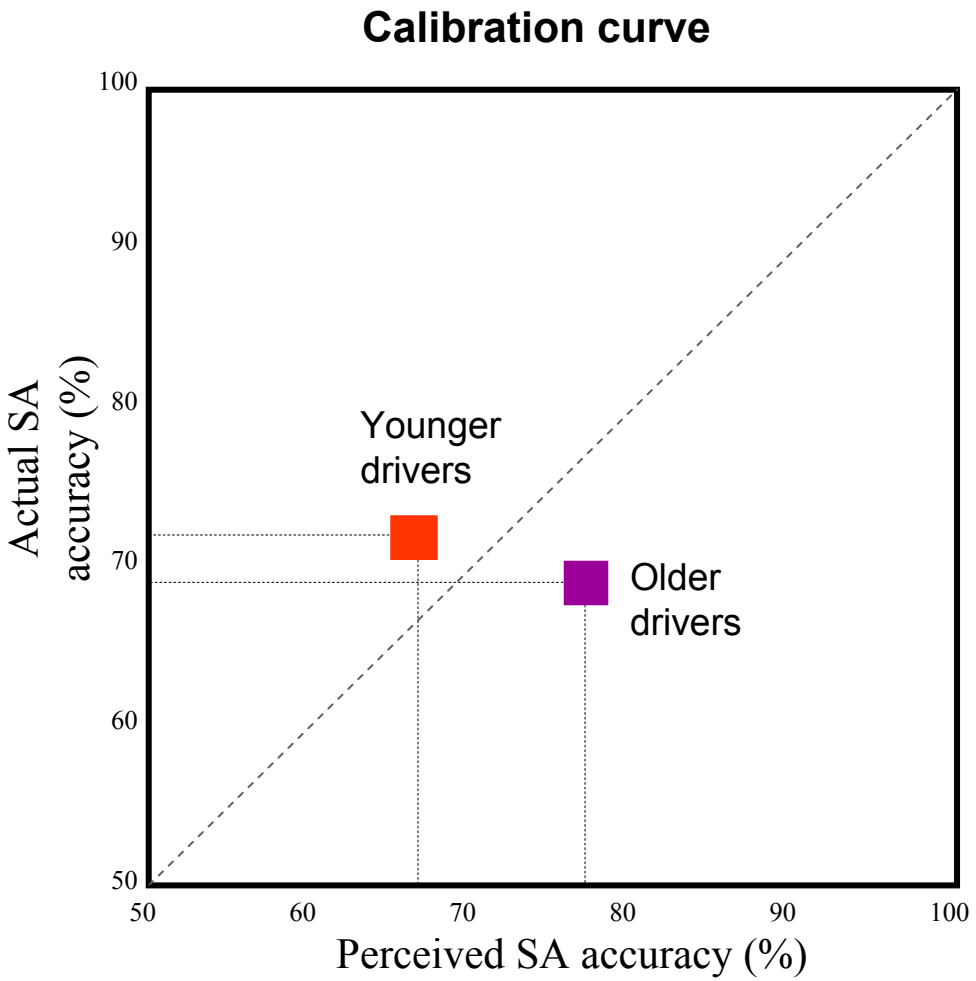
BAE SYSTEMS

Calibration

SA of car drivers presented with safety-related electronic messages by an Advanced Traveller Information System (ATIS).

SA measured using a 2AFC version of SAGAT.

Confidence in each probe response rated on a continuous scale (50%-100



Source

Lee, J.D., Stone, S., Gore, B.F., Colton, C., Macauley, J., Kinghorn, R., Campbell, J.L., Finch, M. & Jamieson, G. (1997).

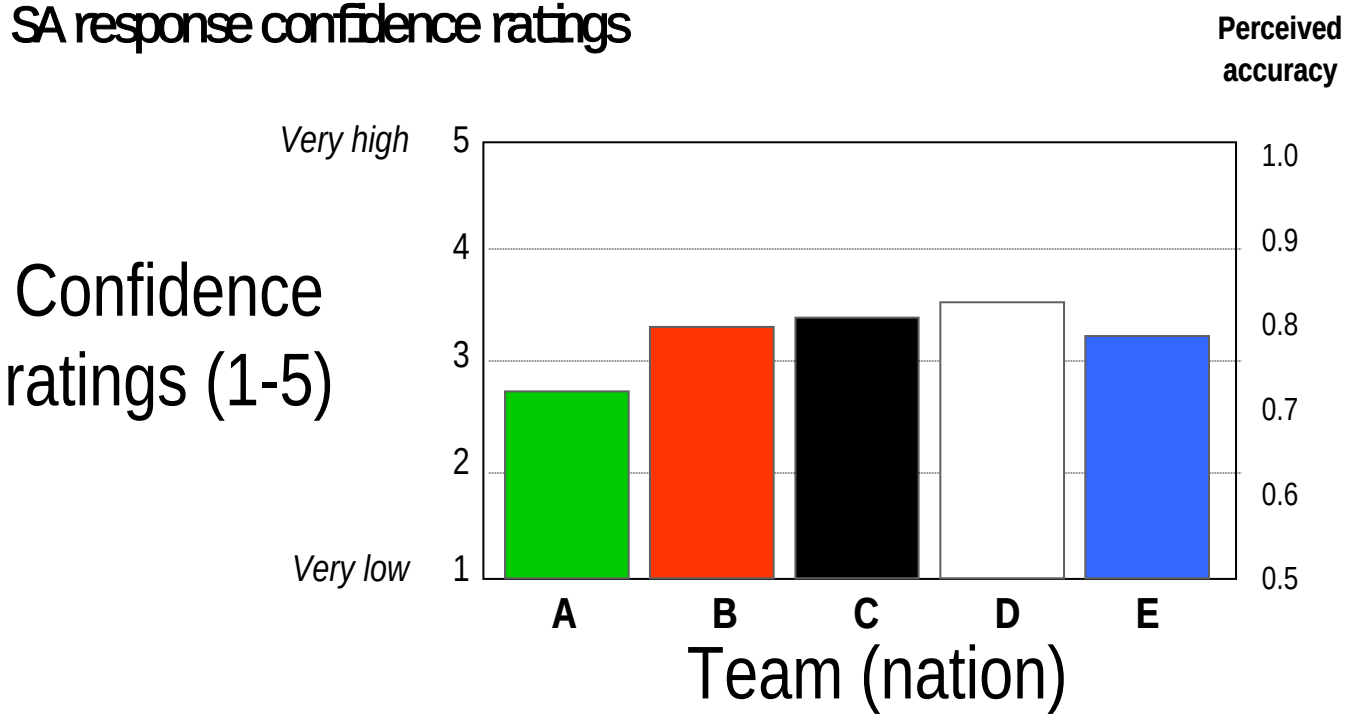
Advanced Traveller Information Systems and Commercial Vehicle Operations Components of the Intelligent Transportation Systems: Design Alternatives for In-Vehicle Information Displays.

