The Validity of Situation Awareness for Performance: A Meta-Analysis

Jonathan Z. Bakdash[a,b*], Laura R. Marusich[c], Katherine R. Cox[d], Michael N. Geuss[e], Erin G. Zaroukian[f], and Katelyn M. Morris[g]

[a] *U.S. Army Combat Capabilities Development Command Army Research Laboratory South at the University of Texas at Dallas, Richardson, TX, United States, ORCiD: 0000-0002-1409-4779*

[b] *Department of Psychology and Special Education, Texas A&M—Commerce, Commerce, TX, United States*

[c] *U.S. Army Combat Capabilities Development Command Army Research Laboratory South at the University of Texas at Arlington, Arlington, TX, United States, ORCiD: 0000-0002-3524-6110*

[d] *U.S. Army Combat Capabilities Development Command Army Research Laboratory, Human Research and Engineering Directorate, Aberdeen Proving Ground, MD, United States, ORCiD: 0000-0002-8351-751X*

[e] *U.S. Army Combat Capabilities Development Command Army Research Laboratory, Human Research and Engineering Directorate, Aberdeen Proving Ground, MD, United States, ORCiD: 0000-0002-2611-7544*

[f] *U.S. Army Combat Capabilities Development Command Army Research Laboratory, Computational and Information Science Directorate, Aberdeen Proving Ground, MD, United States, ORCiD: 0000-0002-1381-085X*

[g] *U.S. Army Combat Capabilities Development Command Army Research Laboratory South at the University of Texas at Arlington, Arlington, TX, United States, ORCiD*: 0000-0003-2396-8578

* Corresponding author

Email: *jonathan.z.bakdash.civ@mail.mil*

This is a preprint, the manuscript was accepted by *Theoretical Issues in Ergonomic Science* on April 11, 2021. There may be minor copy-editing changes in the final journal version.

- **Version 7** (dated May, 2022): Revised supplemental material, fixed errors with detailed only
- Version 6 (dated January 19, 2022): *See revision history below* (see first page of supp for details)
- We submitted the ~~final~~ version (~~this preprint,~~ version 5) to the journal on April 16, 2021.
- Previously, a minor revision of the paper was submitted to the journal on April 5, 2021.

Revision History

Minor changes in version 6 of the preprint: Fixed errors in CI calculations for I^2 and removed z-transform for tau estimates (no impact on results). The CI bounds around I^2 are now narrower: Correct values I^2 = 60.99%, 95% CI [53.00%, 69.16%]. Thanks to Wolfgang Viechtbauer for finding these mistakes. *Note these values are slightly wider in the published paper.*

Minor changes in version 5 of the preprint: To improve clarify, minor edits to the abstract and added a quote about the potential for SA-performance causal circularity from Flach (1995). Fixed typos in the paper and appendices and added the missing legend to Figure F4, Appendix F.

Previous versions
We submitted a revised version (preprint version 4) of this manuscript to a journal on February 26, 2021. The previous version (preprint version 3) was initially submitted to the same journal on December 22, 2020.

Major changes in version 4 of the preprint: We added additional information and analyses in response to reviewers.

- Item 1: Conclusions too strong? (R1). In the response letter, we describe how meta-analysis typically permits stronger and more generalizable inferences than narrative reviews and individual studies.
- Item 2: Value of meta-analysis in general? (R2). Note that Bakdash et al. 2020 is a critique of an existing synthesis, but not a meta-analysis in its own right. The critique points out the need for a correctly-implemented, quantitative meta-analysis based on systematic review, a gap filled by the current work. We also clarify the scientific value of meta-analysis in the response letter (also see Item 1).
- Item 3: *Clarify prediction interval* (R3). We revised the explanation and added equations.
- Item 4: *Drawn relevant connections* (R2 and R3). We have drawn connections in the Discussion to relevant literature/fields suggested by R2 and R3

We also had several minor corrections to the data: we corrected the sample sizes in two papers (Albina, 2019 [N = 30 to N = 28]; Onal et al. 2014 [N = 94 to N = 95]), added coding for a second experiment in one paper where all effects were previously coded as a single experiment (Lafond et al. 2012), and excluded one paper that does not appear to have assessed SA-performance associations (Dong et al. 2003). The meta-analysis now has 678 effects from 77 papers.

Major changes in version 3 of the preprint: New systematic review from scratch, better detection and exclusion of overfit results (excessive degrees of freedom), meta-analysis now has 682 effects from 78 papers, new draw method for estimating ghost results, and analyses of proportions of effects below three meaningful thresholds.

The validity of situation awareness for performance: A meta-analysis

Situation awareness (SA) is a widely used cognitive construct in human factors, often theoretically posited to be a critical causal factor and/or construct for performance. However, there are concerns that SA may not sufficiently capture the psychological processes underlying performance. We address these conflicting perspectives using meta-analysis to evaluate the patterns of associations among SA-performance effect sizes. Specifically, we focus on the validity of SA for performance—how well SA captures the relevant psychological processes for task performance. In our systematic review of the empirical literature, we coded associations of ten unique measures of SA with performance: 678 effects from 77 papers. The meta-analytic means for SA measures were all of approximately medium or lower effect sizes. The overall mean effect, while significant, was also limited in magnitude ($r = 0.26$, $p < 0.001$). Furthermore, there was high unexplained systematic variation with an enormous plausible range for individual effects ($r = -0.15$ to 0.60). The results indicate that SA's validity for performance tends to be, on average, weak with large variations among effects. Interventions that improve SA may not correspond to meaningful improvements in task performance, and it may be appropriate to revise major theories of SA.

Keywords: situation awareness; performance; validity; meta-analysis; systematic review

Relevance to human factors / ergonomics theory

In this work, we used meta-analysis to quantitatively synthesise nearly three decades of previously published papers with associations among situation awareness (SA) and performance. We found limited validity: correlations between measures of SA and performance tended to be of medium or lower effect size, and individual effects exhibited high systematic variability. Overall results were inconsistent with theories positing that SA is typically a meaningful probabilistic factor for performance, let alone that SA is generally critical or fundamental to performance. Theories of SA may need to be revised.

Introduction

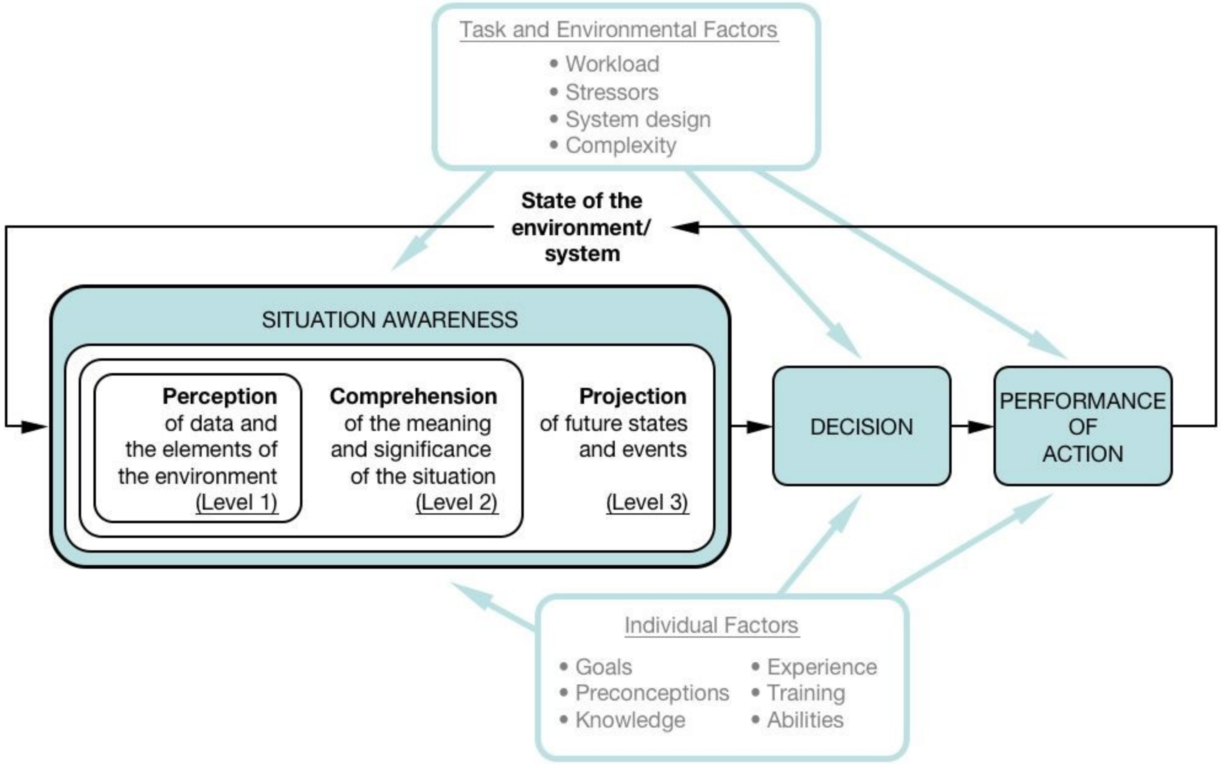
Situation awareness (SA) is a ubiquitous concept and construct in the military and other complex, dynamic, and safety-critical environments, such as aviation and health care (Durso and Sethumadhavan 2008; Endsley 1995a; Endsley 1995b; Endsley 2000a; Endsley 2000b; Endsley 2015a; Endsley 2015b; Parasuraman, Sheridan, and Wickens 2008; Salmon et al. 2008; Stanton, Chambers, and Piggott 2001; Tenney and Pew 2006; Wickens 2008). SA is defined generally as 'knowing what is going on around you' (Endsley 2000b, 4), and more formally as '… the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future' (Endsley 1988, 792). SA has enormous face validity, as it is an intuitive concept (Jones 2015).

