Quantitative Analysis of Situation Awareness (QASA): Modelling and Measuring Situation Awareness using Signal Detection Theory

Graham K. Edgar[1], Di Catherwood[1], Steven Baker[1], Geoff Sallis[1],
Michael Bertels[2], Helen E. Edgar[1], Dritan Nikolla[1], Susanna Buckle[1], Charlotte Goodwin[1] and Allana Whelan[1]

[1]Centre for Research in Applied Cognition, Knowledge Learning and Emotion (CRACKLE),
Department of Natural and Social Sciences,
University of Gloucestershire, UK

[2]IFV Brandweeracademie, Netherlands

Correspondence concerning this article should be addressed to Graham Edgar, Department of Natural and Social Sciences, University of Gloucestershire, Francis Close Hall, Swindon Road, Cheltenham, Gloucestershire, GL50 4AZ, UK. E-mail: gedgar@glos.ac.uk.

Abstract

This paper presents a model of situation awareness (SA) that emphasises that SA is necessarily built using a subset of available information. A technique (Quantitative Analysis of Situation Awareness – QASA), based around signal detection theory, has been developed from this model that provides separate measures of actual SA (ASA) and perceived SA (PSA), together with a feature unique to QASA, a measure of bias (information-acceptance). These measures allow the exploration of the relationship between actual SA, perceived SA, and information acceptance. QASA can also be used for the measurement of dynamic ASA, PSA and bias. Example studies are presented and full details of the implementation of the QASA technique are provided.

Keywords

Dynamic situation awareness, situational awareness, signal detection theory, bias, confidence, team situation awareness, situated situation awareness

Practitioner summary

This paper presents a new model of situation awareness (SA) together with an associated tool (Quantitative Analysis of Situation Awareness – QASA) that employs signal detection theory to measure several aspects of SA, including actual and perceived SA and information acceptance. Full details are given of the implementation of the tool.

Acknowledgements

The authors are very grateful to anonymous reviewers for their unstinting efforts in reviewing this paper and suggesting improvements. The authors are also grateful to all those participants that gave their time freely to participate in the studies described here.

Introduction

In the field of human factors and ergonomics, the concept of situation awareness (SA) has generated a large amount of both research and (often) heated debate. It is beyond the scope of this paper to revisit these debates in depth, particularly given the number of reviews in the area (e.g., Banbury & Tremblay, 2004; Endsley, 2015; Endsley & Garland, 2000; Rousseau, Tremblay, & Breton, 2004; Salmon & Stanton, 2013; Salmon et al., 2008; Stanton, 2010; Stanton, Salmon, Walker, Salas, & Hancock, 2017), and so this paper will outline the existing theory only so far as is necessary to support the model of SA proposed here.

Models of SA tend to consider SA within three levels-of-analysis (Sorensen, Stanton, & Banks, 2011; Stanton et al., 2017). The first level-of-analysis considers SA to be a psychological phenomenon maintained by the cognitive processes (particularly working memory) of the individual. While cognitive processes operate within the environment, SA may be treated as the end product used in decision-making.

The second level-of-analysis is to consider SA as residing in the environment (e.g. within displays and other sources of information, that could include everything from other team members and complex computerised systems, to marks on the ground) and tends to emphasise the importance of information technology for generating SA. As Woods and Sarter (2010) suggest, deploying some new piece of technology will almost inevitably lead to an improvement (or at least a change) in SA.

The final, and probably most practically useful conceptualization of SA, is to take a systems-ergonomics approach and to consider SA to be an emergent property of an individual's *interaction* with their environment (Salmon, Stanton, & Young, 2012; Smith & Hancock, 1995; Stanton, Salmon, Walker, & Jenkins, 2010; Stanton et al., 2006). The importance of this interaction is embodied in the notion of 'distributed cognition' that conceptualizes cognition as a process that must involve interaction with the external world (e.g. Anderson, 2003; Clark, 1997, 2008; Hutchins, 1995; Varela, Thompson, & Rosch, 1991). The external world is used as an extension of cognitive systems and can be used to manipulate and store information; thus lightening the load on human cognition. Distributed situation awareness (DSA) then becomes, '*…adaptive, externally directed consciousness,*' (Smith & Hancock, 1995, p. 138); '*…a dynamic concept that exists at the interface between the agent and its environment,*' (Smith & Hancock, 1995, p. 139).

A variant of DSA is 'situated SA' (SSA) (Chiappe, Rorie, Morgan, & Vu, 2014; Chiappe, Strybel, & Vu, 2011; Chiappe, Vu, & Strybel, 2012). SSA focuses on the individual within the system, but suggests that SA is not maintained entirely 'within' the individual but rather is a dynamic exchange of information between the individual and external systems (e.g. information displays, data storage, other people). In this case, the problem for the individual becomes one of where to focus their attention and information-gathering efforts (Woods & Sarter, 2010). In contrast to the DSA approach, SSA does not need to ascribe cognitive properties to physical artefacts in the systems.

Situated SA is congruent with schema theory (Bartlett, 1932) as developed by Neisser (1976) who proposed the idea of genotype and phenotype schemata. Genotype schemata are cognitive phenomena, developed over time and across many situations. Phenotype schemata are the manifestation of genotype schemata in a particular situation and, in this respect, share some of the features of mental models (for a consideration of the similarities and differences between schemata and mental models see, Plant & Stanton, 2013; Richardson & Ball, 2009) and have sometimes been used interchangeably (e.g. Endsley, 2000). Behaviour in any particular situation results from the cyclical interaction of information from the situation with the genotype schemata in operation at that time.

Tenney, Adams, Pew, Huggins, and Rogers (1992, p. 3) used the cyclical model proposed by Neisser as a framework for understanding SA suggesting that, '*The information that the observer picks up from the environment, or samples from the available information, in turn modifies, or updates, what the perceiver knows about the immediate surroundings and influences what is known about the world in general.*'

We have used the concepts outlined above to develop a cyclical model of SA that is shown graphically in Figure 1 and that underpins the measurement approach outlined in this paper (Quantitative Analysis of Situation Awareness – QASA). Knowledge and experience, built up over time, is held in genotype schemata. The phenotype schema relating to a particular situation is built using information from those genotype schemata, combined with information from the situation. Not all information available from either genotype schemata, or the situation, is necessarily incorporated into the phenotype schema - perhaps due to processing limitations (Stanton, Chambers, & Piggott, 2001), or the possibility that the information available may be inaccurate or simply untrue (e.g. Lampinen, Copeland, & Neuschatz, 2001; Loftus, 1979; Pezdek, Whetstone, Reynolds, Askari, & Dougherty, 1989; Read, Vokey, & Hammersley, 1990; Reinitz & Hannigan, 2001). How much information is accepted for possible inclusion in the phenotype schema is influenced by a bias - a tendency to accept less, or more, information. Such a bias may be moderated by, for example, previous experience, training, emotion, etc. The resulting phenotype schema will then drive an individual's decision-making and behaviour that, in turn, may modify both the situation and existing genotype schemata – and so the cycle continues.

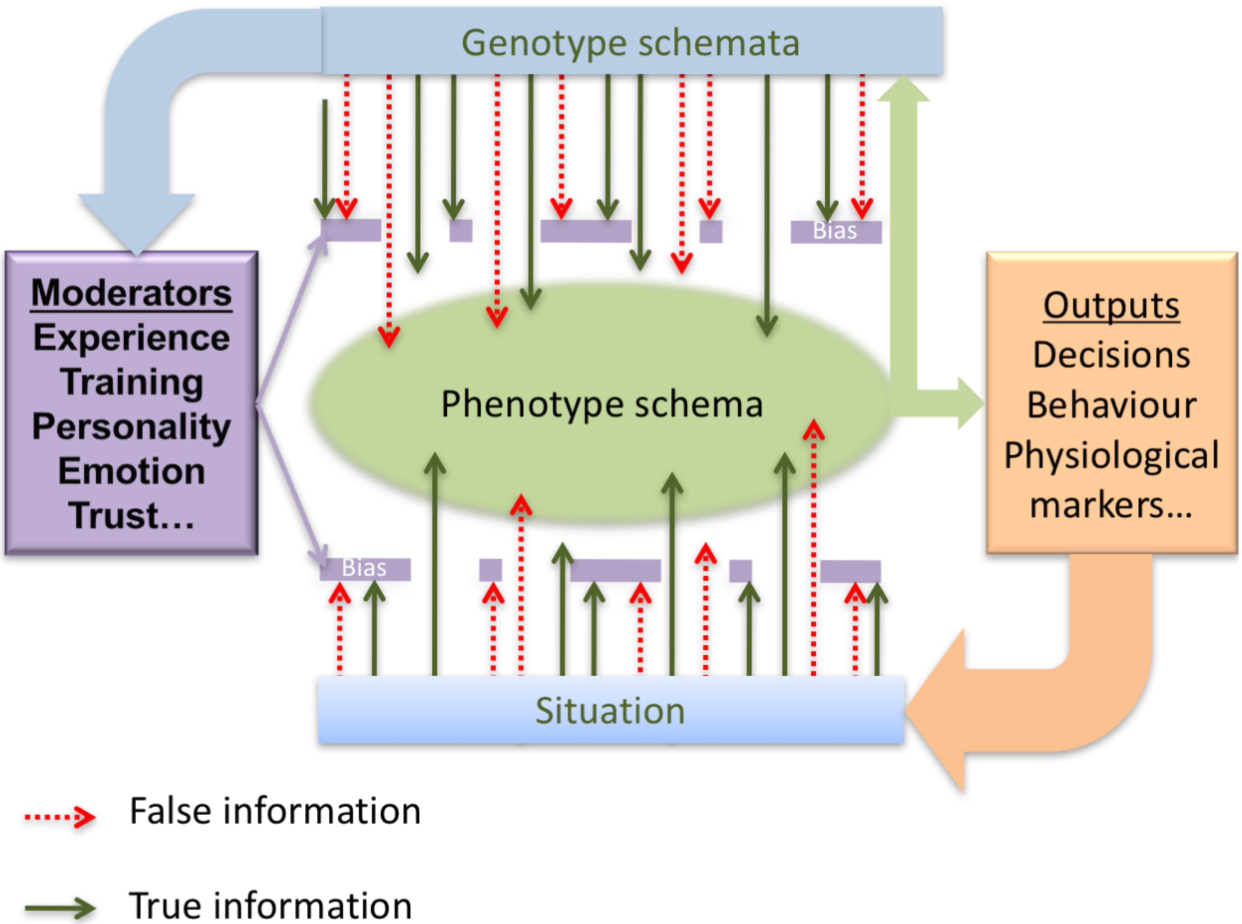

Figure 1. A model of ASA based on schema theory. ASA is underpinned by a phenotype schema that is built using information accepted from genotype schemata, and from the situation. The phenotype schema that is built in any particular situation then drives an individual's decisions and behaviour (and there may be other markers that can be measured) – and contributes to the modification of existing genotype schemata. The amount of information accepted or rejected for possible inclusion in the phenotype schema may be mediated by influences such as experience, personality and emotion.