U.S. Federal Highway Administration technical report FHWA-RD-96-147. McLean, Virginia.

BAE SYSTEMS

QUASA data - LOE 2

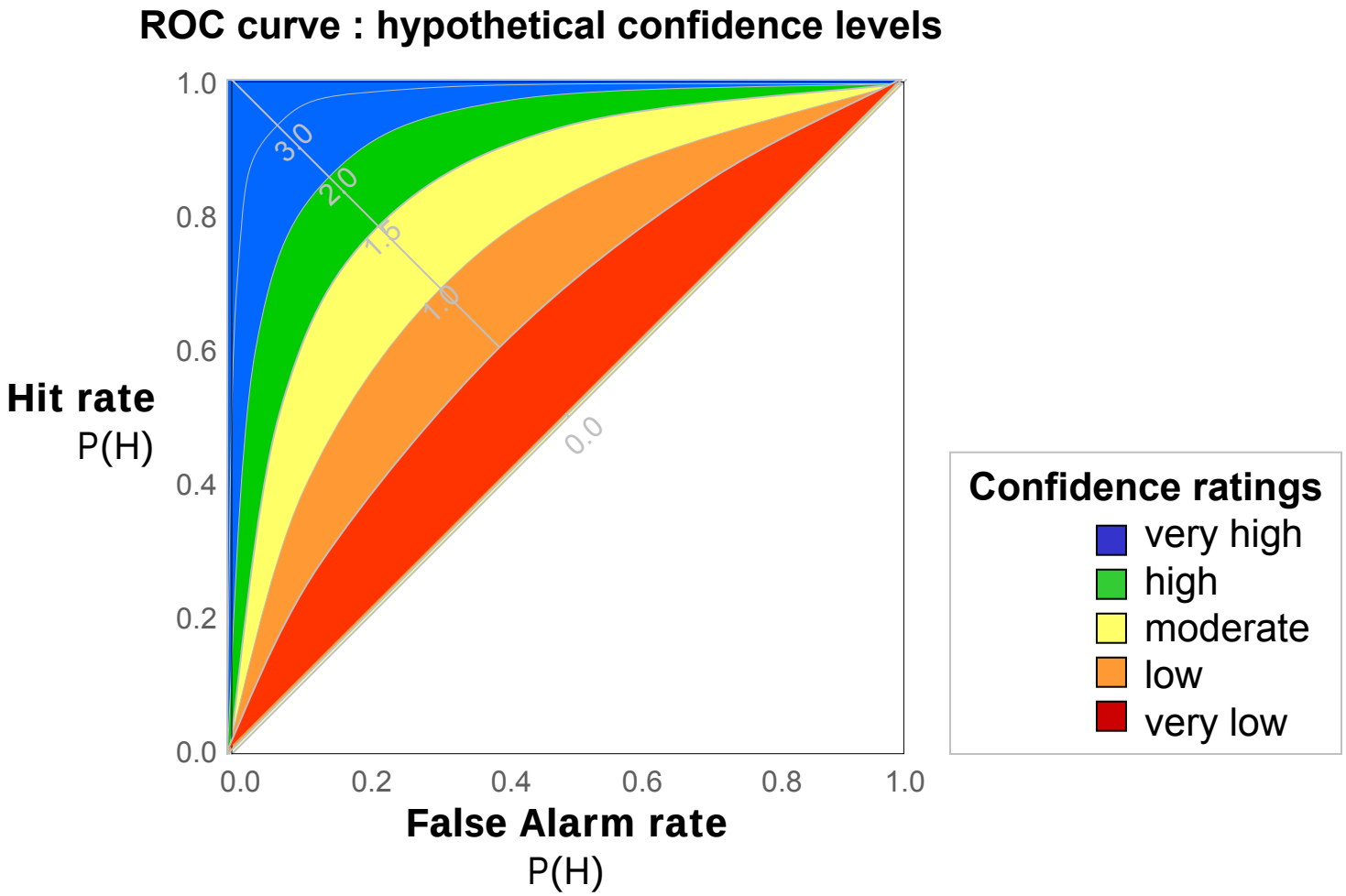
SA response confidence ratings



Mean SA probe response confidence ratings per team in LOE 2.