As implied by the previous definitions and its face validity, SA is theorized to be critical to human performance Endsley (2015a), a relationship illustrated in Figure 1. In this context, performance is typically defined using task-specific measures, including objective (e.g., the number of collisions in a simulated driving task) as well as subjective (e.g., subject matter expert ratings of conflict management in a simulated air traffic control task) measures. The nature of the critical relationship between SA and performance has been described in different ways. For example, SA is posited to be:

(1) A probabilistic factor for performance: 'Good SA … will increase the probability of good performance but cannot necessarily guarantee it' (Endsley 1995a, 41)
(2) A fundamental construct for performance (Endsley et al. 2003), a basis for performance (Endsley and Connors 2008), and a precursor to performance (Endsley 2000b).

SA has also been called a valuable construct "…in understanding and predicting human-system performance in complex systems" (Parasuraman, Sheridan, and Wickens 2008, 140). Consequently, improving SA through design and training is posited to increase performance (Endsley and Jones 2011). Alternatively, others have raised concerns that definitions of SA and its relationship to performance may be too circular and vague to be useful (Billings 1996; Flach 1995; Sarter and Woods 1991; Dekker and Hollnagel 2004; van Winsen and Dekker 2015), but also see (Endsley 2015a). The circularity issue is summarized by Flach (1995): "How does one know that SA was lost? Because the human responded inappropriately. Why did the human respond inappropriately? Because SA was lost." (151).

Figure 1. A widely used model of SA. This model posits a direct link from SA to decisions and performance. This figure was drawn by Lankton (2019), adapted from Figure 1 in Endsley (1995b).


Task and Environmental Factors
• Workload
• Stressors
• System design
• Complexity
State of the environment/ system
SITUATION AWARENESS
Perception of data and the elements of the environment (Level 1)
Comprehension of the meaning and significance of the situation (Level 2)
Projection of future states and events (Level 3)
DECISION
PERFORMANCE OF ACTION
Individual Factors
• Goals
• Preconceptions
• Knowledge
• Experience
• Training
• Abilities

Related to concerns about circularity, Dekker and Hollnagel (2004) offer perhaps the strongest conceptual criticism of SA. They assert that SA is a generic descriptive label because it lacks the appropriate causal psychological mechanisms relevant to performance (also see Billings [1996] and Flach [1995]). In other words, SA does not have meaningful probabilistic associations with performance. Thus, they deem SA a 'folk model' rather than a scientific construct. However, others have disputed this characterization of SA, arguing that SA is supported by theory and empirical evidence, including its relationship with performance as well as its useful applications for training and design (Endsley 2015a; Endsley 2015b; Endsley 2015c; Parasuraman, Sheridan, and Wickens 2008; Wickens 2008).

The diverging perspectives on SA fundamentally address construct validity: What do measures of SA actually assess? (Cronbach and Meehl 1955; Strauss and Smith 2009). Specifically, to what degree do (measures of) SA represent theoretical cognitive processes, strategies, knowledge, and other general and task-specific psychological processes involved in performance? This type of construct validity is construct representation: the psychological mechanisms or processes underlying task performance (Embretson (Whitely) 1983; Strauss and Smith 2009).

One way to address the contradictory views summarised above is to evaluate the validity of SA for predicting or representing performance using correlations among SA measures and corresponding measures of human performance. That is, inferring the probabilistic links among measures of SA and task performance using their associations. Here, we use meta-analysis to quantitatively synthesise SA-performance correlations reported in previous empirical work. Meta-analysis is ideal for addressing specific research questions using quantitative synthesis of evidence in a body of relevant research (Borenstein et al. 2009; Cooper, Hedges, and Valentine 2009; Gurevitch et al. 2018).

There are many narrative reviews and theoretical papers on SA (Durso and Sethumadhavan 2008; Endsley 1995a; Endsley 1995b; Endsley 2000b; Endsley 2015a; Endsley 2015b; Salmon et al. 2008; Stanton, Chambers, and Piggott 2001; Tenney and Pew 2006; Wickens 2008), but only a handful of quantitative syntheses. One was a meta-analysis in health care that assessed the impact of different training methods on SA (Walshe et al. 2019). Similarly, two quantitative syntheses examined the sensitivity of measures of SA to different interface design manipulations (Vidulich 2000; Endsley 2019). Other work by Endsley (2020) synthesised correlations for objective and subjective measures of SA and concluded they were dissociable types of SA. However, results were interpreted only qualitatively rather than compared using inferential statistics.

Most relevant here, the quantitative synthesis by Endsley (2019) found that objective SA measures were strongly correlated with performance, with medium to large pooled effects (using generic conventions from Cohen [1998]: a small effect is r = 0.10, a medium effect is r = 0.30, and a large effect is r = 0.50) ranging from r = 0.41 to 0.53; however, *only* effect sizes reaching significance were included in that synthesis. This type of biased inclusion criterion runs counter to the reason meta-analysis was originally developed (Glass 2015). Furthermore, selecting results based on p-values is circular and will guarantee inflated effect sizes (Bishop 2019; Bishop 2020; Gelman and Loken 2013; Ioannidis et al. 2014; Kriegeskorte et al. 2009; Simmons, Nelson, and Simonsohn 2011; Vosgerau et al. 2019; Vul et al. 2009; Wicherts et al. 2016). Analysis of all effects as-reported, instead of only significance-filtered effects, shows that the filtered mean effect in Endsley (2019) was overestimated by 1.56 times, or 56%; for a detailed critique see (Bakdash, Marusich, Kenworthy, et al. 2020).

Given that SA is widely used and theorised to be critical to human performance, there is a clear need to conduct a quantitative meta-analysis using recommended techniques to assess the relationships among SA measures and task performance. In this paper, we describe the results of such a meta-analysis in which we conducted a systematic review of the literature and took into account all relevant results, as-reported and inferred from included papers, regardless of statistical significance. We evaluated meta-analytic mean effects for correlations from different SA measures, and we synthesised the overall mean meta-analytic effect for SA-performance associations across all SA measure methods. We also quantified the meta-analytic heterogeneity, which indicates the amount of systematic inconsistency among individual effects (Borenstein et al. 2009; Higgins et al. 2003). Finally, we evaluated the distribution of individual effects by quantifying the proportion of effects below/above three relevant thresholds of interest.

Methods

We followed the majority of relevant elements in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist (Moher et al. 2009), see Appendix A in the online supplemental material. Data cleaning and analyses were performed using the statistical programming language *R* (R Core Team 2021); Appendix B lists all *R* packages used. The methods and results are reproducible, see Bakdash, Marusich, Cox, et al. (2021a; 2021b).

Systematic Review

We systematically reviewed the literature across five indexes/databases: DTIC, PsycINFO, SAGE, SCOPUS, and Web of Science. We also conducted supplemental searches via Google Scholar, using the Publish or Perish software (Harzing 2020) to avoid user-specific recommendations; this was necessary because Google Scholar is a search engine, not an index (Gusenbauer 2019). See the supplemental material, Appendix C, for a complete list of databases and search terms used. This work was approved as not human subjects research by the U.S. Army Research Laboratory's Institutional Review Board (ARL 17-293 and 17-293A).