There is evidence that individuals do accept more, or less, information depending on context and experience. Endsley and Smith (1996) found that fighter pilots tended to direct attention to stimuli that they perceived as particularly important to the task. Endsley and Rodgers (1998) found that air traffic controllers reduced attention to less important information as workload increased. Kahneman and Tversky (1973) found that people often did not use all the information given to them in short vignettes when making judgements and Maule and Mackie (1990) found that some subsets of available information may be processed to a lesser extent than others – or ignored completely.

Durso and Gronlund (1999) argue that cognitive processing limitations can be ameliorated by the selection of only important information, a process that becomes more effective with increasing experience (and more refined schemata) related to the task (Sohn & Doane, 2003; Sohn & Doane, 2004). Errors of information selection (rather than lack of available information) may underpin serious errors of judgment, for example friendly-fire (Edgar & Edgar, 2007; Edgar, Edgar, & Curry, 2003).

One issue for our proposed model of SA is identifying the process by which information may be accepted or rejected. We believe that the concept of System 1 and System 2 processing (Stanovich & West, 2000) is helpful here. System 1 processes are *'fast, automatic, or unconscious'* and System 2 processes '*require access to a single, capacity-limited central working memory resource'* (Evans, 2008, p. 270). System 2 processes represent the traditional 'information processing' mode, where a mental model (or phenotype schema) is built explicitly and consciously using information from a particular situation. The process and the output are largely accessible to the individual. System 1 processes are very different, as are the outputs. System 1 may process large amounts of information at a subliminal level, with a suggestion that the processing of emotionally-valenced stimuli is favoured (Öhman, Flykt, & Esteves, 2001). The output may then be an emotional (somatic) marker that moderates the information processing of System 2, as demonstrated by the effect of emotional state on cognitive processing. Current models of SA are heavily based on System 2 processing, although approaches addressing aspects of System 1 processing are being developed (e.g. Croft, Banbury, & Thompson, 2001).

System 1 processing may provide a mechanism for the selection of information represented by a bias as proposed by QASA. '*System 1 continuously generates suggestions for System 2: impressions, intuitions, intentions, and feelings'* (Kahneman, 2011, p. 24). System 1 processing could 'label' items of information, perhaps by highlighting some items at the expense of others, with a concomitant increase in the 'representation strength' of the preferred items. In this way the output of System 1 processing does not represent implicit knowledge in the way conceptualised by Croft et al. (2001), but represents a labelling of information in such a way that it can be incorporated (or not) into System 2 processing.

The rest of this paper describes a technique, based on the model presented in Figure 1, that uses signal detection theory as a basis for a measure of SA and information acceptance (bias). Signal detection theory has been used by a number of researchers to provide a *performance-based* measure of SA (Burge & Chaparro, 2012; Gugerty, 1997; Pritchett & Hansman, 2000), including the Situation Awareness Control Room Inventory (SACRI; Collier & Follesø, 1995; Hogg, Follesø, Strand-Volden, & Torralba, 1995) derived from the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1988) whereby operator responses to queries are compared to parameter trends in the simulators. The approach described in this paper differs from these measures in that it is a knowledge-based approach using signal detection theory, combined with confidence ratings, to give three measures:

1. *Actual situation awareness (ASA).* An individual's knowledge of the situation as compared to the 'ground truth.' An assessment of the phenotype schema.
2. *Bias.* A measure of information acceptance or rejection.
3. *Perceived situation awareness (PSA).* An individual's perception of how good they believe their SA to be.

The technique, Quantitative Analysis of Situation Awareness is referred to as QUASA (Edgar & Edgar, 2007; Edgar et al., 2003; Edgar, Smith, Stone, Beetham, & Pritchard, 2000) if the underlying metrics are parametric and QASA if non-parametric (discussed in more detail below). As we generally use non-parametric signal detection measures, the acronym QASA will be used throughout the rest of this paper when referring to the technique.

The QASA approach has now been applied successfully across a wide range of application domains including military command and control (Edgar & Edgar, 2007; McGuinness, 2007; Nikolla, Edgar, Catherwood, & Matthews, 2017; Rousseau, Tremblay, Banbury, Breton, & Guitouni, 2010; Thomas, 2008; Tremblay, Breton, Vachon, & Allen, 2012), firefighting (Arendtsen et al., 2016; Catherwood, Edgar, Sallis, Medley, & Brookes, 2012), driving (Edgar, Catherwood, & Melhuish, 2009), robotic teleoperation (Gatsoulis, Virk, & Dehghani-Sanij, 2010), and education (Edgar, Catherwood, & Grover, 2007).

The aim of this paper is to present a detailed consideration of the implementation and interpretation of the QASA technique, together with some simple demonstration studies that will be used to illustrate the effectiveness of QASA for tracking elements of SA.

Theoretical basis of QASA.

In any situation the available information could plausibly be modelled as shown in Figure 2. Some information available to the individual is true, some is false (for example the information relates to how the situation *was* not how it is *now*). In this model each item has an associated representation strength that, as discussed above, could be generated by low-level unconscious processing of that information by System 1. Figure 2 shows plausible distributions of representation strengths for true and false items.

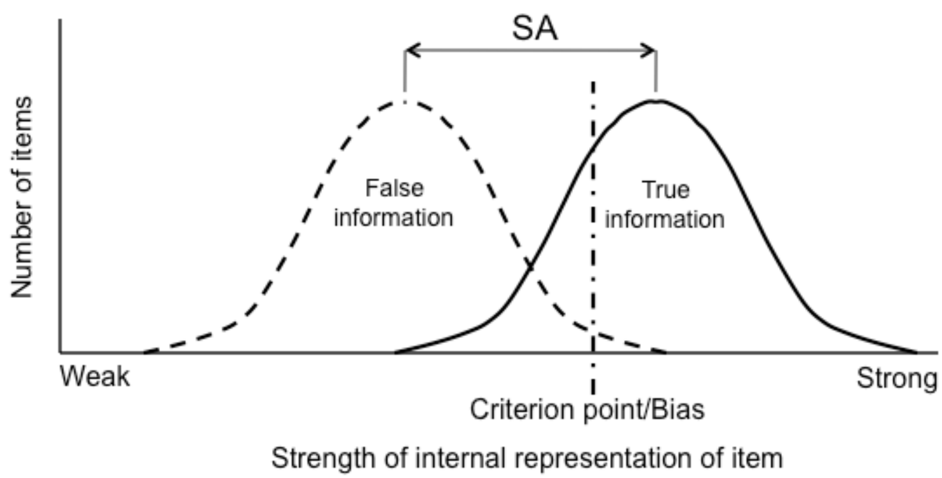

Figure 2. A possible model of the selection of information used to build up a phenotype schema of a situation. Both true and false information may be stored, with each type of information associated with a range of representation strengths. Information with a representation strength above a certain criterion will be accepted for possible use, information below that representation strength will be rejected.

Some overlap in the representation strengths of the two distributions is likely, as the representation strength of false items could be boosted by misinformation, misinterpretation of information, or expectancies of what should be in the situation. Likewise, the representation strength of true items could be reduced by inattention or disbelief. It is even possible for the two distributions to be reversed (the individual completely misunderstands the situation and false items are more strongly represented than true). The form of the distributions shown in Figure 2 are familiar within signal detection theory (for an overview see, Macmillan & Creelman, 1991) , where the task is to distinguish signal from noise.

Signal detection theory can be adapted to assess ASA by measuring an individual's ability to discriminate true (signal) from false (noise) information relating to a particular situation. Signal detection theory also allows another measure of information acceptance to be derived based on the 'criterion point.' The criterion point is a threshold level of, in this case, representation strength. Items of information that have a strength greater than the criterion point will be accepted as true, those that have a strength below the criterion point will be rejected as false.

If the criterion point moves to the right (a high, or positive, bias) then more information is rejected as untrue and less accepted as true. Conversely, if the criterion point moves to the left, then more information is accepted as true and less rejected as false. In signal detection theory, the criterion point and sensitivity (derived from the underlying signal/noise distributions) are independent (for a comprehensive discussion of this issue see Pastore, Crawley, Berens, & Skelly, 2003); that is, a change in one should not drive a change in the other. An important distinction, however, is that the criterion point differs from its derived measure - 'bias'.

As already stated, the criterion point can change independently of the signal and noise distributions and *vice versa*. Such changes may alter the amount of information above or below that criterion point in two ways:

1) The individual's criterion point changes to accept items with a lower or higher representation strength as being true. Such a change has no effect on the underlying signal/noise distributions, just the amount of information accepted or rejected.

2) The individual's criterion point stays the same, but the signal and/or noise distributions move (possibly resulting in a change in sensitivity/ASA). For example if the 'true' distribution (Figure 2) moves to the right (as a result, perhaps, of a change in expectancies), then more information is accepted as true, even though the criterion point has not shifted.

The amount of information accepted or rejected changes as *either* the criterion point or the underlying distributions move, although a change in the criterion point does not affect the underlying distributions. The criterion point is a measure of the *representation strength* above which information is accepted, and below which information is rejected and is independent of the underlying distributions. Bias (as measured by the QASA tool) is a measure of the proportion of information accepted/rejected, based on the position of the criterion point *with respect to* the underlying distributions. Bias is therefore not a measure of the criterion point it is a measure of *information acceptance* although, in practice, measures of bias and criterion point may give very similar results (Nikolla et al., 2017).