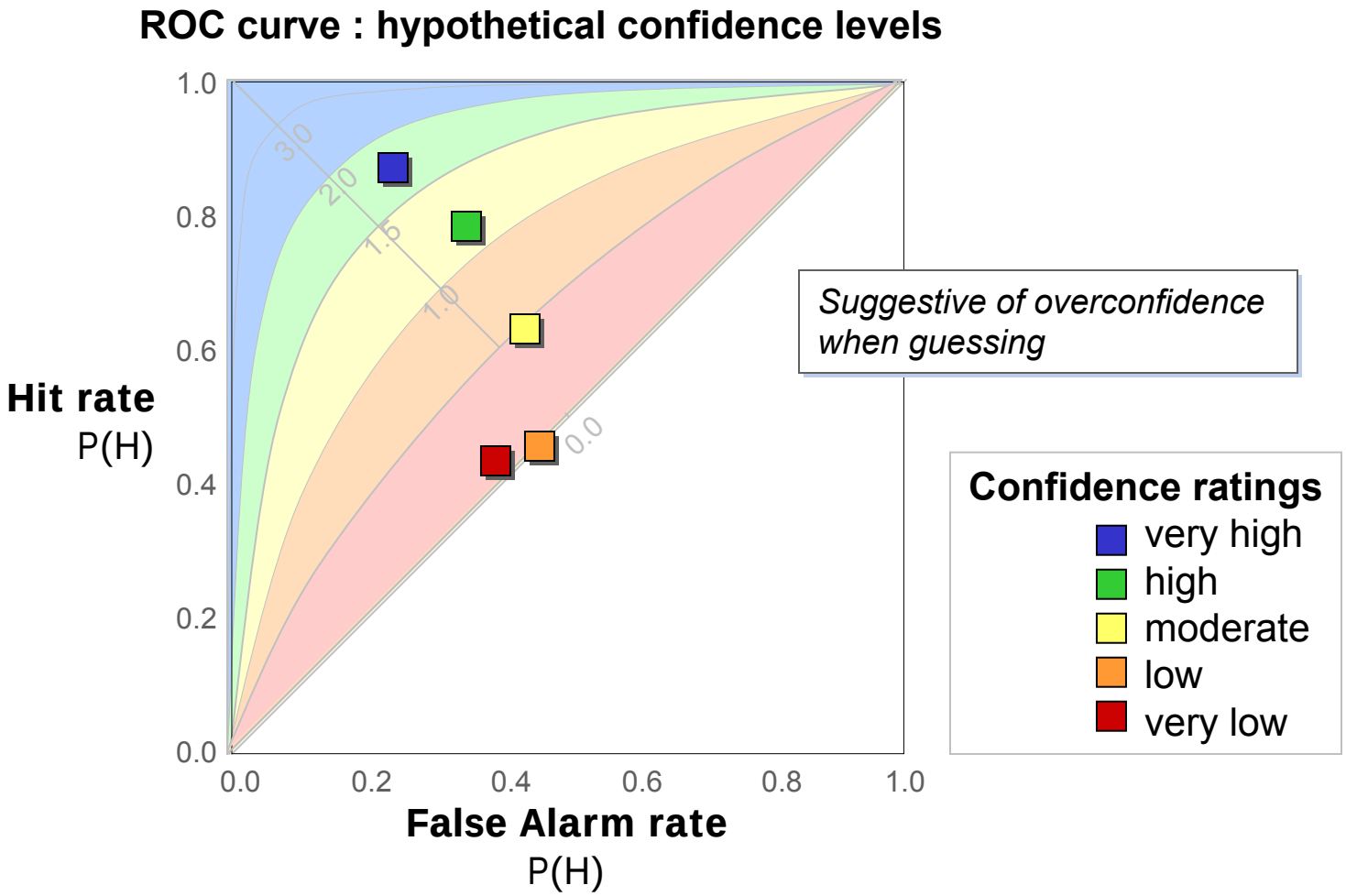
BAE SYSTEMS

QUASA data - LOE 2



BAE SYSTEMS

QUASA data - LOE 2



QUASA data - LOE 2

Calibration scores

– using hit + correct rejection rates as actual accuracy

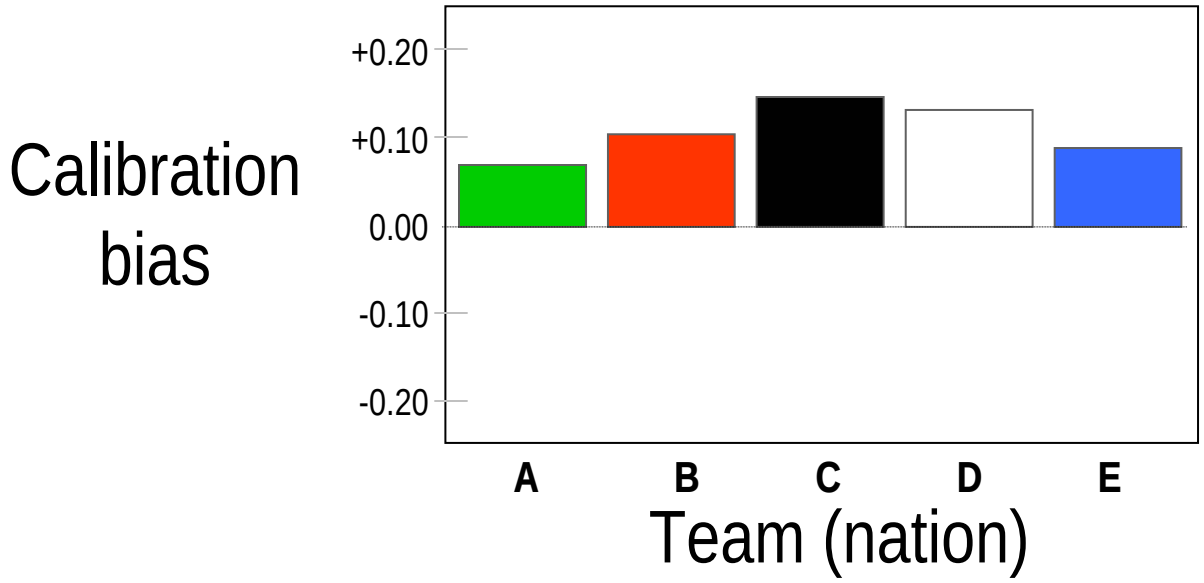
	Team (nation)				
	A	**B**	**C**	**D**	**E**
Perceived accuracy	0.716	0.795	0.803	0.832	0.774
SA accuracy (correct responses)	0.647	0.691	0.656	0.706	0.692
Calibration bias	**+0.07**	**+0.11**	**+0.15**	**+0.13**	**+0.08**

To assess SA calibration, average confidence ratings were transformed (0.5-1.0) and probe accuracy scores (proportion of hits plus correct rejections) were subtracted from the result to provide a calibration bias statistic.