Documents meeting all three of the following criteria were included in the meta-analysis:

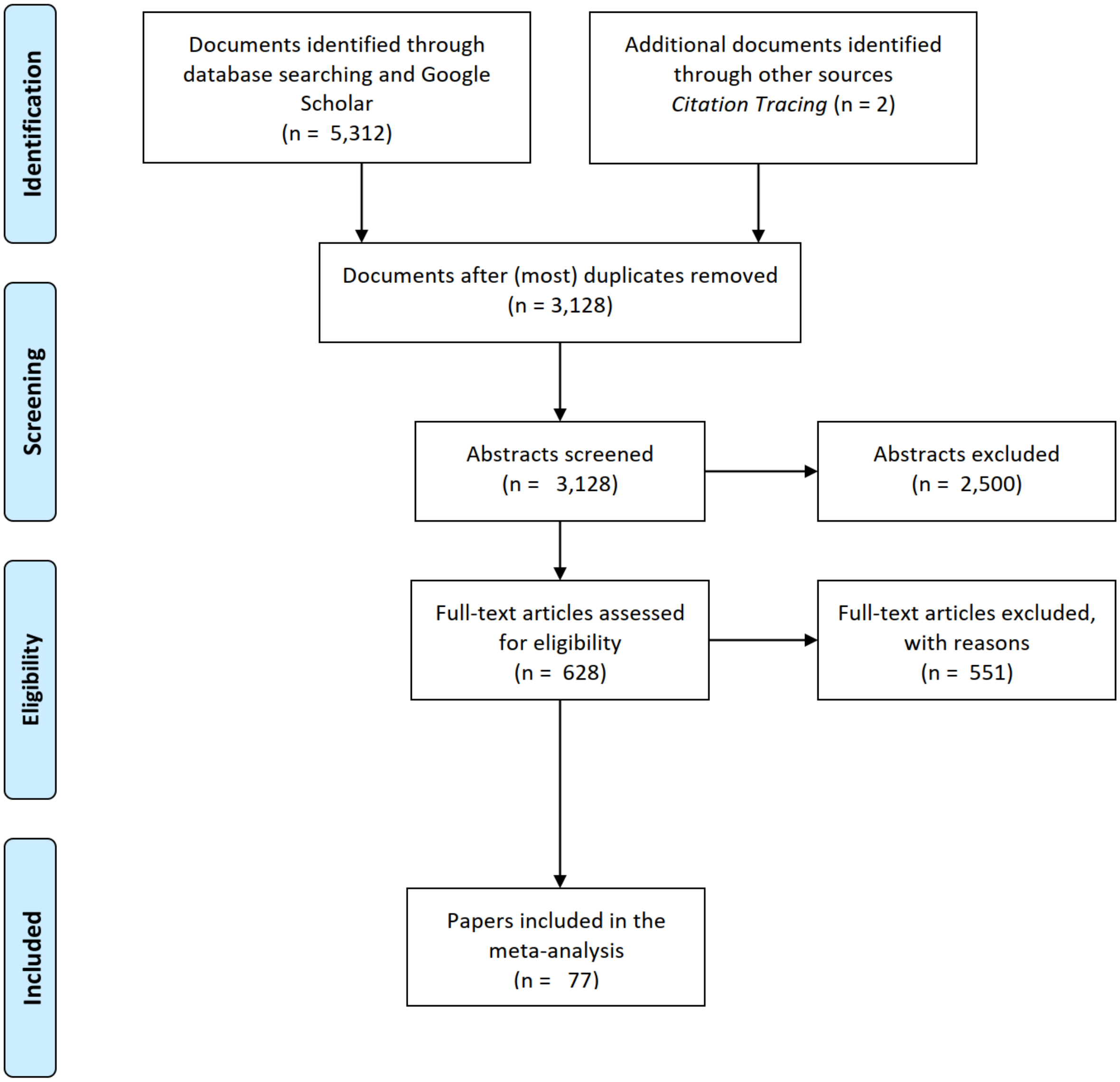
(1) Journal articles, conference proceedings, and technical reports published or in press on or before January 14, 2020.

(2) Empirical research that reported inferential tests for the relationship(s) among (measures of) SA and objective and/or subjective human performance.

(3) Research using state measures of SA (i.e., *not* measures including general cognitive abilities or traits, workload, or similar constructs) at the individual-level only.

The literature searches yielded a total of 5,314 documents; the inclusion/exclusion process is summarized in Figure 2. After removing most duplicate documents, we screened the abstracts of the remaining 3128 documents. Each abstract was coded as 'include', 'exclude', or 'maybe' by a minimum of 3 independent raters. All abstracts (except those from DTIC) were screened using Rayyan (Ouzzani et al. 2016). Inter-rater reliability for abstract screening was moderate, *ICC*(3,1) = 0.63, 95% CI [0.61, 0.65].

Figure 2. Summary of systematic review steps. Flowchart of inclusion/exclusion for papers from the system review. This figure is adapted from the PRISMA flow diagram (Moher et al. 2009).


Identification
Screening
Eligibility
Included
Documents identified through database searching and Google Scholar (n = 5,312)
Additional documents identified through other sources *Citation Tracing* (n = 2)
Documents after (most) duplicates removed (n = 3,128)
Abstracts screened (n = 3,128)
Abstracts excluded (n = 2,500)
Full-text articles assessed for eligibility (n = 628)
Full-text articles excluded, with reasons (n = 551)
Papers included in the meta-analysis (n = 77)

We conducted full-text review for papers with abstracts that received at least one rating of ‘include’ or two ratings of ‘maybe’ (20.08%, or 628 out of 3128). The full-text for each paper was reviewed by at least two raters. Papers were included if it was clear that one or more results assessed the relationship between SA and performance. Specific results were coded by one or more raters and checked by at least one additional rater. Disagreements among raters for inclusion of papers and specific results were resolved through consensus. Of the reviewed fulltext documents, 77 (12.26%) met the inclusion criteria (see Appendix D). For the 551 excluded documents, the reason(s) for excluding a document were agreed upon by at least two raters (see Appendix C for details). The top three reasons a paper was excluded during the full-text review were:

(1) No reported association between SA and performance (290 documents),

(2) No measure of task performance (67 documents), and

(3) SA was a label for performance or no measure of SA (56 documents).

Repeated Measures Designs and Overfitting

All results included in the meta-analysis, even those with a repeated measures experimental design, analysed relationships among SA and performance by treating the data as *between-participants* through either, 1) averaging, or 2) separate analyses to avoid overfitting. For example, Jipp and Ackerman (2016) assessed SA and performance 12 times for each participant (three levels of automation x four trials each) in a simulated air traffic control task. The authors correlated the average SA and average performance by participant (across the 12 trials). Conversely, O'Brien and O'Hare (2007) evaluated associations separately for SAGAT (Level 1 and Level 2/3) with corresponding measures of performance (such as hand-off error, missed approaches, and total errors).

A substantial number of papers (26 documents) reported statistical results from incorrectly modelling repeated measures data as independent observations/participants. Such overfit results, an excessive number of free model parameters (Babyak 2004), are typically uninterpretable because variance is often underestimated but it can also be overestimated (Kenny and Judd 1986). See Figure 6 in Bakdash and Marusich (2017) for a visualization of overfitting and also see Aarts et al. (2014) for an overview of multi-level modeling to address dependencies in data.

Overfitting was detected by excessive degrees of freedom in reported statistics. For example, a paper with a sample of *N* = 13, assessed at four different time points, reported correlations with 50 degrees of freedom, implying a sample size of 52 individual participants. Papers that only reported overfit results were completely excluded from our analyses. In three papers with a combination of overfit and correctly (non-overfit) reported results, we excluded only the overfit results.

Coding Effects and SA Measures

We coded a total of 678 effects from the 77 included papers. Three papers contained multiple experiments, and several papers re-analysed the same experimental data; in effect, there were 79 unique experiments (i.e., different samples of participants) used as 'papers' in the analyses.

Of the 678 included effects, 402 were reported in detail (i.e., exact *p*-value and/or effect size) whereas 276 were stated or implied non-significant results that were not reported in detail (see **Coding Ghost Results**). We used degrees of freedom to accurately determine the number of participants in each analysis (e.g., excluded participants or missing data were not always documented). The calculated sample sizes using the degrees of freedom were limited (median *N* = 30): see Appendix E.

Note the vast majority of included 'papers' or experiments (64 out of 79) reported more than one effect size or correlation for SA and performance: see Appendix E for a graph. Multiple effect sizes arose from different measures of SA (multivariate correlated with different measures of performance, one measure of SA assessed multiple times and correlated with performance, and/or subscales from the same SA measure [e.g., Levels 1, 2, and 3 of the SAGAT] correlated with performance). The number of effects, which is the number of separate SA-performance correlations, is denoted by *k*.

We coded 10 different measures of SA in the dataset; see Table 1 for a summary and Appendix C and the data dictionary for details. Note that there are more than 79 entries in Table 1 due to multiple experiments assessing more than one measure of SA (see Appendix D). The Other measure encompasses a variety of SA measures that occurred too infrequently in the dataset to be coded as a unique measure for meta-analytic model convergence; each unique SA measure was used in at least two papers with a total $k \geq 10$.