Although it may be possible to infer information *use* from the bias measure just as, for example, heart rate variability may be used to infer cognitive workload (Jorna, 1992), bias and information use need not be synonymous. An individual may accept information as true, but not use it to build SA as they consider it, for example, to be irrelevant.

Many measures of sensitivity and bias are available as signal detection theory represents a conceptual framework rather than a single method of analysis (see, Pastore et al., 2003; Wickens, 2001). The underlying signal detection measures chosen should be informed by the data being analysed, in the same way that different statistical tests are used to analyse different types of data. Two measures of sensitivity are considered here that could be used within QASA; one measure represents a ‘parametric’ approach, and the other ‘nonparametric’. The issues discussed also apply to other signal detection measures (Verde, Macmillan, & Rotello, 2006) within the two broad categories (parametric vs nonparametric).

The first measures of sensitivity and bias to be considered are also perhaps the most widely used, namely d' (sensitivity) and (often log) ß (bias). The d' measure is the *z*-score distance between the means of the two distributions. These signal detection measures (d' and ß) do, however, make certain assumptions that the item representation strengths are Gaussian distributions with similar variances. If these assumptions are violated, then the utility of d' and ß are compromised (Pastore et al., 2003). A number of studies suggest that underlying evidence distributions across a variety of situations are likely to be Gaussian, but it appears far less likely that such distributions will have equal variance (Ratcliff, Sheu, & Gronlund, 1992; Stretch &

Wixted, 1998; Swets, 1986). Furthermore, as sensitivity drops to zero (the individual is guessing), the bias also drops to zero (Snodgrass & Corwin, 1988).

The obvious solution to the problems with *d'* and ß is to use a nonparametric SDT measure. Unfortunately, there does not appear to be one as yet. A' (A prime) (Pollack & Norman, 1964) is often reported as a 'nonparametric' measure of sensitivity in signal detection theory; usually together with B" (B double prime) as a measure of bias. Calculations of A' (See, Stanislaw & Todorov, 1999) and B" (Grier, 1971; Hodos, 1970) are based around a straightened 'receiver operating characteristic' (ROC). A' is a more, or less, accurate measure of sensitivity depending on how close the actual ROC curve is to the 'straightened' version. As with *d'*, A' and B" are also not entirely independent. At 'perfect' levels of discrimination (no errors), B" cannot be calculated and, at zero discrimination (the individual is just guessing) B" also drops to zero (Snodgrass & Corwin, 1988). An alternative measure of bias (such as B"D - Donaldson, 1992) computable even when discrimination is zero (See, Warm, Dember, & Howe, 1997) is worth considering to avoid having to enter undefined bias scores as missing data in analysis. Which measure is most appropriate requires further research and understanding of the nature of ASA. Like ß, B" may also change with changes in sensitivity and so is also a measure of bias (information acceptance) rather than criterion-point.

There currently appears no 'perfect' signal detection measure that can be used for the calculation of ASA. *d'* provides a marginally more accurate estimate of sensitivity than A' if the signal and noise distribution variances are equal and Gaussian (Donaldson, 1993). For the purposes of measuring ASA, the authors favour A' (and B") for the following reasons:

1) A' is conceptually related (although not identical) to 'percent correct' as a measure of performance (Pastore et al., 2003), which allows easier, and more intuitive, comparison with other measures of SA, such as SAGAT (Endsley, 1988).

2) *d'* can be undefined if, for example, an individual responds 'yes' to all the signal trials and 'no' to all the noise ones. The value of *d'*, as well as its variance, is affected by the various conventions (discussed in Miller, 1996) used to handle the issue of *d'* being undefined.

3) A' tends to be relatively more robust than *d′* when sample sizes are small (Verde et al., 2006). With small samples, the sampling distribution of *d′* is likely to be neither Gaussian, nor unimodal (Miller, 1996). Given that large samples in ASA measurement are unlikely, due to interference with the situation being assessed, a measure robust to small samples is appropriate.

Perceived/subjective situation awareness

We have seen how QASA can provide a measure of ASA (the ability to tell true from false information), together with a measure of information acceptance (bias). Another facet of SA is an individual's conscious awareness of their own SA and the ability to reflect on that SA. QASA provides a measure of subjective, or perceived, SA (PSA) by asking participants to rate their confidence (on a scale from 'guess' to 'certain') that the answer that they have just given to each True/False probe is correct. A four point scale is used as it allows the scale to be split for the calculation of Type 2 signal detection theory measures (Clarke, Birdsall, & Tanner, 1959; Galvin, Podd, Drga, & Whitmore, 2003; Pollack, 1959), although these further measures will not be described in this paper.

One finding in the SA literature is that measures of actual and perceived SA show varying correlation (e.g. Edgar, Catherwood, Sallis, Brookes, & Medley, 2012; Endsley, Selcon, Hardiman, & Croft, 1998; Sætrevik, 2012). The disparate results that may be obtained for ASA and PSA may provide an explanation for the difficulty in linking SA to task performance. No matter how good an individual's ASA is their decision-making and performance can still be moderated by how good they *believe* their SA to be (see, for example, Bingi, Turnipseed, & Kasper, 2001; Endsley, 1993; Griffin & Tversky, 1992). As Endsley *et al.* (1998) point out, a person's PSA may determine how they act, and this PSA may, or may not, be a reflection of their ASA.

Implementation of QASA

The process for applying QASA is as follows:

1) Generate True/False probe statements drawn from the situation of interest.
2) Collect participant responses to those probe statements and classify them with respect to the 'ground truth,' according to signal detection theory, as hits/misses/false alarms/correct rejections.
3) Collect participant confidence ratings for each T/F response indicating how confident the participant feels that the response that they have just given is correct.
4) Analyse the responses to the T/F statements, using signal detection theory metrics, to give measures of ASA and bias.
5) Analyse the confidence ratings to give a measure of PSA.

We now provide a detailed consideration of how probe statements may be generated and administered within QASA. This is then followed by two controlled studies that demonstrate how the data collected by QASA may be analysed and interpreted.

Probe generation

The basis of the QASA approach is the generation of a number of statements concerning the situation of interest. Typically half the statements are true (although this is not a requirement) and half are false. Generating the true statements is relatively straightforward as they directly reflect the situation. Generating the false statements can be more difficult. The question is, 'How false should the false statements be?' Some statements are essentially binary (the undercarriage of the aircraft is either locked down or it is not) and some lie along a dimension. For example, if you want to test a pilot's awareness of their flying height with a false statement, how far from the true height should the false statement be? A task analysis or subject matter expert (SME) is invaluable for specifying the constraints for good lures. For instance, if it is determined that a pilot should know their flying height to within 50m, then a false statement could give the flying height ±51m.

Another issue with probe generation is whether probes should be framed positively or negatively. We have always used positively-framed probes (e.g., 'There is an enemy unit at position x.' rather than negatively-framed (e.g., 'There is *no* enemy unit at position x.'). This fits with the theoretical background to the approach that is based on the acceptance or rejection of information, rather than the lack of information. Using negatively framed questions, however, may give insights into another aspect of SA, and would be worth further research.

Definition of the situation of interest is important to any technique for measuring ASA (Flach, Mulder, & van Paassen, 2004). The probe statements within the QASA technique, as with any probe technique (such as SAGAT) do more than test awareness of a situation, the probes *define* that part of the situation represented in the ASA measure. One advantage of probe techniques, however, is that the probes can be inspected and the situation tested can be deduced from the probes used (Lau, Jamieson, & Skraaning Jr, 2014).

QASA is a psychometric test inasmuch as it has been constructed based on psychological theory. QASA, however, differs from psychometric tests of attributes such as personality and intelligence that can, and do, present the same items in every administration of the test. QASA is designed to measure awareness of a potentially infinite range of situations and so, every time awareness of a different situation is assessed, a new set of probes will be used. We present here a description of a process we have developed for generating and refining probes, and demonstrate the application of that process to one situation.

The situation was a simulation of a fireground (a factory fire) using a sequence of videos and pictures, as described in Catherwood et al. (2012). Forty two probes were generated by a senior officer within the UK Fire and Rescue Service. As a part of the process for generating each probe, the officer was asked to evaluate and record the consequences of the possible answers to each probe in terms of signal detection theory. For each probe the officer considered the importance of giving a correct answer (hit or correct rejection) and the costs of an incorrect answer (miss or false alarm). For example, consider the probe statement:

'The factory had an asbestos roof.'

The officer recorded that a correct response would indicate an awareness of whether there was the possibility of asbestos contamination, and whether collapsing the roof would be a viable firefighting option. The consequences of getting it wrong would be that firefighters may be exposed to asbestos contamination, or reject useable firefighting options.

The officer was also asked to rate on a scale from 0 (not important) to 10 (crucial) how important it would be, in relation to effective performance, for a firefighter to answer each probe correctly. The rationales and ratings for each probe were then given to two serving firefighters who were familiar with the scenario. The two firefighters then independently rated each probe as to the importance of a correct response for a firefighter's ASA in that scenario. The mean importance rating across the two raters was 6.96 (SE = 0.31). High importance ratings are pivotal for content validity as, if a probe is not considered to test anything important to ASA, then its inclusion in the test will impact the content validity of that test.

Interrater reliability was assessed using the intraclass correlation coefficient (ICC, Shrout & Fleiss, 1979). Using a two-way random model (consistency definition), the average measured ICC was 0.86 with a 95% confidence interval from 0.74 to 0.92 $F(41,41) = 7.20$, $p < 0.001$). Interrater reliability of the probe importance ratings can be considered very high.

The scenario was run on 35 (all male, mean age 42.1 years, SD 7.6 years) serving firefighters. Internal consistency of the probes was then assessed using KR-20 (Kuder & Richardson, 1937). KR-20 was calculated to be -0.01. The scores on the probes were not correlated. Whether or not probes should be correlated in measuring SA is a debate that is beyond the scope of the current paper.

The percentage of correct responses to each probe was then calculated. If participants were just guessing, the percent correct should have been approximately 50%. If the percentage of correct answers dropped below this there may have been a problem with the probe. This distinguished a difficult probe (where one might expect the percent correct to be around 50%) from a poor probe.