BAE SYSTEMS

QUASA data - LOE 2

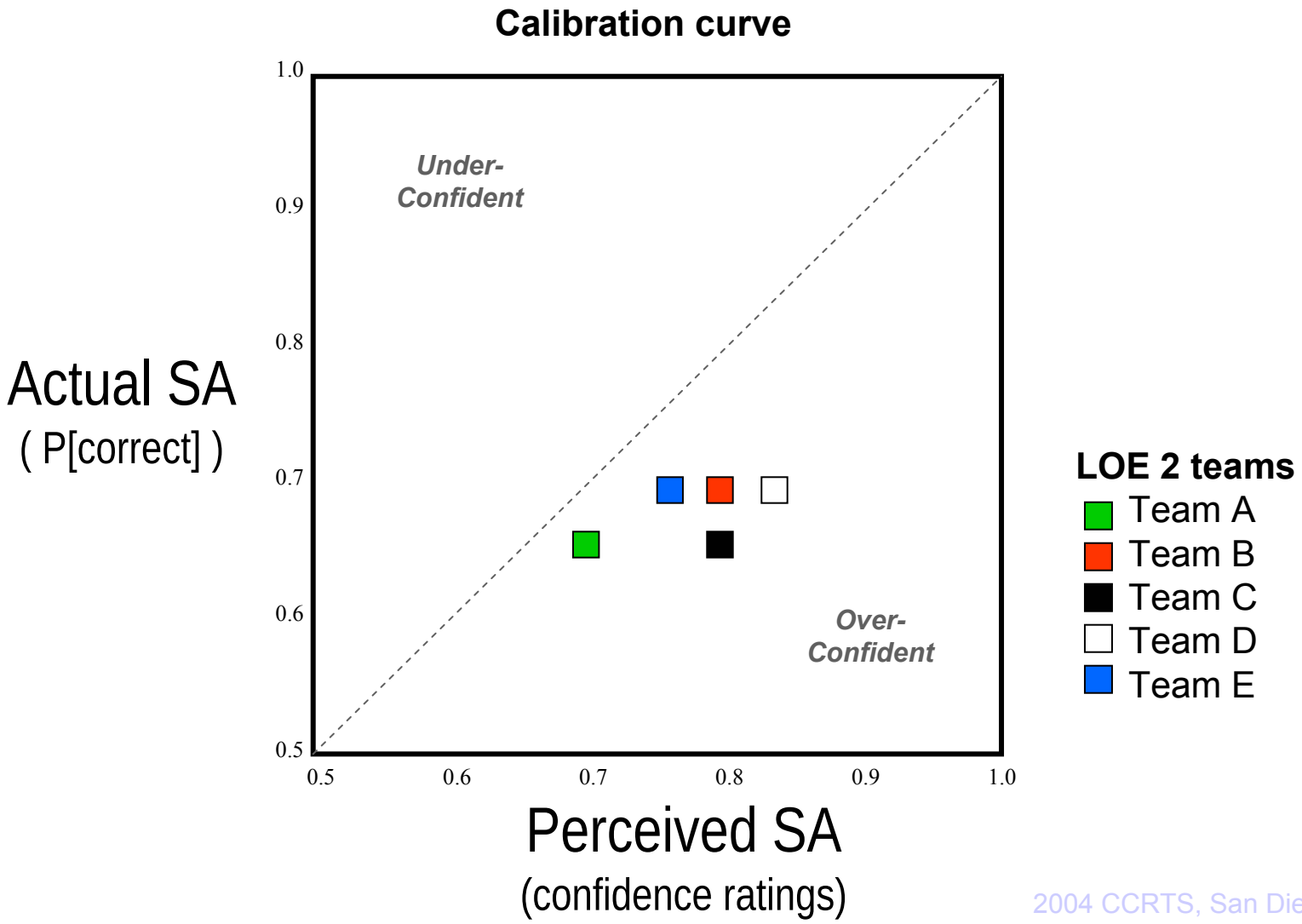
Calibration scores



Mean SA probe hit rates per team in LOE 2.

BAE SYSTEMS

QUASA data - LOE 2



BAE SYSTEMS

QUASA data - LOE 2

Summary

- Team A had lowest overall confidence ratings in their SA responses
- Confidence ratings were transformed into “perceived SA” scores and calibrated with actual SA scores
- Calibration analysis revealed general overconfidence
- Team A was actually best calibrated

BAE SYSTEMS

Summary & conclusions

QUASA

- Technique for SA assessment
- Combines true/false SA probes with simultaneous self-ratings of confidence for each probe response.
- SDT analysis is applied to probe responses
  - Differentiates between actual SA accuracy (sensitivity) and response bias when uncertain
- Calibration analysis examines the relationship between actual SA and perceived SA.

Conclusions

- **QUASA yields potentially insightful quantitative results**
- **SDT statistic can be used as measure of actual SA accuracy.**
- **Subjects appear to be generally well-calibrated for SA**

BAE SYSTEMS

Lessons learned

- **T/F probes need objective referent (‘groud truth’)**
  - Can be used to assess awareness of empirical information (objective environment & features, type of situation, actions)
  - Cannot be used to assess awareness of non-empirical information (future possibilities, intentions)
- **T/F probes need very careful construction & pre-testing**
  - Avoid ambiguity in language
  - Avoid bias in likelihood
- **In a dynamic situation, T/F probes may need to be constructed on the fly**

BAE SYSTEMS

Outstanding issues

- Does response criterion/bias obtained with probes reflect a similar criterion/bias of the subject in assessing the real situation?
- How many probes / responses needed?
- How does this compare with other metrics?
- What about time to respond to probe? (= distance from criterion?)

BAE SYSTEMS

Research directions

- **Perform calibration analysis with Fuzzy SDT and/or Type 2 SDT**
- **Address team / shared SA**

BAE SYSTEMS

Quantitative Analysis of Situational Awareness (QUASA)

Applying Signal Detection Theory to True/False Probes and Self-Ratings

Barry McGuinness

Principal Scientist
Human Factors Dept
Advanced Technology Centre
BAE Systems
Bristol, UK

barry.mcguinness @ baesystems.com

BAE SYSTEMS

BACKUP SLIDES

BAE SYSTEMS

Characteristics of SA

- Mode of cognition that facilitates effective action
  - Critical in situations that are potentially complex, demanding, high-tempo, uncertain and/or unpredictable.
- Consists of mental representations of a situation and its implications:

- **OBJECTIVE AWARENESS :**

  The operational environment and the constellation of elements within it
  - terrain, weather, buildings, platforms, people; locations, movements, actions, states
  - derived from observations or data in context

- **SITUATIONAL UNDERSTANDING :**

  The global characteristics of the situation -- type and status
  - *Hijack situation? Hostage situation? Safe? Problematic? Critical?*
  - inferred from current awareness in context

- **OPERATIONAL APPRECIATION :**

  The implications of the situation w.r.t. one's operational goals / plans / tasks
  - Getting better or worse? Critical points ahead? Need a new course of action?
  - inferred from situational understanding in context