Table 1. Summary of SA measures included in the meta-analysis. The table has 10 different SA measures with their median sample size, number of papers, and number of effects.

SA Measure	**Median Sample Size (*N*)**	**Number of Papers**	**Number of Effects (*k*)**
Direct-SR (Self Report)	16	2	12
Explicit Recall	16	4	63
General Knowledge	20	14	107
Mission Awareness Rating Scale (MARS)	35	4	30
Situational Awareness Behaviourally Anchored Rating Scale (SABARS)	40	3	24
Situation Awareness Global Assessment Technique (SAGAT)	20	34	170
Situational Awareness Rating Scale (SARS)	40	2	18

SA Measure	Median Sample Size (*N*)	Number of Papers	Number of Effects (*k*)
Situation Awareness Rating Technique (SART)	21	17	84
Situation Presence Assessment Measure (SPAM)	50	14	127
Other	34	12	43

At least two coders independently determined the measure of SA, sample size, degrees of freedom, and effect size. Effect sizes that were not reported as a Pearson correlation coefficient (*r*) were converted to *r* values. We also coded assessment methods (e.g., freeze probe or real-time probe) for SA measures, adapted from Salmon et al. (2006), see Appendix C.

Recoding effect size sign

In some studies, higher positive values indicated better performance (e.g., accuracy, number of tasks completed), whereas in others, higher positive values indicated worse performance (e.g., response times, number of misses). The same was true for different measures of SA. Consequently, the sign of correlations between performance and SA had different meanings across the dataset. To resolve these discrepancies, we recoded the sign of effects so that positive correlations always indicated that better performance was associated with higher SA, and negative correlations indicated better performance associated with lower SA. In total, we recoded the sign of 119 effect sizes in 31 papers.

Coding ghost results

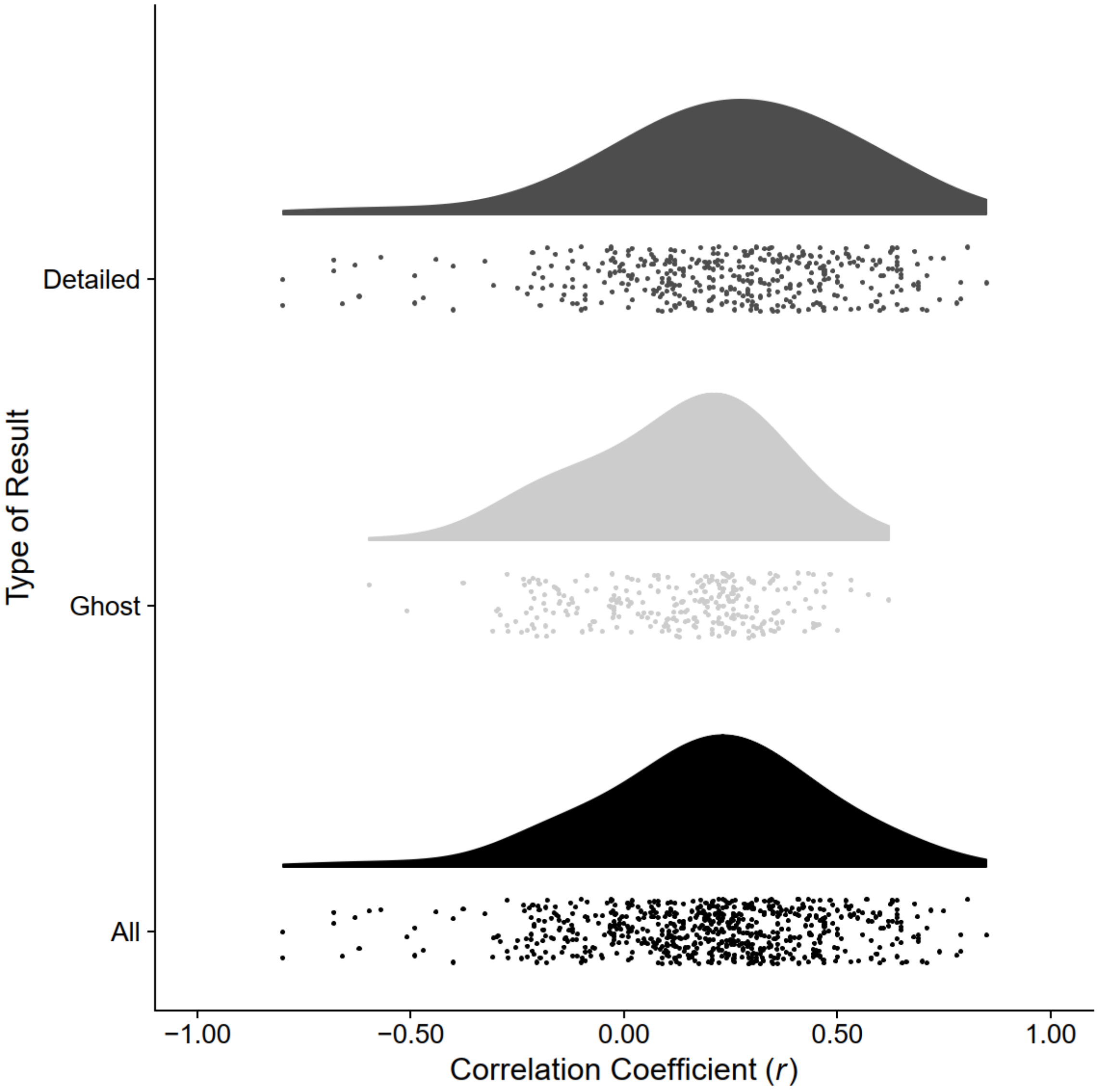
For results that did not reach statistical significance, often no *p*-value or effect size was reported; we refer to these cases as ghost results, adapting the concept and terminology from Bishop and

Thompson (2016). We do not intend to criticize past work by noting their presence; complete and detailed reporting of ghost results was not required in the past and may have even been discouraged by some editors and reviewers. Selective reporting is still a common issue (Bishop 2019; 2020); nevertheless, these ghost results should be taken into account in the meta-analysis, otherwise, results could be positively biased toward the statistically significant and larger effect sizes more likely to be reported in detail.

We coded two types of ghost results that were encountered in the included papers. For the first type (148 effects from 14 papers), the authors either explicitly stated: 1) that a *particular* result was analysed and found to be non-significant, without providing a specific *p*-value and effect size, or 2) that only significant results were reported (and/or non-significant results were *not* reported). The second type of ghost result (128 effects from 14 papers) was implied by specific patterns of omissions in reporting with no direct statements about what results were reported or not reported. Detailed examples of ghost results are provided in the data dictionary, see Bakdash, Marusich, Cox, et al. (2021a).

In order to include these ghost results in the meta-analysis, we imputed *p*-values and effect sizes using the distribution of non-significant effects that were reported in detail (see Figure 3). For each ghost result, we randomly drew, with replacement, a *p*-value (and sign) from the distribution of non-significant results reported in detail, then converted it to an effect size with corresponding variance calculated using the appropriate sample size. We caution that this method of addressing ghost results is likely conservative. The random sampling method used data reported in detail, which is unlikely to be representative of selectively omitted results (see **Limitations**). In addition, there were likely undetected ghost results that were neither stated nor implied. Note that because of the prevalence of small sample sizes there were ghost results that, counter-intuitively, had non-trivial effects.

Figure 3. Raincloud plots of effects by type of result. The smoothed distribution of effects is represented by the ‘cloud’ at the top of each result, the ‘rain’ below consists of dots representing individual effects for each result (detailed, ghost, and all). Dots were randomly vertically jittered to improve visibility.



Effect size transformation

Given the number of papers with small samples, we applied the Fisher *r*-to-*z* transformation to help stabilize variance for estimating meta-analytic mean effects (Cooper, Hedges, and Valentine 2009). The transformation was reversed (*z*-to-*r*) for presented results.

Results

We used multilevel modelling to account for dependencies among effects (repeated/multivariate assessments of SA and performance using the same sample of participants), matching the structure and properties of our dataset as closely as possible (Cooper, Hedges, and Valentine 2009; Scammacca, Roberts, and Stuebing 2014; Konstantopoulos 2011; Viechtbauer 2010). Effect sizes were nested within the grouping variable of 'paper.' Random effect meta-analysis was used to model varying true effects (Borenstein et al. 2009; Cooper, Hedges, and Valentine 2009).

Mean effects and their uncertainty were estimated using cluster robust variance estimation (CRVE), specifying paper as the cluster unit for the small sample adjustment using the CR1p as the specific CRVE estimator. CRVE is a small sample size correction which addresses the dependencies among effects originating from the same sample/paper without requiring knowledge of the exact covariance structure for the sampling error among effects (Hedges, Tipton, and Johnson 2010; Imbens and Kolesar 2016; Pustejovsky and Tipton 2018; Tipton 2015). See Appendix F for additional meta-analytic models using an alternative estimator.