The findings from the pilot study showed an even spread of percent correct scores from approximately 30% up to 100% across the probes. This spread is important as it gives an indication that the sensitivity of the test is appropriate. Percent correct scores bunched around 50% or 100% would indicate floor or ceiling effects, either of which would suggest that the test may not be sensitive to variations in SA. There was one probe statement, however, for which the percent correct was approximately 10%. This is far below the, 'guessing rate' of 50% and suggests a problem with the wording of the probe, which was, 'There was only a small amount of storage.' Perhaps the participants' judgement as to what constituted, 'a small amount of storage,' was different to that of the officer that devised the probes. This probe was removed.

The importance ratings given by the two raters were then inspected, and any probes for which the importance ratings differed by 3 or more, were removed. Four probes were removed for this reason. Any probes with a mean importance rating of less than 5 were also removed. Six probes were removed for this reason, leaving 31 probes remaining. The cut-off points were arbitrary, but can be refined by further research. For example, it would be possible to set a criterion importance rating, and to keep refining and filtering the probes until this rating is achieved. This process was conducted *post hoc.*

One advantage of filtering the probes *post hoc* is that participant responses can be used, as we have, to identify poorly worded questions. A beneficial consequence of filtering probes *post hoc* is also that a proportion of the probes used in a test may, either accidentally or deliberately, be irrelevant to performance in the situation under test. The irrelevant probes may be presented to the participant to mask the situation under test, and then removed for the analysis. This has the advantage that it lowers the face validity of the measure (to the participant) while keeping content validity, making it more difficult for participants to use the probes to inform or guide their own ASA.

Following the removal of the probes, the mean importance rating was recalculated as 7.87 (SE = 0.24), an improvement on the previous mean (6.96). Interrater reliability was also reassessed. Using a two-way random model (consistency definition), the average measured ICC was 0.84 with a 95% confidence interval from 0.66 to 0.92 $F(30,30) = 6.16$, $p < 0.001$). This value is very close to that obtained before the probes were filtered and may reflect the fact that, although probes with low agreement were removed, some of the probes on which the raters most closely agreed were also removed as they were rated low in importance.

This method can be applied to any situation and, we believe, allows important aspects of sensitivity, reliability and validity to be established for QASA for each situation.

Probe Administration

One issue with QASA, along with other measurement techniques that use probes to assess aspects of ASA is when, and how, to administer the probes. Probes can either be delivered during the task or they can be administered *post hoc*. *Post hoc* administration of probes does not interfere with performance on the task, but does introduce an increased memory component into the measurement. Presenting the probes within the task reduces the memory load, but the individual may use the probes to guide how they build ASA. Including 'masking' probes that do not probe the situation of interest may reduce this effect.

If presentation of probes within a task is used, then there is also the choice of whether to stop the task while the individual completes the probes. SAGAT, for example, uses a 'freeze technique' whereby the situation is suspended while the probes are answered. Endsley (1995) has reported studies that showed no significant effect of briefly freezing a situation on performance. McGowan & Banbury (2004), however, found that interrupting a driving hazard-perception test impaired performance on that test and so it is unwise to assume that interrupting a task is always without cost in terms of SA. The Situation Present Assessment Method SPAM (Durso & Dattel, 2004) (Durso et al., 1995) utilises the idea that most tasks have some 'slack time' within them and makes the probes available without stopping the task. The time taken to access the probes is taken as a measure of workload. QASA probes can be presented during a freeze in the task (as in SAGAT), concurrently with the task (as in SPAM), or *post hoc*.

As to how many probes should be administered when using QASA, the short answer is, 'As many as possible.' We regard 20 probe statements per scenario as the absolute minimum.

Studies (Catherwood et al., 2014; Catherwood et al., 2012; Edgar & Edgar, 2007, and below) suggest that, if less than 20 probes are used; a few responses can have a disproportionately large influence on both ASA and bias.

Examples of the application of QASA

We now present two studies to illustrate the use of QASA. The situations described in other papers using QASA were quite complex, making it difficult to establish that the figures provided by QASA were a true measure of ASA. These two studies were designed to assess the ability of QASA to track changes in ASA, PSA, and bias in a highly controlled laboratory situation where we knew at what point a change in ASA had occurred. Ecological validity has been sacrificed for precise control of changes in ASA. The studies should be considered, therefore to be *demonstrations* of the capabilities of QASA to measure and track changes in ASA, PSA, and bias.

Study 1

In the first study the task chosen was an abstract category-learning task based loosely on the Wisconsin card sorting task (WCST) (Berg, 1948; Grant & Berg, 1948). Using an abstract task made it unlikely that participants were able to draw on pre-existing genotype schemata to build the phenotype schema related to the situation.

Participants were shown a series of abstract images (an example is shown in Figure 3) and were asked to determine whether images were representative of a predefined category (defined by two aspects of the stimulus patterns, e.g. colour and shape) based on feedback.

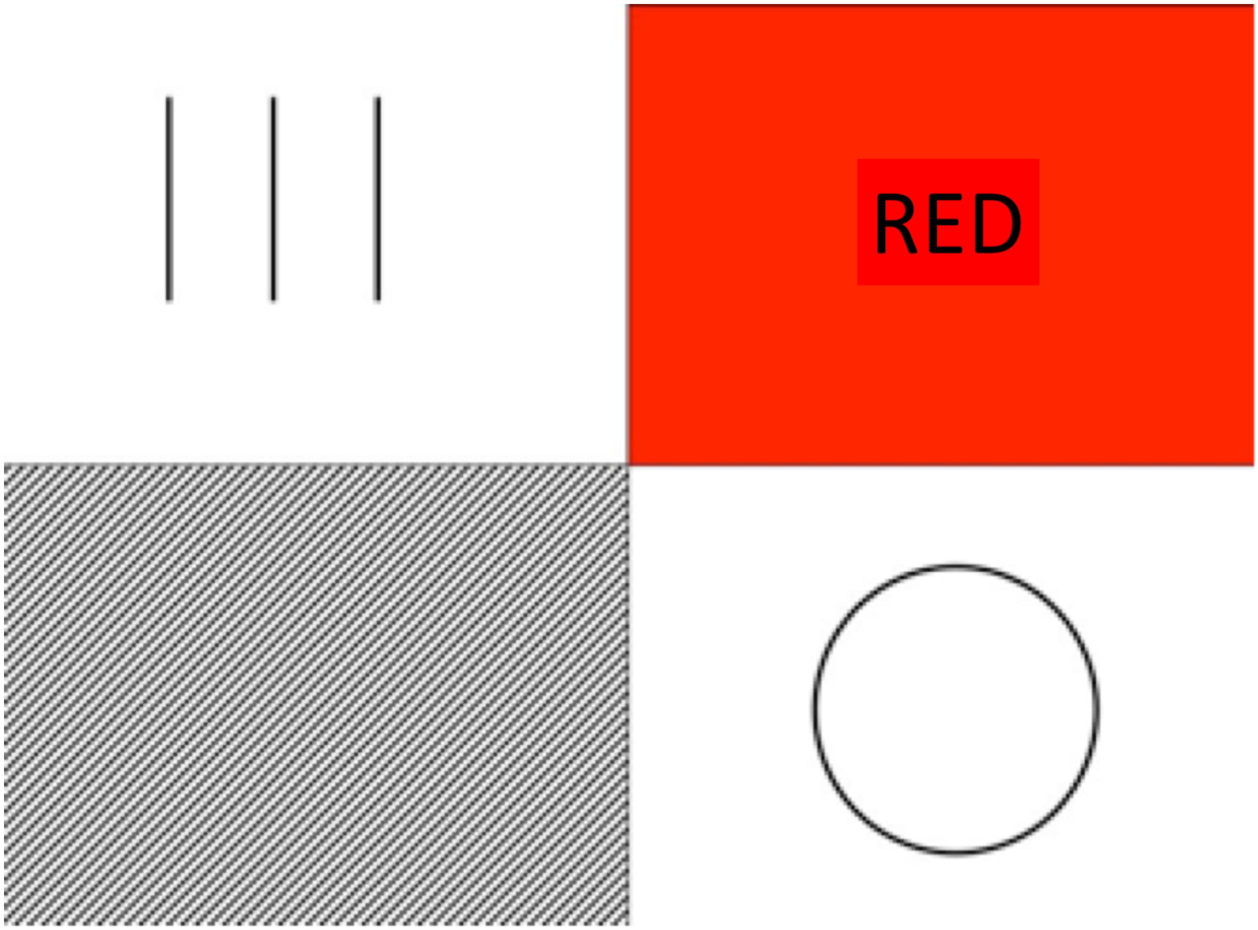

Figure 3. An example of a stimulus used in Study 1. Each stimulus consisted of four quadrants and each quadrant could show one of three states. In the top left quadrant there could be one, two, or three lines. The top right quadrant could be coloured red, green, or blue. The bottom left quadrant contained lines that could be oriented vertically, horizontally, or obliquely and the bottom right quadrant could contain a diamond, a circle, or a square.

In complex situations, a drop in measured ASA could be due to a loss of ASA, or the measurement technique not measuring the relevant aspect of the situation (i.e., a lack content validity). In this study, good ASA was clearly defined (awareness of the underlying category) and a loss of ASA was *forced* midway through the task by changing the target category without informing the participant. To ensure that each participant did show a *loss* of ASA we excluded any participants that did not reach a criterion ASA prior to the category change.

Method

Participants. Sixty-four participants (46 female, 18 male) were tested. Modal age range was 18-30 years. All participants had self-reported normal or corrected-to-normal vision.

Stimuli. Participants were presented with a series of 104 slides, an example of one of which is given in Figure 3.

A ‘target category’ was defined by a combination of *two* aspects of the slide. For example in the first 52 trials the target category was, ‘three lines and red’. After 52 trials the category was switched and, in this case, in the next 52 trials the target category was diagonal lines paired with a circle. The order of presentation of the categories was counterbalanced across participants. Although the presentation order was random within each block of 52 trials, the second block of 52 trials contained exactly the same slides as the first 52; only the target category changed. The target category was presented on 50% of trials.