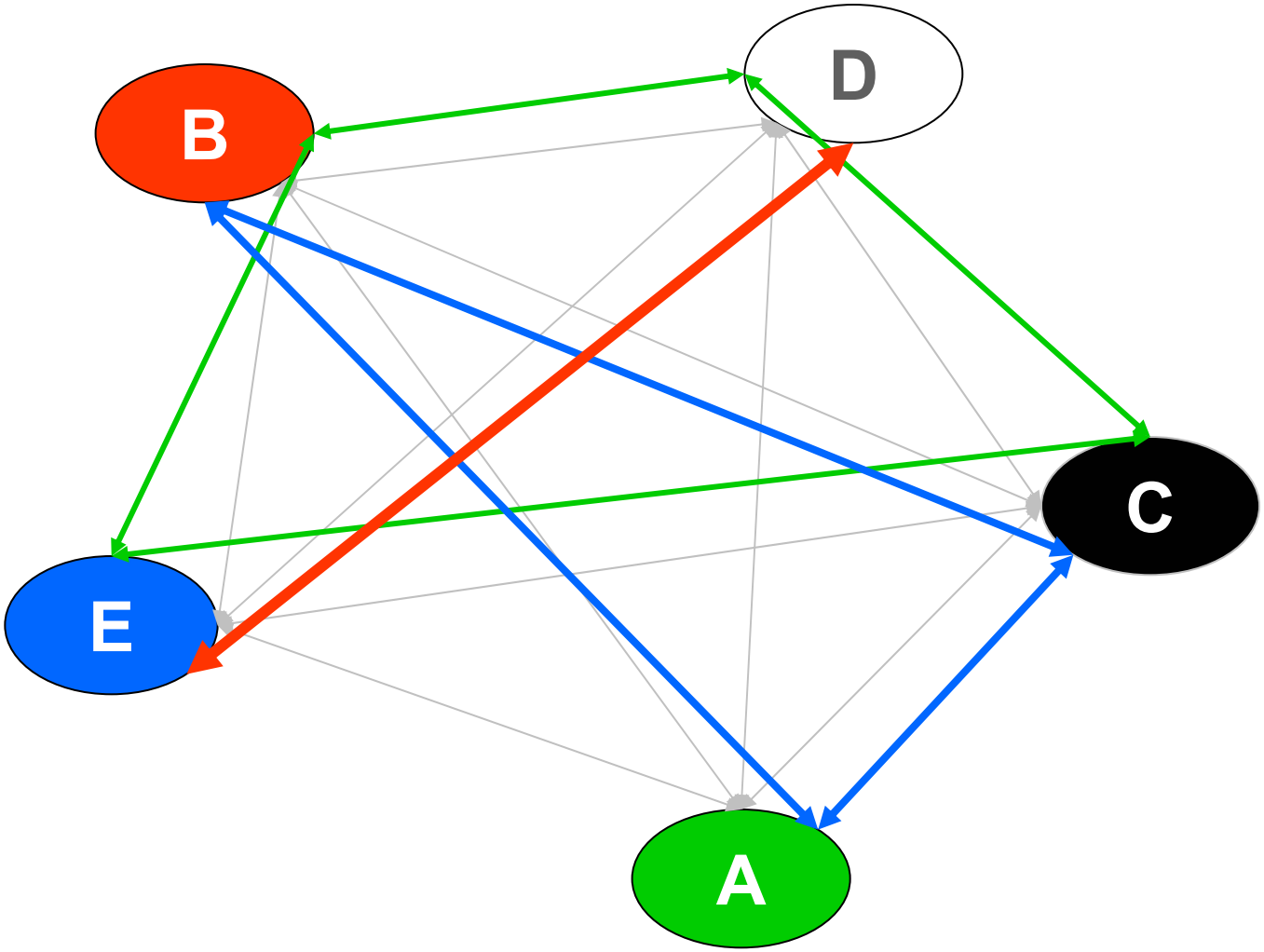
BAE SYSTEMS

LOE 2 information sharing agreements

Country	ML	TL	BL_1	BL_2	Coalition	Private	Total
A		X		X	X	X	4
C	X	X			X	X	4
B	X	X			X	X	4
D	X		X		X	X	4
E	X		X		X	X	4