Confidence intervals (CIs) for mean effects were estimated with the Knapp-Hartung (2003) method using a *t*-distribution. This method is recommended for all random effects meta-analysis, particularly if heterogeneity is present (IntHout, Ioannidis, and Borm 2014).

Systematic uncertainty for all effects was quantified using heterogeneity in a prediction interval (PI).

The confidence interval for the overall meta-analytic effect does not necessarily capture how individual effects are distributed because it is an estimate of the pooled or mean effect; this does not include heterogeneity. To estimate the plausible range of *individual* effects, we used the prediction interval which includes heterogeneity (systematic variation among effects).

The equations for the PI and CI follow (adapted from Borenstein, 2009):

$$Prediction\ Interval = M^{*} \pm t\sqrt{\tau^{2} + \ Variance_{M}}$$

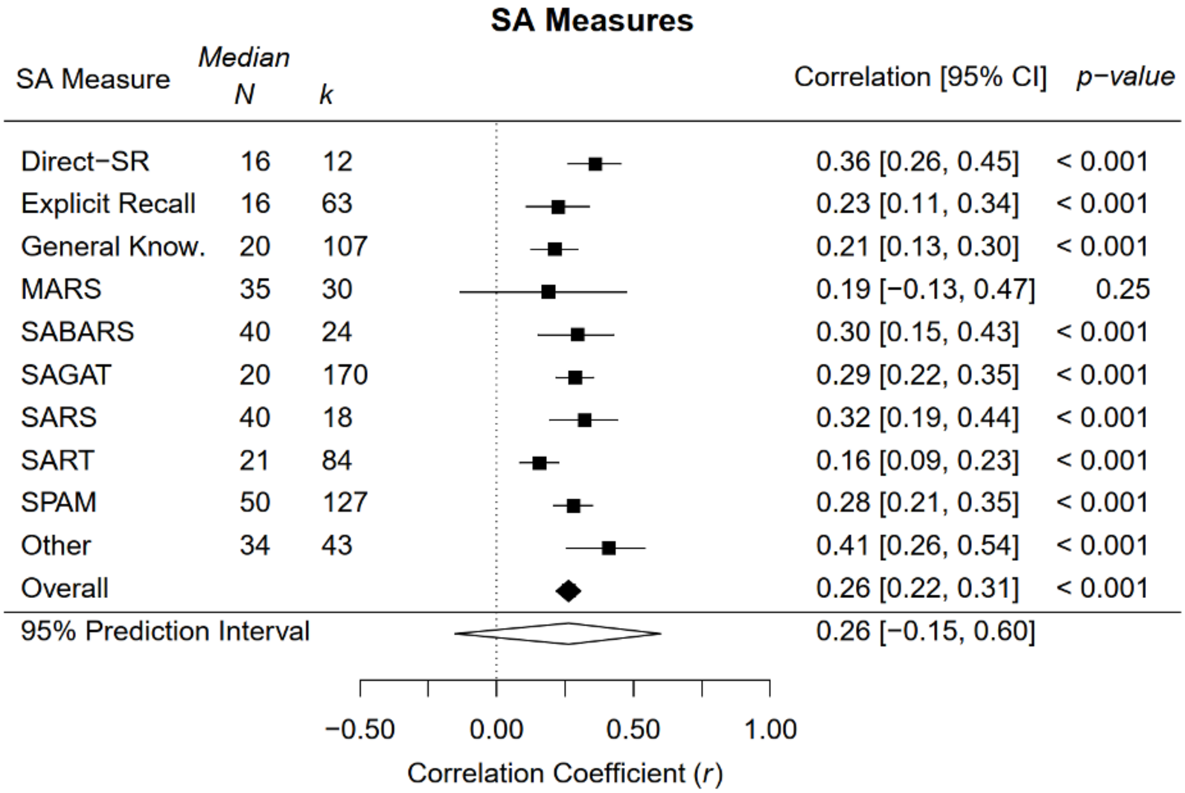
$$Confidence\ Interval = M^{*} \pm t\sqrt{\ Variance_{M}}$$

*M** is the overall meta-analytic mean effect, *t* is the *t*-distribution calculated using the meta-analytic model degrees of freedom (it can be approximated using *z* as 1.96), τ^{2} is the variance of the overall heterogeneity, and $Variance_{M}$ is the variance of the overall meta-analytic effect. When heterogeneity is estimated to be zero, the CI and PI are equivalent.

SA measures

Figure 4 shows the results of the multilevel meta-analytic model that includes SA measure type as a moderator. This forest plot depicts the estimated mean correlations (transformed back from *z* to *r* for ease of interpretation) between SA and performance for each SA measure (with CIs), as well as the estimated mean overall effect (with CI), and the PI for the distribution of individual effects.

Figure 4. Forest plot for the validity of SA measures and performance. Each black square represents the mean estimated meta-analytic effect for each SA measure and the bars depict the width of the 95% CI. The overall mean effect and its 95% CI (plausible distribution of mean effects) is represented by the black diamond. The mean and 95% PI of individual effects is represented by the white diamond. The PI is a Bayesian method; it does not have a frequentist *p*-value.



Nine out of ten SA measures, all but MARS, had mean effects significantly greater than zero; all were medium or smaller in magnitude. The magnitude of the overall mean effect was less than medium ($r = 0.26$) but still significantly greater than zero, 95% CI [0.22, 0.31]. However, the wide coverage of the PI indicates large heterogeneity, 95% PI [-0.15, 0.60], see below.

Meta-analytic heterogeneity

In addition to estimating mean effects, we also quantified heterogeneity, which is the inconsistency or dispersion of effects not due to random error (Borenstein et al. 2009; Cooper, Hedges, and Valentine 2009). Here, heterogeneity is always meta-analytic so it refers to systematic or non-random variance among effects. We assessed total heterogeneity (represented in the PI), as well as within- and between-paper heterogeneity (Table 2). We used the *Q*-test for heterogeneity to test if predictors from the meta-analytic model shared a single common true effect size (Borenstein et al. 2009). There was clear evidence for systematic variations in effect sizes: $\chi^2(668) = 1238.11$, $p < 0.001$.

Table 2. Meta-analytic heterogeneity for the overall model. Values and confidence intervals for several estimates of heterogeneity.

Source	Estimated Parameter [95% CI]
$\hat{\tau}$ = Standard deviation of true effects (due to total heterogeneity)~~, interpret as *r* value~~	0.21 [0.18, 0.25]
$\widehat{\tau_1}$ = Standard deviation of true effects (due to between-paper heterogeneity)~~, interpret as *r* value~~	0.17 [0.14, 0.22]
$\widehat{\tau_2}$ = Standard deviation of true effects (due to within-paper heterogeneity)~~, interpret as *r* value~~	0.12 [0.10, 0.14]
I^2 = Index of dispersion: Variance due to heterogeneity relative to total variance	60.99% [53.00%, 69.16%] ~~[49.59%, 70.94%]~~

The standard deviation for total heterogeneity ($\hat{\tau} = 0.21$) neared the overall mean effect ($r = 0.26$), and there was substantial heterogeneity both between and within-papers. We used the I^2 statistic to determine '…what proportion of the observed variance reflects real differences in effect sizes?' (Borenstein et al. 2009, 117). Sixty percent of the estimated variance was due to (total) non-random variations in true effects. Although meta-analytic heterogeneity could not be evaluated with CRVE, total heterogeneity was approximately the same under different assumptions (see the next section and Appendix F).