Procedure.

There was only one probe statement that remained the same all the way through the study:

‘This slide represents a member of the target category. True/False?’

Each trial consisted of the presentation of a fixation marker for between 1 and 1.5 seconds (the timing was randomised within this range) followed by one of the ‘quadrant’ slides. Participants were asked to press ‘T’ on the keyboard if they believed the target category was present or ‘F’ if they believed it was absent. Following the participant’s response, they were asked to indicate on a four-point scale ranging from 1 (guess) to 4 (certain that answer was correct) how confident they were that the response they had given was correct. Following their confidence rating, participants were given feedback as to whether or not their response was correct. The fixation spot then reappeared and the next trial began. Trials were self-paced in that participants had as long as they wished to make a response, although the feedback did include a measure of their reaction time. Participants were given 10 practice trials and, if prepared to continue, were presented with the 104 trials.

Analysis

The True/False responses were analysed using the QASA approach. A' and B" were calculated using formulae described by Stanislaw and Todorov (1999):

$$A' = 0.5 + \left(sign(H-F)\frac{(H-F)^2 + |H-F|}{4\max(H,F) - 4HF} \right)$$

$$B'' = sign(H-F)\frac{H(1-H) - F(1-F)}{H(1-H) + F(1-F)}$$

H = ‘Hit’ rate
F = ‘False alarm’ rate
Max(*H,F*) = Either *H* or *F*, whichever is greater.

Using the formulae above, values of A' will be in the range 0 to 1 (with 0.5 indicating an inability to tell true from false information) and B" will be in the range -1 to +1. The confidence (PSA) scores will be in the range 1 to 4. For ease of plotting and interpretation the QASA tool rescaled all measures to a scale running from -100 to +100. As the form of the underlying distribution for the calculation of B" is logistic (Macmillan & Creelman, 1990), a (natural) logarithmic transformation was applied to the bias values to give a linear relationship between bias and the underlying distribution (although this has not been used in previously published papers, e.g. Catherwood et al., 2014; Catherwood et al., 2012; Edgar & Edgar, 2007). The implications of the scores for ASA, PSA, and bias, and confidence are given in Table 1.

Table 1. Interpretations of the ASA, PSA, and bias scores provided by the QASA tool.

Score	Actual Situation Awareness (ASA) - A'	Bias - B''	Perceived situation awareness (PSA) - Confidence
Positive (max +100)	Good ASA. Can tell true information from false: higher score is better.	'Strict' bias. Tendency to reject information as false even if true. The higher the score the greater the tendency to reject information as false.	Indicates a *belief* that the responses given are correct, suggesting a belief that SA is good. A higher score represents greater confidence.
Zero	No ASA – guessing?	No bias towards accepting or rejecting information. A 'neutral' attitude.	Neither high nor low confidence.
Negative (max -100)	Misguided. Believes false information is true and *vice versa*. More negative is worse.	'Lax' bias. Tendency to accept information as true even if false. The more negative the score the greater the tendency to accept information.	Indicates a *belief* that the responses given are wrong, suggesting a belief that SA is poor.

Measures of dynamic ASA, PSA, and bias were obtained by using a 'rolling' twenty-trial 'window' advancing four trials at a time. The first data point was calculated using the block of trials from 1-20. The next data point was calculated using trials, 5-24, then 9-28, and so on. Although each True/False probe was administered only once, most were used in more than one, 'window'. Any participants that failed to achieve an ASA score of at least 50 (on the -100 to +100 scale) in the final block before the category was changed (forcing a loss of ASA) were excluded from further analysis. Out of 64 participants, 19 were excluded from the analysis for failing to reach this criterion.

Results

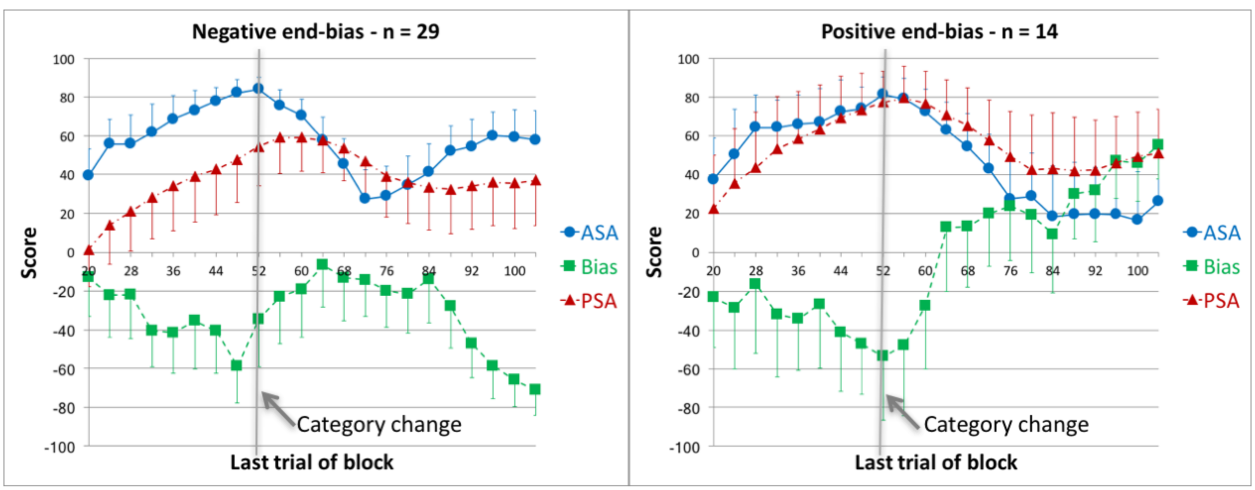

Figure 4. Rolling values for ASA, PSA, and bias in Study 1 averaged across participants and classified into two groups depending upon whether the bias calculated in the final block of trials was positive or negative. Error bars indicate 95% confidence interval.

Mean dynamic ASA, PSA, and bias scores are plotted in Figure 4. All participants that achieved the criterion ASA level (n = 45) showed the same pattern of ASA in the learning phase, with a steadily increasing ASA score; followed by a drop in ASA score when the category was changed. The relationship between bias and situation awareness *before* the category change was consistent across participants, with increasing ASA scores associated with a generally negative bias.

After the category change two distinct groups were identified from inspection of the data. This was an exploratory classification, the aim of which was to establish the ability of the QASA method to reveal different patterns of responding in the data. Other classifications were possible, the one chosen was designed to show that QASA can reveal different patterns of SA and bias. No *a priori* theoretically-informed predictions were made.

The two groups were classified using the measured bias score for the final 20-trial block in the trial to define a 'negative-bias' (n = 29) and a 'positive-bias' (n = 14) group. If the bias was zero that participant was excluded from the analysis (two participants were excluded from the analysis for this reason) as there was no theoretical reason to assign them to either the negative or positive bias group. With a larger sample, it may have been possible to create a third group, but it would have been impractical to include a group with only two participants in this analysis.

In both groups the bias initially became more positive post-change (representing a 'stricter' criterion and a tendency to reject information as false) with the loss of ASA, but in one group the positive bias shift was greater and maintained. In the other group, after an initial slight rise, the bias decreased (became more lax). Differences in the ASA, bias, and PSA scores for the negative- and positive-bias groups were analysed using Mann-Whitney U (two-tailed) tests as the data were not normally distributed. Some bias scores could not be calculated if the ASA score reached 100 and were entered into the analysis as missing data. No Bonferroni corrections were applied in any analyses as this would be inappropriate for an exploratory study of this nature (Bender & Lange, 1998; Perneger, 1998). For the scores in the block of trials immediately preceding the change in category there was no significant difference between the two groups in ASA ($U = 178.50$, $N_1 = 29$, $N_2 = 14$, $p = 0.520$), bias ($U = 91.00$, $N_1 = 19$, $N_2 = 11$, $p = 0.582$) or

PSA (U = 247.50, N_1 = 29, N_2 = 14, $p = 0.240$) scores. In the final 20-trial block of the study there was a significant difference in ASA (U = 127.00, N_1 = 29, N_2 = 14, $p = 0.049$) and bias (U = 378.00, N_1 = 29, N_2 = 14, $p < 0.001$) scores, but no significant difference in PSA scores (U = 227.00, N_1 = 229, N_2 = 14, $p = 0.530$).

Spearman's rho (two-tailed) was used to assess the correlation between ASA and PSA for negative- and positive-bias groups. The correlation between ASA and PSA, across the duration of the whole scenario (pre- and post-category change) approached significance at $p < 0.05$ (r_s = 0.412, N = 22, $p = 0.056$) in the negative-bias group and was significant in the positive-bias group (r_s = 0.756, N = 22, $p < 0.001$).

Discussion of Study 1

This study used the QASA tool to measure dynamic ASA loss. Both groups showed a similar pattern of data in the first half of the study but a different pattern following the forced loss of ASA. In the first half of the study the QASA tool tracked a steady increase in ASA and confidence, with a steady drop in bias. When a loss of ASA was forced, the measured ASA, PSA, and bias all changed. Both groups initially lost ASA at a similar rate and showed a positive-shift in bias (more pronounced in the positive end-bias group). Changes in PSA tended to lag slightly behind changes in ASA, perhaps reflecting the time taken for the loss of ASA to be appreciated by participants and to be reflected in their PSA scores.

One group, post loss of ASA, maintained a positive bias and *did not* regain ASA; the other group showed a negative-shift in bias and *did* regain ASA. Note that the classification of participants into groups was as a result of inspection of the data. It is possible to split participants into groups using different criteria (e.g. bias gradient following a loss of ASA) but this was an exploratory study and, until more data are gathered using the QASA approach, it is difficult to define groups *a priori*.

Changes in bias, interpreted as a tendency to accept or reject information, could provide insight into why the participants in one of the identified groups tended to regain ASA, whereas those in the other group did not. A positive-going bias suggests a stricter criterion, with a reduction in the amount of information considered when building ASA. Focusing on a restricted range of information may inhibit performance on this task. For example, focusing down on a single aspect of the stimuli (e.g. colour or shape) would not allow a recovery of ASA, as a combination of two criteria is necessary to deduce the underlying category. Those participants that showed a more ‘restricted’ bias generally did not regain ASA, those that showed a more ‘liberal’ bias did suggesting, perhaps that bias does reflect (at least in this study) not only information acceptance, but information use when building ASA.