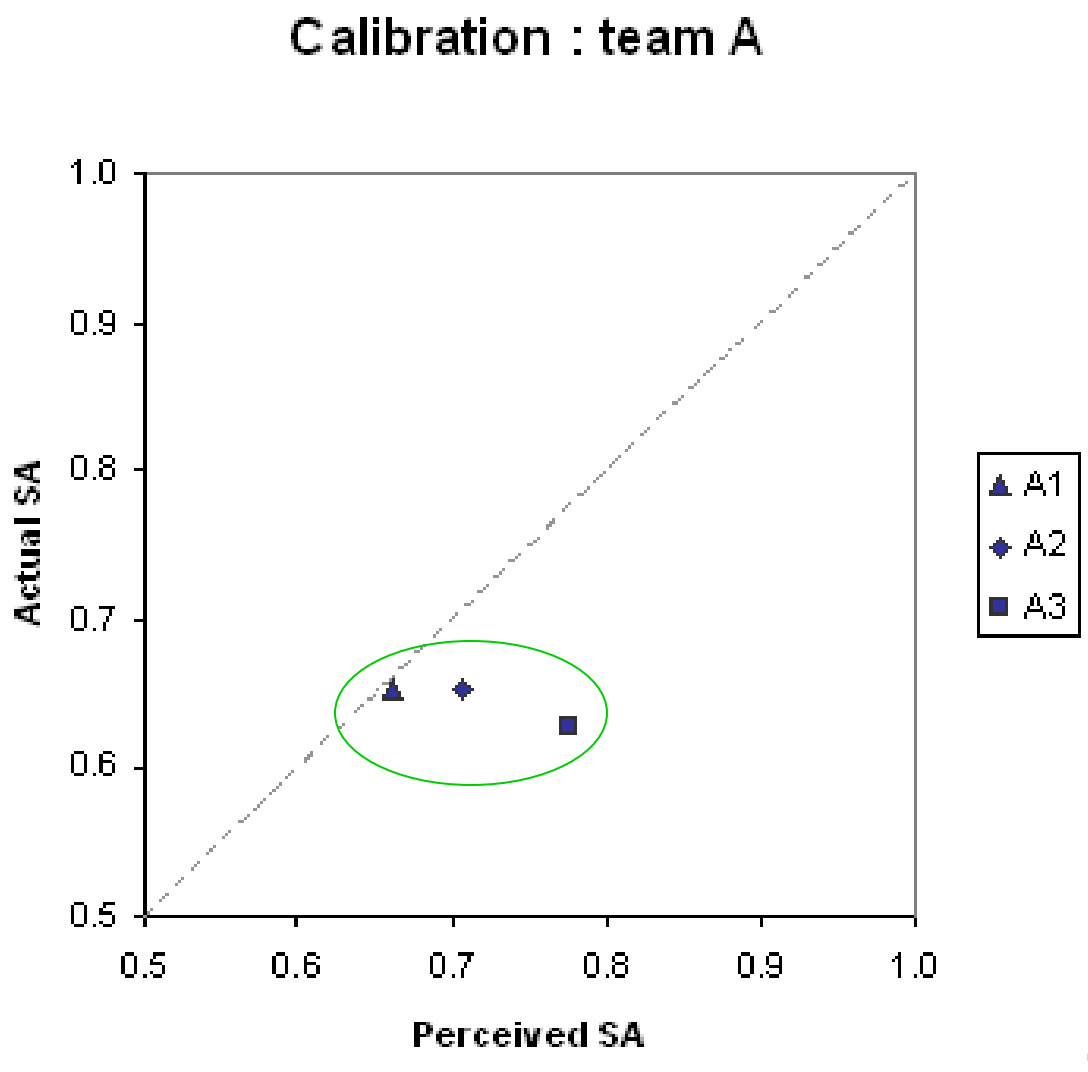
BAE SYSTEMS

LOE 2 information sharing agreements



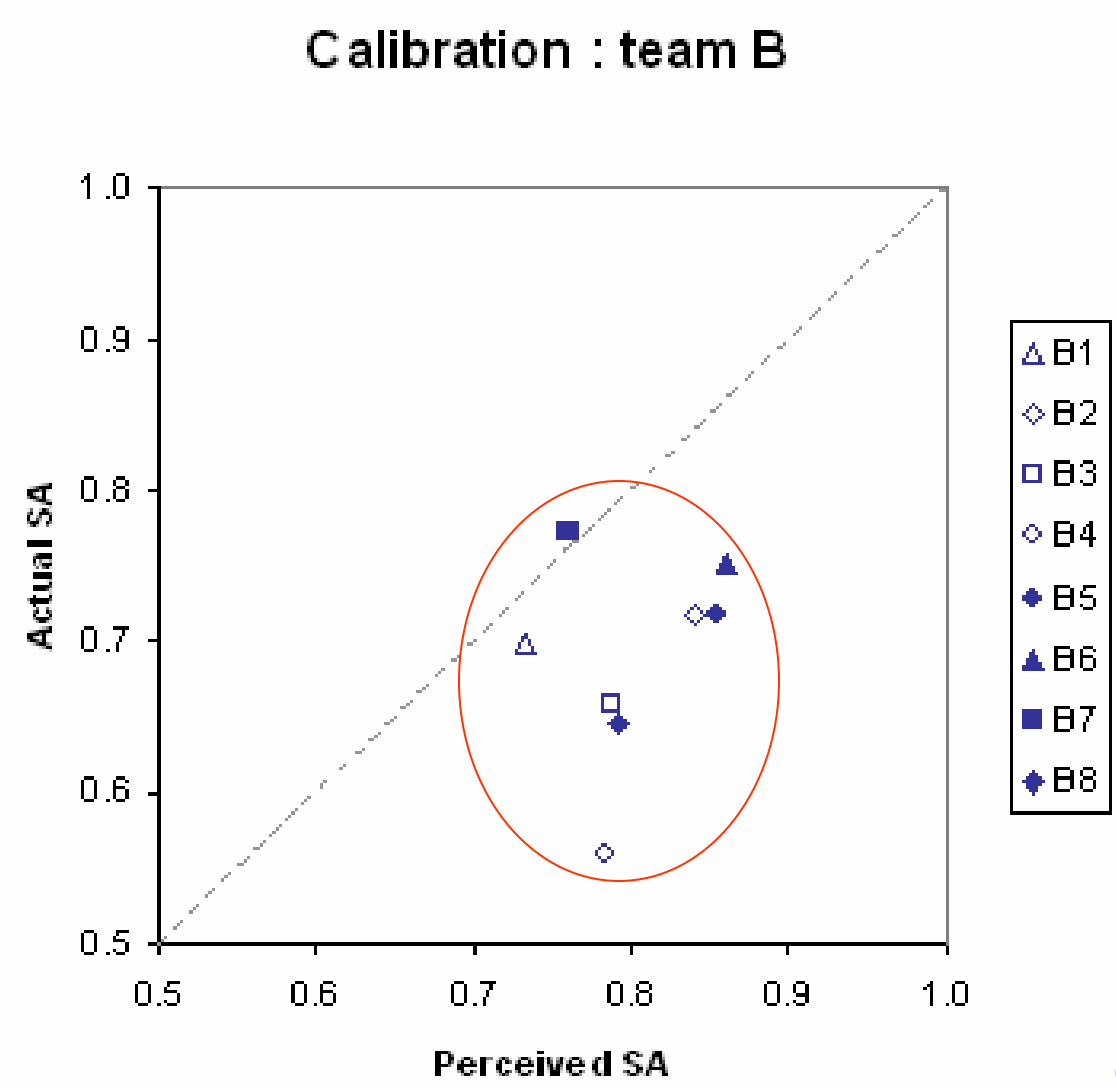
BAE SYSTEMS

QUASA data - LOE 2



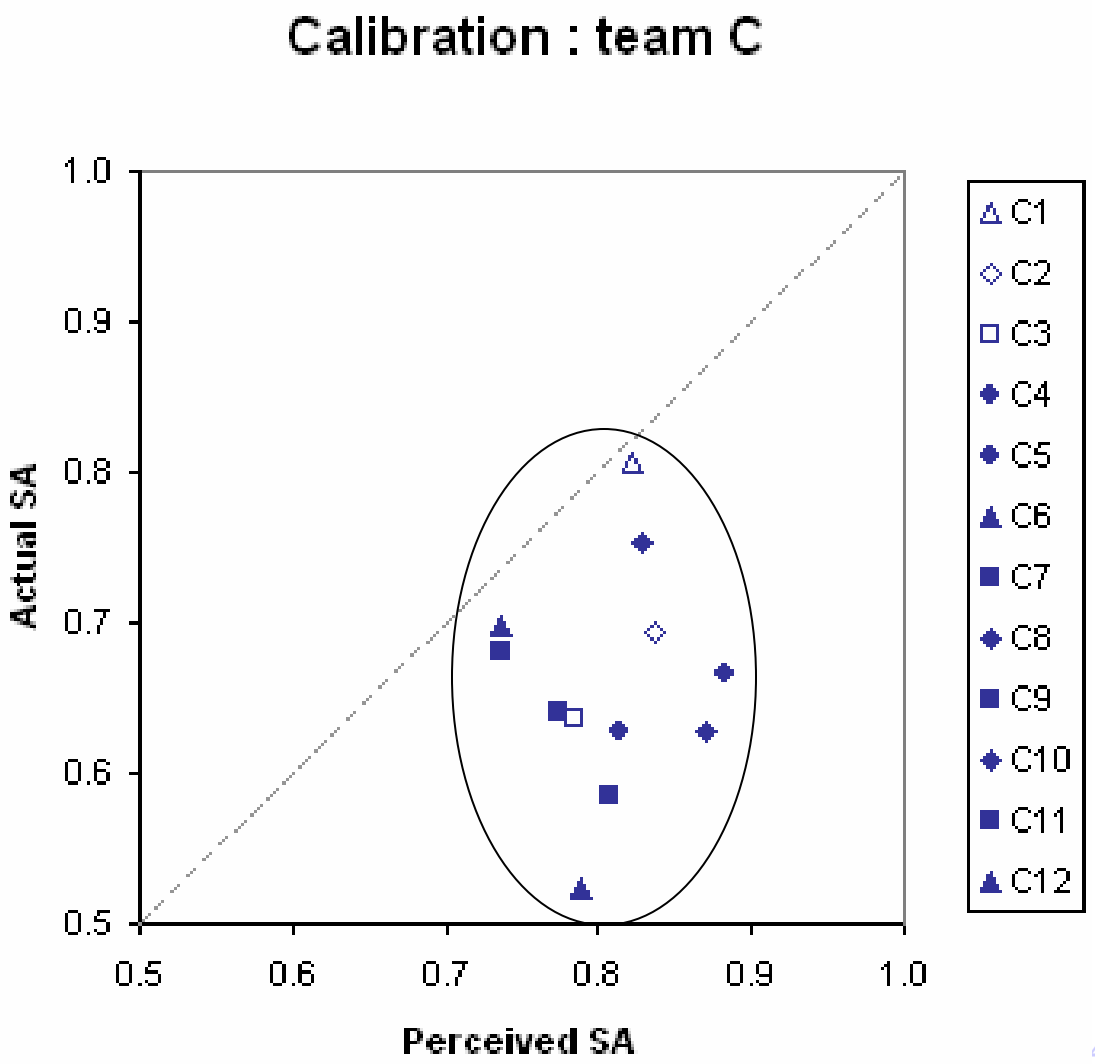
BAE SYSTEMS

QUASA data - LOE 2



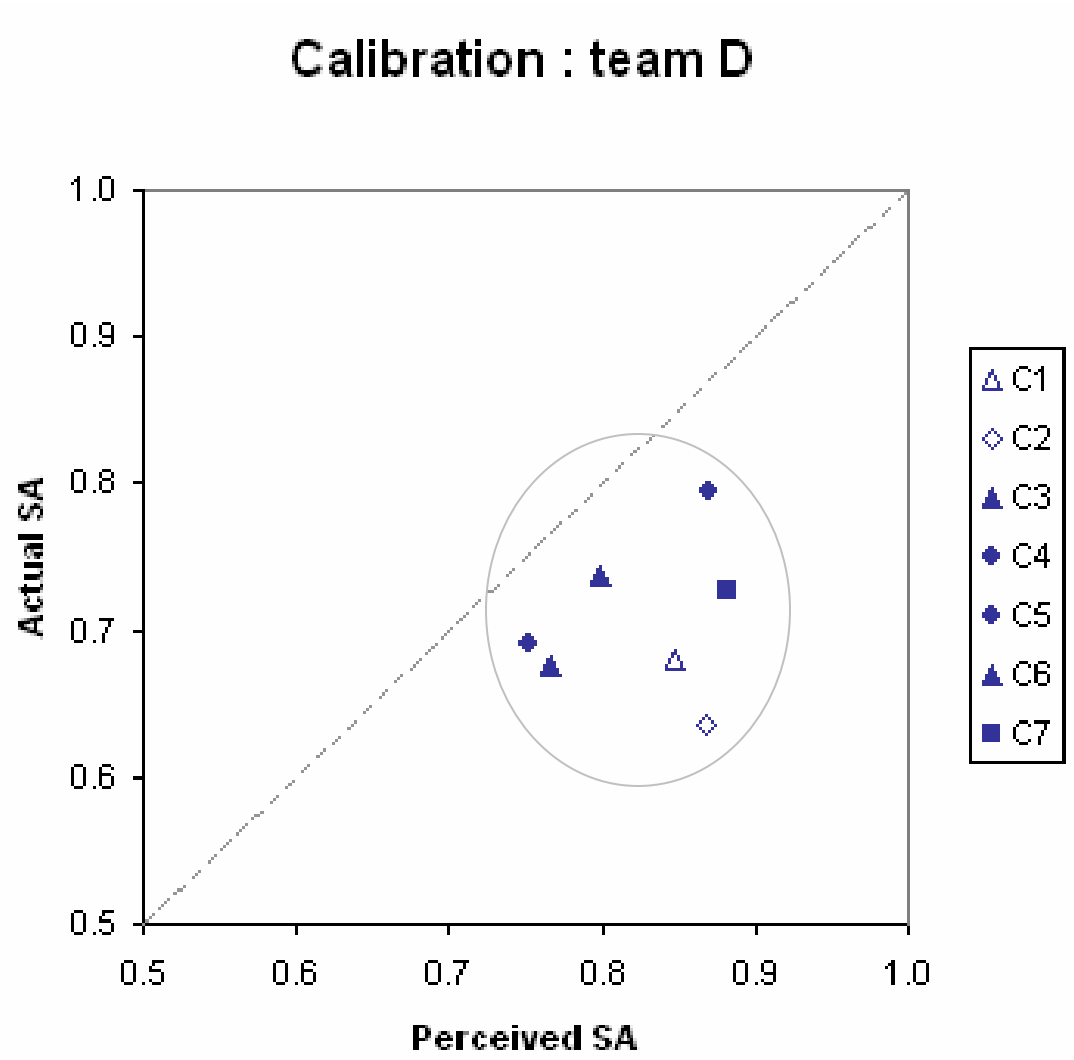
BAE SYSTEMS

QUASA data - LOE 2



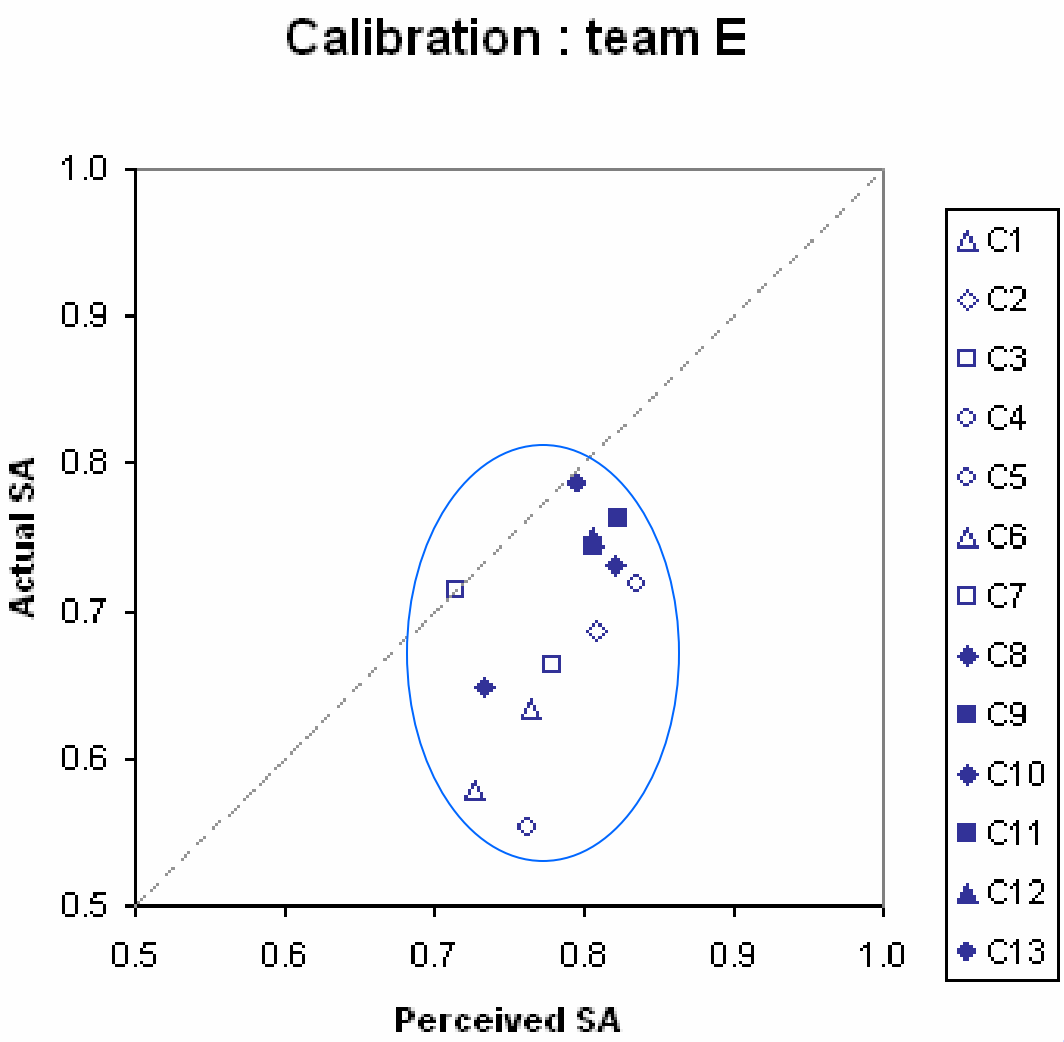
BAE SYSTEMS

QUASA data - LOE 2