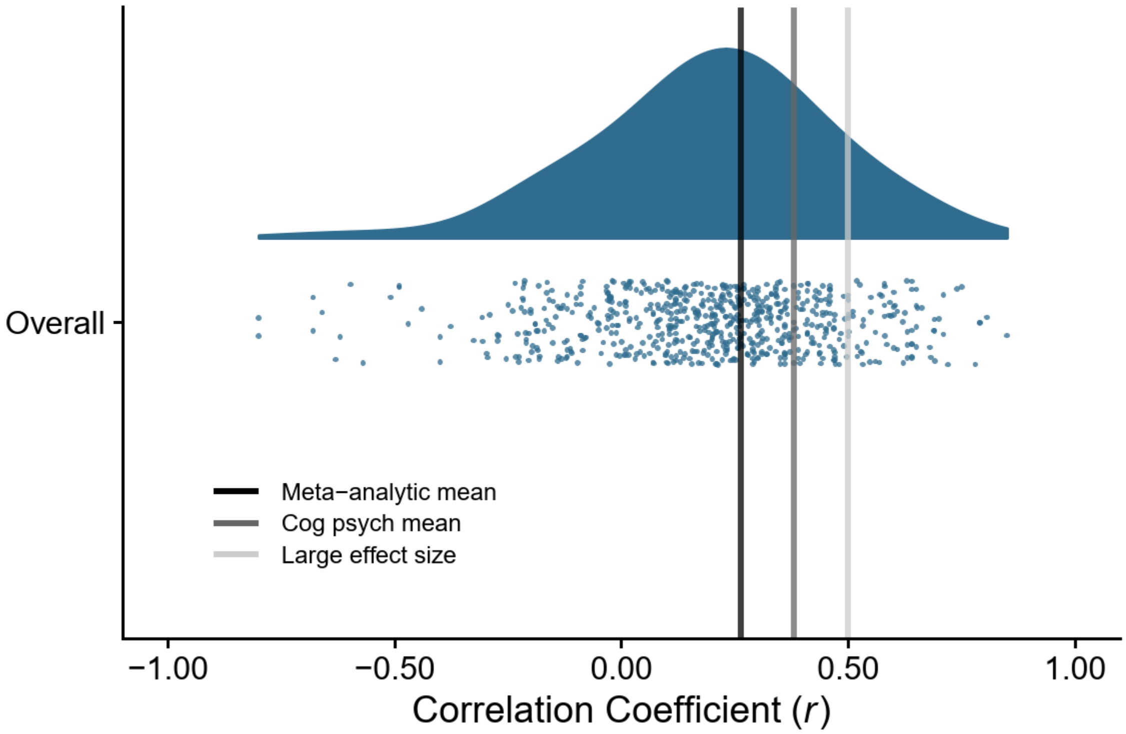
Proportions of effects

Because meta-analytic means might be misleading for non-normally distributed effects and/or highly heterogeneous effects, we also quantified the proportion of effects below/above three meaningful thresholds (Mathur and VanderWeele, 2019, 2020a); for details about how proportions were estimated for non-independent effects see Bakdash et al. (2020) and Mathur and VanderWeele (2020b). These three thresholds were:

1) The estimated overall meta-analytic mean (r = 0.26),
2) The typical mean effect size in cognitive psychology (r = 0.38; Kühberger, Fritz, and Scherndl 2014), and
3) A large effect size (r = 0.50) approximately where SA is interpreted as "highly predictive of performance" (Endsley 2019, 11).

We visualize the distribution of all effects below/above these three thresholds using a raincloud plot (Allen et al. 2019), see Figure 5. These plots show some non-normality in the dataset of all untransformed effects. It is apparent from the figure that the vast majority of effects are below the less stringent threshold of a 'typical' effect in cognitive psychology and even more so for the threshold of a large effect size. Next, we provide statistical evaluation for proportions of effects below/above these thresholds.

Figure 5. Raincloud plot of individual effects and key thresholds. The smoothed distribution of effects is represented by the ‘cloud’ at the top, the ‘rain’ below consists of dots representing individual effects. Dots were randomly vertically jittered to improve visibility. The three thresholds are depicted by vertical lines: the estimated meta-analytic mean is shown with vertical black line and the other two thresholds are shown as vertical grey lines.



Almost two-thirds of all effects (65% or 441 out of 678) were below the overall meta-analytic mean. This finding of a larger-than-50% proportion of effects below indicates that the overall meta-analytic mean was somewhat overestimated. In general, if the meta-analytic mean were perfectly representative, and the underlying data were normally distributed, 50% of effects would be below the mean and 50% of effects above it.

In addition, most effects (86%, or 585 out of 678) were below a representative effect size from cognitive psychology (*r* = 0.38). Likewise, the vast majority of effects (96%, or 651 out of

678) were below the threshold of a large effect size (r = 0.50). As we describe in the **Discussion**, these comparisons to meaningful thresholds provide some perspective on the strength of relationships among SA and performance in the literature.

Additional analyses

We performed a number of additional analyses exploring results in more detail and under different assumptions. Results tended to align with those presented in the paper, demonstrating that the results were generally robust (see Appendix F). First, we performed an additional meta-analysis using SA assessment techniques as a moderator. Second, we used post-hoc tests to compare meta-analytic means among SA measures. There were a small number of significant differences (six out of 45) among meta-analytic means by SA measure, with no discernable patterns that particular measures consistently had significantly higher or lower effects than other measures. Third, a similar post-hoc analysis with assessment methods also did not reveal clear patterns. Fourth, we tested if the meta-analytic results meaningfully differed depending on assumptions (Borenstein et al. 2009; Cooper, Hedges, and Valentine 2009), using sensitivity analyses comparing: a different CRVE estimator, varying fixed sampling correlation errors, and a variety of fixed values for ghost results and no ghost results. Fifth, we visualize distributions of effects for SA measures and assessment techniques. Last, we evaluate and visualize proportions of effects excluding ghost results below three thresholds and also performed additional analyses based on comments from reviewers.

Discussion

The meta-analytic results combined with analyses using proportions provide strong evidence, in terms of effect sizes, that SA is rarely fundamental or critical to performance. Instead, we found that SA and performance tend to have weak probabilistic relationships on average with high variations. This should be interpreted as limited practical significance, but it does not imply there is no true relationship among SA and performance.

We could only evaluate SA and performance as defined and empirically assessed in the included literature spanning nearly 30 years. Different measures of SA may vary in what they are actually assessing (i.e., construct validity). For example, subjective measures of SA might reflect confidence in SA rather than SA itself (Endsley, 1995b). Nonetheless, patterns of limited validity were generally robust under a number of ways of categorizing and partitioning the data including suggestions by reviewers (e.g., non-expert vs. expert samples and objective vs. subjective SA, see Appendix F).

In sharp contrast to most theories and previous narrative reviews, positing that SA is a meaningful, critical, or fundamental probabilistic factor for performance, our findings indicate limited validity. Many existing theories of SA may need to be revised, as the meta-analytic findings here more closely align with the long-standing theoretical and conceptual concerns about SA (Flach 1995; Billings 1996; Dekker and Hollnagel 2004; van Winsen and Dekker 2015; Sarter and Woods 1991).

Ultimately, interpretation of validity should rely on effect sizes (Smith 2005), and we have interpreted the meta-analytic results accordingly. We found that SA measures typically had limited mean associations with performance, with an overall effect of $r = 0.26$. A comparison may be drawn to the meta-analytic mean of another well-studied relationship: the association between overall task satisfaction (a subjective measure) and objective task performance measured using time and errors (both have meta-analytic effects of $r = 0.23$; Sauro and Lewis 2009). Preferences are posited to be only weakly predictive of task performance, and they exhibit some degree of dissociation (Andre and Wickens 1995; Nielsen and Levy 1994). The similar meta-analytic mean for SA and performance indicates that SA has comparably limited construct representation for performance. There are also possible parallels with observed dissociations among workload and performance which "… occur more frequently than extant explanatory theories imply" (Hancock and Matthews 2019, 374), although to our knowledge there is not yet a meta-analysis assessing the associations among workload and performance. Consistent with a partial dissociation, there were high systematic variations in individual SA-performance effects. Total heterogeneity ($\hat{\tau}$ = 0.21) neared the overall meta-analytic mean effect, yielding a wide plausible range for individual effects: PI [-0.15, 0.60]. Heterogeneity was, at least partially, attributable to a greater than expected proportion of effects below the meta-analytic mean (see Figure 5).

Validity of SA: Implications and falsification

The results here raise numerous unanswered questions; foremost is why does SA generally lack strong and consistent construct representation for psychological processes and/or mechanisms underlying task performance? Returning to Figure 1, what psychological processes actually comprise measures of SA? Specifically, how does SA relate to lower-level cognitive processes (e.g., attention, perception, memory, and decision-making) as well as higher-level processes (e.g., meta-cognition and reasoning; Lichacz 2017; Tremblay 2017)? Little work has empirically investigated potential cognitive processes in SA, for exceptions see (Durso, Bleckley, and Dattel 2006; Özcan and Çakır 2013; Rousseau et al. 2010). Perhaps SA is a higher-order construct in cognition that also (partially) reflects training, knowledge, skills, and abilities? But if SA is a higher-order, multidimensional construct that, to some extent, subsumes both low and high-level psychological processes and other performance-relevant factors (such as the individual, task, and environmental factors shown in Figure 1), then why does SA have tend to have limited construct representation for performance with high variability in associations?

Our findings have general implications relevant to SA-based approaches for design and training (Endsley and Jones 2011). Based on effect sizes here, these approaches are unlikely to translate into meaningful increases in performance *if* SA and *only if* SA is improved. But, if approaches for improving SA also positively affect other causal factors for performance (e.g., training to reduce divided attention by minimizing task switching and dual-tasking, worker scheduling that reduces fatigue and stress) then they will also improve performance. This raises the questions: What makes SA-based approaches different from design and training that do not explicitly incorporate SA? Does all (effective) system design and training also tend to increase SA? These questions and the previous ones prompt concerns about the falsifiability of SA as a construct, see Dekker and Hollnagel (2004). We have demonstrated that at least some degree of falsification for SA is possible using meta-analysis. For strong falsification, future research should address the issues identified here: small samples, overfitting, and selective reporting including ghost results.