Perceived SA, as measured by the confidence scores, tended to track ASA, and to be correlated with it, particularly for the positive-bias group. The time lag between changes in ASA and changes in PSA would lower the correlation between them. As already discussed, a relationship between actual and perceived SA is not always found but, in this study, the immediate feedback on response accuracy may help align actual and perceived SA. The lag in PSA compared to ASA may also explain why a link between actual and perceived SA is not always found.

This Study allowed the QASA tool to be tested using a tightly defined and abstract situation, with minimal influence from top-down factors such as prior experience or expectations driven by genotype schemata. The next study investigated changes in ASA, PSA, and bias using a more realistic task that might be expected to tap into genotype schemata to a greater extent than Study 1.

Study 2

Study 2 was designed to be as similar as possible to Study 1 but used more realistic stimuli – real world (urban street) scenes. Moving from abstract stimuli to real-world scenes, allowed a greater possibility that participants would use pre-existing genotype schemata in building ASA.

Participants were briefed that they were in the role of monitoring incoming pictures from surveillance cameras in the context of a possible terrorist threat. Participants were shown a series of pictures of individual's in urban contexts (see, e.g. Figure 5) and their task was to identify whether or not they believed the main subject of the picture to be a 'threat'. Thus the category that had to be learned was what aspects of the pictures defined a threat.


DORAL
DORAL

Figure 5. An example of a stimulus used in Study 2. In one half of the study the category was defined by the main subject of the picture carrying a bag; in the other half, the category was defined by the main subject not carrying a bag.

Initially, this study was designed to be a direct analogue of Study 1, but using more realistic stimuli and greater affective content. A threat was originally defined by a combination of two characteristics of the main subject of the picture (e.g. a *man* carrying a *bag*) and this definition was used in a pilot study. In the pilot study, however, we found what we believe to be clear evidence of genotype schemata influencing, and impairing, performance on this task. A number of participants commented that they used their pre-existing ideas of what might constitute a threatening individual (such as, for example, aspects of gender or ethnicity) rather than using the feedback provided in the task to inform their choice. The impact of such stereotypical responses was such that we encountered pronounced floor effects in the pilot study – participants failed to gain ASA at all. Given that we hoped to study loss of ASA (for comparison with Study 1), it was necessary for participants to be able to gain ASA. We therefore simplified the task. The defining characteristic for a threat was the presence (or absence) of a bag carried by the main subject of the photograph.

Method

Participants. Forty two participants (29 female, 13 male) were tested, modal age range 18-30 years. Seventeen participants were excluded from the analysis as they failed to achieve the criterion level of SA (a score of 50 or more in the final pre-change block). As in the pilot study, feedback on debrief suggested that these participants were influenced by stereotypical factors (nobody mentioned bag/no-bag as a stereotypical threat) in attempting to identify a category – and this hindered their learning of the category using the feedback provided. All participants had self-reported normal or corrected-to-normal vision.

Stimuli. Participants were presented with a series of 104 slides, an example of one of which is given in Figure 5. Each slide contained a picture of an urban scene (taken from open-access sources on the internet) with a clear central figure.

A ‘target category’ was defined by one aspect of the central figure. For example, in the first 52 trials, in half the runs, the target category was defined by whether the central figure was carrying a bag of any kind. After 52 trials the category was switched (the category was defined by the absence of a bag). The order of presentation of the categories was counterbalanced across participants. Although the presentation order was random within each block of 52 trials, the second block of 52 trials contained exactly-the-same slides as the first 52; only the target category changed. The target category was presented on 50% of trials.

Procedure The general procedure was exactly-the-same as in Study 1, except that the single True/False probe statement used for all trials was:

‘The person in this slide represents a potential threat. True/False?’

Feedback indicated whether the participant’s response was correct or not and included an arbitrary count of lives saved/lost by the participant’s responses. Correct answers saved lives and incorrect ones lost them. The number of lives saved or lost was linked only to the participant’s reaction time, and not to any aspect of the image they had just viewed. Following the participant’s response, they were asked to indicate on a four-point scale ranging from 1 (guess) to 4 (certain that answer was correct) how confident they were that the response they had given was correct.

Analysis.

The analysis was conducted in the same way as for Study 1.

Results

As in Study 1, two groups were identified using the measured bias score for the final 20-trial block to define a ‘negative bias’ (n = 10) and a ‘positive bias’ (n = 13) group. If the bias was zero that participant was excluded from the analysis, for the same reason as in Study 1 (two participants were excluded from the analysis for this reason). Mean dynamic ASA, PSA, and bias scores for the two groups are plotted in Figure 6.

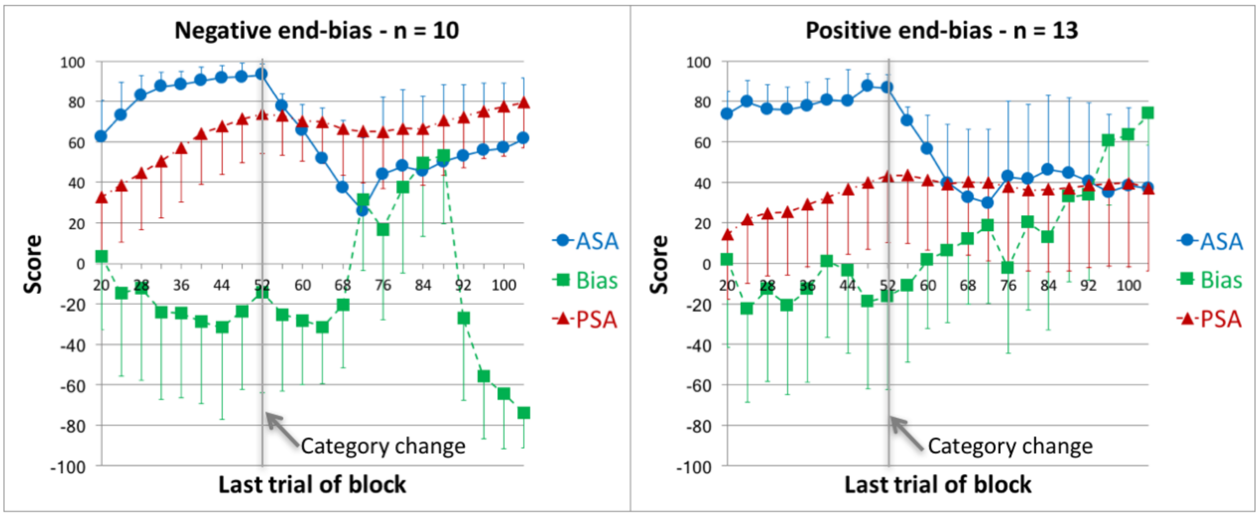

Figure 6. Rolling values for ASA, PSA, and bias in Study 2 averaged across participants and classified into two groups depending upon whether the bias calculated in the final block of trials was positive or negative. Error bars indicate 95% confidence interval.

In this study, a greater proportion of participants fell into the positive-bias group than in Study 1, although the difference only approached statistical significance ($\chi^2(1, N=66) = 3.56$, p = 0.059).

Data were analysed using Mann-Whitney U (two-tailed) tests. For the scores in the block of trials immediately preceding the change in category there was no significant difference between the two groups in ASA (U = 41.00, N_1 = 10, N_2 = 13, p = 0.135), bias (U = 49.00, N_1 =

7, $N_2 = 12$, $p = 0.54$) or PSA ($U = 41.5$, $N_1 = 10$, $N_2 = 13$, $p = 0.141$) scores. In the final 20-trial block of the study there were no significant difference in ASA ($U = 48.00$, $N_1 = 10$, $N_2 = 13$, $p = 0.291$) and PSA scores ($U = 44.00$, $N_1 = 10$, $N_2 = 13$, $p = 0.167$) but there was a significant difference in bias scores ($U = 130.00$, $N_1 = 10$, $N_2 = 13$, $p < 0.001$) as would be expected as the groups were defined by bias score.

For the negative-bias group, the relationship between ASA and bias was similar to that for the same group in the previous study, although the positive shift post-change was more pronounced. Before the category change, the positive bias shift group showed a similar pattern to those in the previous study, although the overall level of bias was more positive. As in Study 1, ASA showed little sign of recovery post-change in this group.

Spearman's rho (two-tailed) was used to assess the correlation between ASA and PSA for negative- and positive-bias groups. The correlation between ASA and PSA was not significant at $p < 0.05$ for either the negative-bias group ($r_s = 0.065$, $N = 22$, $p = 0.774$) or the positive-bias group ($r_s = -0.258$, $N = 22$, $p = 0.165$).

Discussion of Studies 1 and 2.

The participants in Study 2, as in Study 1, could be classified into positive- and negative-bias groups. Both groups in Study 2 showed similar patterns of ASA and bias to those found in Study 1, although with some variation.

The negative-bias group in Study 2 show a larger positive bias shift following the category change than in Study 1. As with Study 1, the negative-bias group in Study 2 showed a recovery of ASA, the positive-bias group did not - possibly reflecting an acceptance of too narrow a range of information to support acquisition of the new category.

The difference in final ASA between the negative- and positive-bias groups was not significant, potentially due to (i) the lower power of the second study (i.e., smaller sample size) and (ii) a positive bias being relatively more effective in Study 2. Focusing down and concentrating on a single criterion of the situation would not work in Study 1 but might work in Study 2 – *if* the single criterion chosen was the right one (bag/no bag).

The negative-bias group in Study 2 showed consistently higher PSA than the positive bias group. Perceived SA for both groups also stayed relatively constant, and did not 'track' the ASA in the same way as in Study 1. The reference to a, 'terrorist threat' in Study 2 may have triggered genotype schemata, priming participants to focus on certain stereotypical aspects of the pictures driven by what they believed would identify a terrorist. Such priming may have increased participants' confidence in their responses as they may have felt such responses were corroborated by previous knowledge – and so were less sensitive to the feedback provided.