BAE SYSTEMS

QUASA data - LOE 2



BAE SYSTEMS

QUASA data - LOE 2

Calibration scores

– using A' as actual accuracy

	Team (nation)				
	A	**B**	**C**	**D**	**E**
Perceived accuracy	0.716	0.795	0.803	0.832	0.774
SA accuracy (correct responses)	0.647	0.691	0.656	0.706	0.692
SA accuracy (A' score)	0.744	0.776	0.737	0.792	0.778
Calibration bias	**- 0.03**	**+0.03**	**+0.07**	**+0.03**	**+0.01**

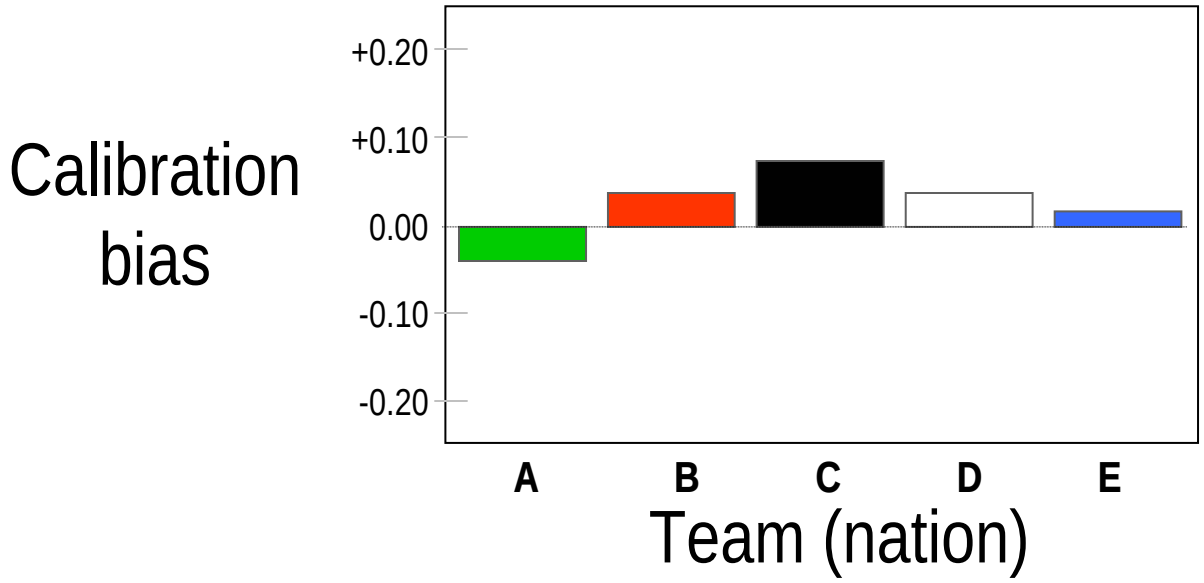
To assess SA calibration, average confidence ratings were transformed (0.5-1.0) and probe accuracy scores (A', a measure of sensitivity) were subtracted from the result to provide a calibration bias statistic.

BAE SYSTEMS

QUASA data - LOE 2

Calibration scores

– using A' as actual accuracy



Mean SA probe hit rates per team in LOE 2.

BAE SYSTEMS

QUASA data - LOE 2