Limitations

When possible, we sought to mitigate limitations, but many were inherent to the available data, the systematic review, and the capabilities of current statistical techniques. The major limitations are described in detail below.

Comparing apples to oranges?

A meta-analysis that is too broad in scope and inclusion criteria may be comparing apples to oranges. That is, comparisons across papers may be non-equivalent due to the use of different measures with distinct operationalizations as well as dissimilar experimental designs (Rosenthal and DiMatteo 2001). This is a relevant concern here because, unlike measures of many constructs, the majority of both SA and performance measures are tailored to a particular task and task domain, and they are typically varied from experiment to experiment. However, we contend that our criteria were appropriate because they were consistent with theories and narrative reviews. The measures of SA that we used were labelled as such in the papers analysed and met our stated inclusion criteria. Furthermore, quantitatively synthesising the relevant literature can be viewed as a strength for determining the degree of generalizability to specificity of effects (Rosenthal and DiMatteo 2001).

While methodological differences among papers (e.g., experimental design, assessment methods for SA) can contribute to meta-analytic heterogeneity, we did not find evidence this was the sole cause for systematic variations here. We found meaningful heterogeneity both between- and within-papers, see Table 2. Moreover, meta-analytic results, including heterogeneity, were similar to Bakdash, Marusich, Kenworthy et al. (2020) which simply re-analysed a smaller and largely different sample of papers (the papers included in Endsley [2019] but without selection using statistical significance). Last, we again emphasize, overall results here were generally robust (see Appendix F).

Selection/reporting bias

In general, selection or reporting bias inflates effect sizes and spuriously raises the number of significant results (Ioannidis et al. 2014). Here, we focus on two types of reporting bias; each was only partially addressed in our meta-analysis. The first type of bias is selective reporting, which is the tendency to report only analyses and/or measures that are significant and consistent with the hypothesis (Ioannidis et al. 2014), leaving out non-significant ghost results.

The second type of reporting bias is publication bias (i.e., the file drawer problem: papers with significant results are more likely to be published than papers with non-significant results).

Ghost results. When ghost results are encountered, we unequivocally recommend inclusion of *all results,* not just those reported in detail. Although we identified widespread selective reporting in the literature, we found that including ghost results had minimal impact on the meta-analytic results here. This may be because ghost results were calculated based on the available distribution of non-significant results that *were* reported in detail, which may represent only a subset of the range of non-significant results that authors actually obtained. For example, authors may be more likely to report in detail non-significant results that are in the expected direction but only reach marginal significance, and they may be less likely to report in detail non-significant negative correlations because they do not conform to hypotheses. An alternative possibility is that ghost results tend to have a limited true impact (Head et al. 2015), but also see (Bishop and Thompson 2016; Friese and Frankenbach 2020) and the next section. Regardless, ghost results indicate a substantial research quality issue in the relevant literature.

Publication bias. Papers with significant results, and thus effects with greater magnitudes, are more likely to be submitted and accepted for publication compared to papers with non-significant results (Rosenthal 1979; Ferguson and Heene 2012). Our systematic review had some coverage of the unpublished or grey literature, technical reports and conference papers found using DTIC and Google Scholar (see Appendix D). Still, this is a potential limitation because pervasive publication bias may drastically change meta-analytic results (Carter and McCullough 2014; Ferguson 2015).

Another limitation is that we were unable to implement the standard techniques in meta-analysis for evaluating and adjusting for publication bias (Carter et al. 2019); these techniques have limited power with multivariate/repeated effects (Rodgers and Pustejovsky 2020). Even with independent effects, adjusting for publication bias is problematic with non-trivial heterogeneity (Carter et al. 2019).

Between-participants versus within-participants

All results we included here evaluated correlations between-participants, either averaged across participants or separate analysis of each SA-performance pairing. A handful of papers did evaluate relationships for SA-performance within-participants, but results were varied (Loft et al. 2018; O'Hagan et al. 2019; Strybel et al. 2009). We were unable to include the small number of within-participant results because the output of multilevel models could not be converted to correlation coefficients or the results were overfit. It is possible the within-individual relationship(s) for SA and performance could meaningfully differ from the between-individual relationships (Bakdash and Marusich 2017; Fisher, Medaglia, and Jeronimus 2018; Molenaar and Campbell 2009). To evaluate patterns within-individuals, we recommend using repeated measures correlation (Bakdash and Marusich 2017) and/or multilevel modelling which can evaluate patterns between- and within-individuals simultaneously (Aarts et al. 2014; Gelman, Hill, and Yajima 2012). For multilevel modelling, sharing raw data will likely be necessary for meta-analysis.

Overfitting

We identified numerous papers with overfit results (26 papers with all results overfit and 3 papers with partial overfitting), erroneously treating repeated measures from individuals as independent in their analyses, for details see (Bakdash, Marusich, Cox, et al., 2021a). This indicates that overfitting was common in the literature relevant to this meta-analysis. Although we had to exclude a substantial amount of relevant work here, previous work with a smaller and largely different dataset of papers that included overfit results as-reported produced similar results (Bakdash, Marusich, Kenworthy, et al. 2020). Nevertheless, as with ghost results, overfitting is a pervasive research quality problem. There are multiple ways to avoid overfitting, including averaging and separate analyses as well as techniques described in the previous paragraph.

Conclusion

We have shown that there is a substantial disconnect between SA's compelling descriptive and theoretical strength and its weak and highly varied meta-analytic quantitative validity for performance. Despite several caveats associated with the included literature (e.g., limited sample sizes and thus low power, selection/reporting bias, and varying design and research quality), meta-analysis still provides the best available estimate for true effects (Button et al. 2013). This is not the first time that meta-analytic results diverged from narrative literature reviews and qualitative interpretations of empirical research (Mann 1994) as well as expert opinion (Mann 1990).

SA has enormous breadth: it combines a theory with a construct indicated by multiple measures to assess a process/state, a method for improving training and system design, and a desired end-goal. As a description, SA has undeniable intuitive appeal. It is perhaps the most succinct term for describing the combination of human cognition, task performance, and human interaction with systems in complex environments (Byrne 2015; Jones 2015). Yet, the descriptive appeal and breadth of SA may also make it 'too neat' and 'too holistic' (Billings 1996). Face validity is the weakest type of construct validity (Drost 2011).

Two recent meta-analyses also found that other intuitive psychological constructs fell short of their high face validity and respective theoretical claims. First, mindsets (i.e., beliefs that abilities, such as intelligence, can grow with effort) generally had much weaker effects on academic performance compared to typical effect sizes in other educational interventions (Sisk et al. 2018). Second, grit (i.e., high perseverance and resilience toward goals) was a weak and also a non-unique predictor of academic performance; moreover, grit also lacked a theoretically posited higher-order structure accounting for other personality traits (Credé, Tynan, and Harms 2017).

Paralleling the strong face validity of SA for performance, usability preferences (such as beliefs, predispositions, and satisfaction for using particular products or systems) also have high face validity with task performance (Nielsen and Levy 1994). However, preferences and performance can be dissociable and thus conflict, so preferences should never be assessed alone as a proxy for performance in usability research (Andre and Wickens 1995; Nielsen and Levy 1994). Likewise, we strongly caution against using measures of SA alone as a proxy or surrogate variable for inferring performance (and vice versa), especially if actual performance can be measured, which is not always possible for real-world tasks. Similarly, interventions designed to improve SA may not necessarily translate to meaningful improvements in performance.

Given the meta-analytic results, we instead recommend alternative approaches for assessing and improving real-world performance — for example, investigating all relevant systems for people, technology, and the environment: individual, team, and organization-levels, tasks and outcomes, technological systems, and the work environment. Such broader system approaches and frameworks include the Swiss cheese model of human error with potential hazards and using preventive defences as barriers (Reason 1990), distributed cognition (Hutchins 1995), distributed situation awareness (Stanton, Salmon, and Walker 2015), socio-technical systems (Perrow 2011) and natural systems (Durso and Drews 2010), and task-technology fit to appropriately match tasks and technology together (Goodhue and Thompson, 1995).

Acknowledgments

We thank the reviewers and the editor for helpful comments, including additional analyses. The authors also thank Nancy Simini for editing a previous version of this manuscript.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation.

Disclosure statement

The authors declare no potential conflict of interest.

Funding

Not applicable.