General Discussion

Although ecological validity has been sacrificed for experimental control, we believe that the two studies reported in this paper have served the intended purpose of demonstrating that QASA can effectively track changes in SA (actual and perceived) and bias.

Tracking of ASA revealed that a loss of ASA was almost always followed by a rise in bias (greater likelihood of rejecting information). We cannot say with any certainty what precipitated the change in bias following the loss of ASA in this case, but previous research (Nikolla et al., 2017) has demonstrated that such changes may be precipitated by a change in affective state.

Changes in arousal and affective state may influence the scope of attention; with negative affect leading to a narrowing of attentional focus (e.g. Chajut & Algom, 2003; Derryberry & Reed, 1998; Easterbrook, 1959; Fenske & Eastwood, 2003; Gable & Harmon-Jones, 2010; Storbeck, 2013), and positive affect leading to a broadening of focus (e.g. Derryberry & Tucker, 1994; Easterbrook, 1959; Rowe, Hirsh, & Anderson, 2007) – although there is evidence that the link between affect and attention may have some flexibility (Huntsinger, 2012). If the category change (forcing a loss of ASA) induced a negative affective state in participants, this may have precipitated a narrowing of attention that, as discussed above, manifested as an increasingly positive bias.

In each study, two groups were identified *post hoc* based on changes in bias following the forced loss of ASA. We acknowledge that these two groups were identified by inspection of the data and that other groupings or divisions were possible. The aim of this grouping was to illustrate that QASA can reveal discriminable groups in the data. The two groups identified (in both studies) showed clear differences in the relationship between bias and ASA. One group, following the forced loss of ASA, reverted (after the initial drop) to a wider bias. This group showed a recovery of ASA. The other group, that maintained a narrower bias, generally did not regain ASA. Given that reacquiring ASA necessitated identifying new aspects of a category, the efficacy (in this case) of a broader bias in reacquiring ASA is plausible.

In summary, we believe that the QASA approach provides measures of information use, actual SA, and perceived SA based on a theoretically plausible model. The model incorporates aspects of both System 1 and System 2 processing in the selection of information and construction of ASA and is currently unique in this conceptualisation of the building and maintenance of ASA. Information selection is considered to be part of a dynamic cyclical process for building and maintaining ASA. ASA is an emergent, and dynamic, property of this process.

References

Arendtsen, B., Baker, S., Bertels, M., Brookes, D., Catherwood, D., Christiansen, K., . . . Wenarski, G. (2016, November). Firemind: Trialling a new tool for training fire and rescue service decision-making. *International Fire Professional,* 14-17.

Banbury, S., & Tremblay, S. (Eds.). (2004). *A Cognitive Approach to Situation Awareness: Theory and Application*: Ashgate.

Bartlett, F. C. (1932). *Remembering: a study in experimental and social psychology*. Cambridge, England: University Press.

Bender, R., & Lange, S. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal, 316*, 1236-1238.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *Journal of General Psychology, 39*, 15-22.

Bingi, P., Turnipseed, D. L., & Kasper, G. M. (2001). The Best Laid Plans of Mice and Men: The Role of Decision Confidence in Outcome Success. *North American Journal of Psychology, 3*(1).

Burge, R., & Chaparro, A. (2012). *The effects of texting and driving on hazard perception*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Catherwood, D., Edgar, G. K., Nikolla, D., Alford, C., Brookes, D., Baker, S., & White, S. (2014). Mapping brain activity during loss of situation awareness: an EEG investigation of a basis for top-down influence on perception. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 56*(8), 1428-1452.

Catherwood, D., Edgar, G. K., Sallis, G., Medley, A., & Brookes, D. (2012). Fire Alarm or False Alarm?! Decision-making "Bias" of Firefighters in Training Exercises. *International Journal Emergency Services, 1*(2), 135-158.

Chajut, E., & Algom, D. (2003). Selective attention improves under stress: Implications for theories of social cognition. *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology, 85*(2), 231-248.

Chiappe, D. L., Rorie, C., Morgan, C., & Vu, K. P. L. (2014). A Situated approach to the acquisition of shared SA in team contexts. *Theoretical Issues in Ergonomics Science, 15*(1), 69-87.

Chiappe, D. L., Strybel, T. Z., & Vu, K.-P. L. (2011). Mechanisms for the acquisition of situation awareness in situated agents. *Theoretical Issues in Ergonomics Science, iFirst*, 1-23. doi:10.1080/1463922x.2011.611267

Chiappe, D. L., Vu, K.-P. L., & Strybel, T. (2012). Situation Awareness in the NextGen Air Traffic Management System. *International Journal of Human-Computer Interaction, 28*(2), 140-151. doi:10.1080/10447318.2012.634768

Clarke, F. R., Birdsall, T. G., & Tanner, W. P. (1959). Two Types of ROC Curves and Definitions of Parameters. *J. Acoust. Soc. Am., 31*, 629-630.

Collier, S. G., & Follesø, K. (1995). *SACRI: A mesure of situation awareness for nuclear power plant control rooms*. Paper presented at the International conference on experimental analysis and measurement of situation awareness, Daytona Beach, Florida, USA.

Croft, D., Banbury, S., & Thompson, D. (2001). *The development of an implicit situation awareness toolkit*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Derryberry, D., & Reed, M. A. (1998). Anxiety and attentional focusing: trait, state and hemispheric influences. *Personality and Individual Differences, 25*(4), 745-761. doi:10.1016/S0191-8869(98)00117-2

Derryberry, D., & Tucker, D. M. (1994). Motivating the focus of attention. In P. M. Neidenthal & S. Kitayama (Eds.), *The heart's eye: Emotional influences in perception and attention* (pp. 167-196). San Diego, CA: Academic Press.

Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General, 121*(3), 275.

Donaldson, W. (1993). Accuracy of d' and A' as estimates of sensitivity. *Bulletin of the Psychonomic Society, 31*, 271-274.

Durso, F. T., & Dattel, A. R. (2004). SPAM: The Real-Time Assessment of SA. In S. Banbury & S. Tremblay (Eds.), *A Cognitive approach to situation awareness: theory and application*. Aldershot, UK: Ashgate.

Durso, F. T., & Gronlund, D. J. (1999). Situation awareness. In F. Durso (Ed.), *Handbook of Applied Cognition* (pp. 283-314): John Wiley and Sons Ltd.

Durso, F. T., Truitt, T. R., Hackworth, C. A., Crutchfield, J. M., Nikolic, D., Moertl, P. M., . . . Manning, C. A. (1995). Expertise and Chess: A Pilot Study Comparing Situation Awareness Methodologies. In O. Garland & M. R. Endsley (Eds.), *Experimental Analysis* *and Measurement of Situation Awareness* (pp. 295-304). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review, 66*, 183-201.

Edgar, G. K., Catherwood, D., & Grover, R. (2007). *A novel approach to formative assessment: Assessing student learning and attitude to information using signal detection theory*. Paper presented at the ISSOTL Conference, University of NSW, Sydney, Australia.

Edgar, G. K., Catherwood, D., & Melhuish, J. (2009). Seeing the past. Situation awareness is influenced by old information when workload is high. *Perception, 38*(3), 461-462.

Edgar, G. K., Catherwood, D., Sallis, G., Brookes, D., & Medley, A. (2012). "I always know what's going on." Assessing the Relationship between Perceived and Actual Situation Awareness across Different Scenarios. *World Academy of Science, Engineering and Technology, 71*, 1480-1481.

Edgar, G. K., & Edgar, H. E. (2007). Using Signal Detection Theory to Measure Situation Awareness: The Technique, The Tool (QUASA), the TEST, the Way Forward. In M. Cook, J. Noyes, & Y. Masakowski (Eds.), *Decision making in complex environments* (pp. 373-385). Aldershot, UK: Ashgate.

Edgar, G. K., Edgar, H. E., & Curry, M. B. (2003). Using signal detection theory to measure situation awareness in command and control *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (Vol. 47(18), pp. 2019-2023). Denver, Colorado: SAGE Publications.

Edgar, G. K., Smith, A. J., Stone, H. E., Beetham, D. L., & Pritchard, C. (2000). QUASA: QUantifying and Analysing Situational Awareness. *Proceedings of the IMCD People in Digitized Command and Control Symposium, RMCS Shrivenham, UK.*

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *Aerospace and Electronics Conference, 1988. NAECON 1988*, 789-795 vol. 783. doi:10.1109/NAECON.1988.195097

Endsley, M. R. (1993). A survey of situation awareness requirements in air-to-air combat fighters. *The International Journal of Aviation Psychology, 3*(2), 157-168.

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

Endsley, M. R. (2000). Theoretical underpinning of situation awareness: a critical review. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 147-173). Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, M. R. (2015). Situation awareness misconceptions and misunderstandings. *Journal of Cognitive Engineering and Decision Making, 9*(1), 4-32.

Endsley, M. R., & Garland, D. J. (Eds.). (2000). *Situation Awareness Analysis and Measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, M. R., & Rodgers, M. D. (1998). Distribution of attention, situation awarenes, and workload in a passive air traffic control task: implications for operational errors and automation. *Air Traffic Control Quarterly, 6*(1), 21-44.

Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998). *A comparative analysis of SAGAT and SART for evaluations of situation awareness*. Paper presented at the 42nd annual meeting of the Human Factors & Ergonomics Society, Chicago, Il.

Endsley, M. R., & Smith, R. P. (1996). Attention distribution and decision making in tactical air combat. *Human Factors, 38*(2), 232-249.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol., 59*, 255-278.

Fenske, M. J., & Eastwood, J. D. (2003). Modulation of focused attention by faces expressing emotion: evidence from flanker tasks. *Emotion, 3*(4), 327-343.

Flach, J., Mulder, M., & van Paassen, M. M. (2004). The concept of the situation in psychology. In S. Banbury & S. Tremblay (Eds.), *A Cognitive Approach to Situation Awareness: Theory and Application* (pp. 42-60): Ashgate.

Gable, P. A., & Harmon-Jones, E. (2010). The Blues Broaden, but the Nasty Narrows Attentional Consequences of Negative Affects Low and High in Motivational Intensity. *Psychological Science, 21*(2), 211-215.

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 and Review, 10*(4), 843-876.

Gatsoulis, Y., Virk, G. S., & Dehghani-Sanij, A. A. (2010). On the measurement of situation awareness for effective human-robot interaction in teleoperated systems. *Journal of Cognitive Engineering and Decision Making, 4*(1), 69-98. doi:10.1518/155534310X495591

Grant, A. D., & Berg, A. (1948). A behavioral analysis of degree of reinforcement and case of shifting to new responses in a Weigi-type card-sorting problem. *Journal of Experimental Psychology, 38*, 404-411.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*, 424-429.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive psychology, 24*, 411-435.

Gugerty, L. (1997). Situation awareness during driving: Explicit and implicit knowledge in dynamic spatial memory. *Journal of Experimental Psychology: Applied, 3*(1), 42.

Hodos, W. (1970). Nonparametric index of response bias for use in detection and recognition experiments. *Psychological Bulletin, 74*(5), 351-354. doi:10.1037/h0029933

Hogg, D. N., Follesø, K., Strand-Volden, F., & Torralba, B. (1995). Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms. *Ergonomics, 38*(11), 2394-2413.

Huntsinger, J. R. (2012). Does positive affect broaden and negative affect narrow attentional scope? A new answer to an old question. *Journal of Experimental Psychology: General, 141*(4), 595-600.

Jorna, P. G. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology, 34*(2), 237-257.

Kahneman, D. (2011). *Thinking, Fast and Slow*. London, England: Allen Lane.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160.

Lampinen, J. M., Copeland, S. M., & Neuschatz, J. S. (2001). Recollections of things schematic: Room schemas revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1211-1222.

Lau, N., Jamieson, G. A., & Skraaning Jr, G. (2014). Inter-rater reliability of query/probe-based techniques for measuring situation awareness. *Ergonomics, 57*(7), 959-972.

Loftus, E. F. (1979). The malleability of human memory: Information introduced after we view an incident can transform memory. *American Scientist, 67*, 312 320.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and" nonparametric" indexes. *Psychological Bulletin, 107*(3), 401.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: a user's guide*. Cambridge: Cambridge University Press.

Maule, A. J., & Mackie, P. M. (1990). A componential investigation of the effects of deadlines on individual decision making. In K. Borcherding, O. I. Larichev, & D. M. Messick (Eds.), *Contemporary issues in decision making* (pp. 449–461). Amsterdam: Elsevier.

McGowan, A. M., & Banbury, S. (2004). Evaluating interruption-based techniques using embedded measures of driver anticipation. In S. Banbury & S. Tremblay (Eds.), *A Cognitive Approach to Situation Awareness: Theory and Application* (pp. 176-192). Aldershot: Ashgate.

McGuinness, B. (2007). Signal detection theory and the assessment of situation awareness. In M. Cook, J. Noyes, & Y. Masakowski (Eds.), *Decision making in complex environments* (pp. 299-310). Aldershot, UK: Ashgate.

Miller, J. (1996). The sampling distribution of d'. *Perception and Psychophysics, 58*(1), 65-72.

Neisser, U. (1976). *Cognition and reality: principles and implications of cognitive psychology*. San Francisco: W.H. Freeman.

Nikolla, D., Edgar, G., Catherwood, D., & Matthews, T. (2017). Can bottom-up processes of attention be a source of "interference" in situations where top-down control of attention is crucial? . *British Journal of Psychology*, Online. doi:10.1111/bjop.12251

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: detecting the snake in the grass. *Journal of Experimental Psychology: General, 130*(3), 466.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review, 10*(3), 556-569.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ: British Medical Journal, 316*(7139), 1236.

Pezdek, K., Whetstone, T., Reynolds, K., Askari, N., & Dougherty, T. (1989). Memory for real world scenes: The role of consistency with schema expectation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 587 595.

Plant, K. L., & Stanton, N. A. (2013). The explanatory power of Schema Theory: theoretical foundations and future applications in Ergonomics. *Ergonomics, 56*(1), 1-15.

Pollack, I. (1959). On indices of signal and response discriminability. *The Journal of the Acoustical Society of America, 31*(7), 1031-1031.

Pollack, I., & Norman, D. A. (1964). Non-parametric analysis of recognition experiments. *Psychonomic Science, 1*, 125-126.

Pritchett, A. R., & Hansman, R. J. (2000). Use of testable responses for performance-based measurement of situation awareness. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 189-209).

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*(3), 518-535.

Read, J. D., Vokey, J. R., & Hammersley, R. (1990). Changing photos of faces: Effects of exposure duration and photo similarity on recognition and the accuracy confidence relationship. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 870 882.

Reinitz, M. T., & Hannigan, S. L. (2001). Effects of simultaneous stimulus presentation and attention switching on memory conjunction errors. *Journal of memory and language, 44*, 206-219.

Richardson, M., & Ball, L. J. (2009). Internal representations, external representations and ergonomics: towards a theoretical integration. *Theoretical Issues in Ergonomics Science, 10*(4), 335-376.

Rousseau, R., Tremblay, S., Banbury, S., Breton, R., & Guitouni, A. (2010). The role of metacognition in the relationship between objective and subjective measures of situation awareness. *Theoretical Issues in Ergonomics Science, 11*(1/2), 119-130.

Rousseau, R., Tremblay, S., & Breton, R. (2004). Defining and modeling situation awareness: a critical review. In S. Banbury & S. Tremblay (Eds.), *A Cognitive Approach to Situation Awareness: Theory and Application* (pp. 3-21): Ashgate.

Rowe, G., Hirsh, J. B., & Anderson, A. K. (2007). Positive affect increases the breadth of attentional selection. *Proceedings of the National Academy of Sciences of the United States of America, 104*(1), 383-388.

Sætrevik, B. (2012). *A controlled field study of situation awareness measures and heart rate variability in emergency handling teams*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Salmon, P. M., & Stanton, N. A. (2013). Situation awareness and safety: Contribution or confusion? Situation awareness and safety editorial. *Safety Science, 56*(Complete), 1-5.

Salmon, P. M., Stanton, N. A., Walker, G. H., Baber, C., Jenkins, D. P., McMaster, R., & Young, M. S. (2008). What really is going on? Review of situation awareness models for individuals and teams. *Theoretical Issues in Ergonomics Science, 9*(4), 297–323.

Salmon, P. M., Stanton, N. A., & Young, K. L. (2012). Situation awareness on the road: review, theoretical and methodological issues, and future directions. *Theoretical Issues in Ergonomics Science, 13*(4), 472-492.

See, J. E., Warm, J. S., Dember, W. N., & Howe, S. R. (1997). Vigilance and signal detection theory: An empirical evaluation of five measures of response bias. *Human Factors, 39*(1), 14-29.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420.

Smith, K., & Hancock, P. A. (1995). Situation Awareness Is Adaptive, Externally Directed Consciousness. *Human Factors, 37*(1), 137-148.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*(1), 34.

Sohn, Y. W., & Doane, S. M. (2003). Roles of working memory capacity and long-term working memory skill in complex task performance. *Memory & cognition, 31*(3), 458-466.

Sohn, Y. W., & Doane, S. M. (2004). Memory processes of flight situation awareness: Interactive roles of working memory capacity, long-term working memory, and expertise. *Human Factors, 46*(3), 461-475.

Sorensen, L. J., Stanton, N. A., & Banks, A. P. (2011). Back to SA school: contrasting three approaches to situation awareness in the cockpit. *Theoretical Issues in Ergonomics Science, 12*(6), 451-471. doi:10.1080/1463922x.2010.491874

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

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645-665.

Stanton, N. A. (2010). Situation awareness: where have we been, where are we now and where are we going? *Theoretical Issues in Ergonomics Science, 11*(1/2), 1-6.

Stanton, N. A., Chambers, P. R. G., & Piggott, J. (2001). Situational awareness and safety. *Safety Science, 39*(3), 189-204.

Stanton, N. A., Salmon, P., Walker, G., Salas, E., & Hancock, P. (2017). State-of-Science: Situation Awareness in individuals, teams and systems. *Ergonomics*(just-accepted), 1-33. doi:http://dx.doi.org/10.1080/00140139.2017.1278796

Stanton, N. A., Salmon, P. M., Walker, G. H., & Jenkins, D. P. (2010). Is situation awareness all in the mind? *Theoretical Issues in Ergonomics Science, 11*(1/2), 29-40.

Stanton, N. A., Stewart, R., Harris, D., Houghton, R. J., Baber, C., McMaster, R., . . . Green, D. (2006). Distributed situation awareness in dynamic systems: theoretical development and application of an ergonomics methodology. *Ergonomics, 49*(12-13), 1288-1311. doi:10.1080/00140130600612762

Storbeck, J. (2013). Negative affect promotes encoding of and memory for details at the expense of the gist: Affect, encoding, and false memories. *Cognition & Emotion, 27*(5), 1-20.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1397-1410. doi:10.1037/0278-7393.24.6.1397

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin, 99*, 181-198.

Tenney, Y. J., Adams, M. J., Pew, R. W., Huggins, A. W. F., & Rogers, W. H. (1992). *A principled approach to the measurement of situation awareness in commercial aviation* (NASA Contractor Report 4451). Retrieved from http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19920021063.pdf 28th December, 2015

Thomas, J. A. (2008). *Network Science: Observations from the Omni Fusion 2007 Digital Warfighter Exercise Simulation Experiment*. Paper presented at the 13th International Command and Control Research and Technology Symposia (ICCRTS 2008), 13th International Command and Control Research and Technology Symposia (ICCRTS 2008).

Tremblay, S., Breton, R., Vachon, F., & Allen, D. (2012). *Support of collaborative work in battlespace management: Shared (loss) of situation awareness*. Paper presented at the Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2012 IEEE International Multi-Disciplinary Conference on.

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of‚Ä≤, A z, andA‚Äô. *Perception & Psychophysics, 68*(4), 643-654.

Wickens, T. D. (2001). *Elementary Signal Detection Theory*: Oxford University Press.

Woods, D. D., & Sarter, N. B. (2010). Capturing the dynamics of attention control from individual to distributed systems: the shape of models to come. *Theoretical Issues in Ergonomics Science, 11*(1/2), 7-28.