Biographical note

Dr. Bakdash is a research psychologist for the U.S. Army Research Laboratory. He received his PhD in cognitive psychology from the University of Virginia in 2010.

Dr. Marusich is a research psychologist at the U.S. Army Research Laboratory. She received her PhD in cognitive psychology from the University of Texas at Austin in 2011.

Dr. Cox is a research psychologist at the U.S. Army Research Laboratory. She received her PhD in cognitive psychology from Georgetown University in 2014.

Dr. Zaroukian is a cognitive scientist for the U.S. Army Research Laboratory. She received her PhD in cognitive science from Johns Hopkins University in 2013.

Dr. Geuss is a research psychologist for the U.S. Army Research Laboratory. He received his PhD in cognitive psychology from the University of Utah in 2014.

Ms. Morris is a College Qualified Leader at the U.S. Army Research Laboratory. She is a senior finishing her B.A. of Psychology at the University of Texas, Arlington.

Dataset availability

The data that support the findings of this paper are openly available on the Open Science Framework: https://doi.org/10.17605/OSF.IO/4K7ZV (Bakdash, Marusich, Cox, et al. 2021a) and a Code Ocean Capsule: https://doi.org/10.24433/CO.1682542.v4 (Bakdash, Marusich, Cox, et al. 2021b).

References

Aarts, Emmeke, Matthijs Verhage, Jesse V Veenvliet, Conor V Dolan, and Sophie van der Sluis. 2014. "A Solution to Dependency: Using Multilevel Analysis to Accommodate Nested Data." *Nature Neuroscience* 17 (4): 491–496. doi:10.1038/nn.3648.

Allen, Micah, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, and Rogier A. Kievit. 2019. "Raincloud Plots: A Multi-Platform Tool for Robust Data Visualization." *Wellcome Open Research* 4 (April): 63. doi:10/gfxr7w.

Andre, A. D., and C. D. Wickens. 1995. "When Users Want What's Not Best for Them." *Ergonomics in Design: The Quarterly of Human Factors Applications* 3 (4): 10–14. doi:10.1177/106480469500300403

Babyak, Michael A. 2004. "What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models." *Psychosomatic Medicine* 66 (3): 411–421. doi:https://doi.org/10.1097/01.psy.0000127692.23278.a9.

Bakdash, Jonathan Z., and Laura R. Marusich. 2017. "Repeated Measures Correlation." *Frontiers in Psychology* 8: 1–13. doi:https://doi.org/10.3389/fpsyg.2017.00456.

Bakdash, Jonathan Z., Laura R. Marusich, Katherine Cox, Michael Geuss, Erin Zaroukian, and Katelyn Morris. 2021a. "The Validity of Situation Awareness for Performance: A Meta-Analysis (Systematic Review, Data, and Code for Results)." Open Science Framework (OSF): https://doi.org/10.17605/OSF.IO/4K7ZV

Bakdash, Jonathan Z., Laura R. Marusich, Katherine Cox, Michael Geuss, Erin Zaroukian, and Katelyn Morris. 2021b. "The Validity of Situation Awareness for Performance: A Meta-Analysis." Code Ocean Capsule: https://doi.org/10.24433/CO.1682542.v4

Bakdash, Jonathan Z., Laura R. Marusich, Jared Kenworthy, Elyssa Twedt, and Erin Zaroukian. 2020. "Statistical Significance Filtering Overestimates Effects and Impedes Falsification: A Critique of Endsley (2019)." *Frontiers in Psychology* 11 (3669): 1–12. doi: 10.3389/fpsyg.2020.609647

Billings, C. E. 1996. “Situation Awareness Measurement and Analysis: A Commentary.” In *Situation Awareness Measurement and Analysis*, 1–5. Daytona Beach, FL: Center for Aviation/Aerospace Research. doi:http://www.dtic.mil/dtic/tr/fulltext/u2/a522540.pdf.

Bishop, Dorothy VM. 2019. “Rein in the Four Horsemen of Irreproducibility.” *Nature* 568 (April): 435–435. doi:10.1038/d41586-019-01307-2.

Bishop, Dorothy VM. 2020. “The Psychology of Experimental Psychologists: Overcoming Cognitive Constraints to Improve Research: The 47th Sir Frederic Bartlett Lecture.” *Quarterly Journal of Experimental Psychology* 73 (1): 1–19. doi:10.1177/1747021819886519.

Bishop, Dorothy VM, and Paul A. Thompson. 2016. “Problems in Using P-Curve Analysis and Text-Mining to Detect Rate of p-Hacking and Evidential Value.” *PeerJ* 4 (February): e1715. doi:10.7717/peerj.1715.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons. doi:10.1002/9780470743386.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. “Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience.” *Nature Reviews Neuroscience* 14 (5): 365–376. doi:10.1038/nrn3475.

Byrne, Evan. 2015. “Commentary on Endsley’s ‘Situation Awareness Misconceptions and Misunderstandings’.” *Journal of Cognitive Engineering and Decision Making* 9 (1): 84–86. doi:10.1177/1555343414554703.

Carter, Evan C., and Michael E. McCullough. 2014. “Publication Bias and the Limited Strength Model of Self-Control: Has the Evidence for Ego Depletion Been Overestimated?” *Frontiers in Psychology* 5: 1–11. doi:10.3389/fpsyg.2014.00823.

Carter, Evan C., Felix Schönbrodt, Will M. Gervais, and Joseph Hilgard. 2019. “Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods.” *Advances in Methods and Practices in Psychological Science* 2: 115–144. doi:10.1177/2515245919847196.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Second Edition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, Harris, Larry V. Hedges, and Jeffrey C. Valentine. 2009. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd Edition. New York: Russell Sage Foundation.

Credé, Marcus, Michael C. Tynan, and Peter D. Harms. 2017. “Much Ado about Grit: A Meta-Analytic Synthesis of the Grit Literature.” *Journal of Personality and Social Psychology* 113 (3): 492–511. doi:10.1037/pspp0000102.

Cronbach, Lee J., and Paul E. Meehl. 1955. “Construct Validity in Psychological Tests.” *Psychological Bulletin* 52 (4): 281–302. doi:10.1037/h0040957.

Dekker, Sidney, and Erik Hollnagel. 2004. “Human Factors and Folk Models.” *Cognition, Technology & Work* 6 (2): 79–86. doi:10.1007/s10111-003-0136-9.

Drost, Ellen A. 2011. “Validity and Reliability in Social Science Research.” *Education Research and Perspectives* 38 (1): 105–123.

Durso, F. T., M. Kathryn Bleckley, and Andrew R. Dattel. 2006. “Does Situation Awareness Add to the Validity of Cognitive Tests?” *Human Factors* 8 (4): 721–733. doi:10.1518/001872006779166316.

Durso, F. T., and F. A. Drews. 2010. "Health Care, Aviation, and Ecosystems: A Socio-Natural Systems Perspective." *Current Directions in Psychological Science* 19 (2): 71–75. doi:10.1177/0963721410364728.

Durso, F. T., and A. Sethumadhavan. 2008. "Situation Awareness: Understanding Dynamic Environments." *Human Factors* 50 (3): 442–448. doi:10.1518/001872008X288448.

Embretson (Whitely), Susan E. 1983. "Construct Validity: Construct Representation versus Nomothetic Span." *Psychological Bulletin* 93 (1): 179–197. doi:10.1037/0033-2909.93.1.179.

Endsley, Mica R. 1988. "Situation Awareness Global Assessment Technique (SAGAT)." In *Proceedings of IEEE National Aerospace and Electronics Conference*: 789–795. doi:10.1109/NAECON.1988.19509

Endsley, Mica R. 1995a. "Toward a Theory of Situation Awareness in Dynamic Systems." *Human Factors* 37 (1): 32–64. doi:10.1518/001872095779049543

Endsley, Mica R. 1995b. "Measurement of Situation Awareness in Dynamic Systems." *Human Factors* 37 (1): 65–84. doi:10.1518/001872095779049499.

Endsley, Mica R. 2000a. "Direct Measurement of Situation Awareness: Validity and Use of SAGAT." In *Situation Awareness: Analysis and Measurement*, edited by M.R. Endsley and D.J. Garland, 131–157. Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, Mica R. 2000b. "Theoretical Underpinnings of Situation Awareness: A Critical Review." In *Situation Awareness: Analysis and Measurement*, edited by M.R. Endsley and D.J. Garland, 3–28. Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, Mica R. 2015a. "Situation Awareness: Operationally Necessary and Scientifically Grounded." *Cognition, Technology & Work* 17 (2): 163–167. doi:10.1007/s10111-015-0323-5.

Endsley, Mica R. 2015b. "Situation Awareness Misconceptions and Misunderstandings." *Journal of Cognitive Engineering and Decision Making* 9 (1): 4–32. doi:10.1177/1555343415572631.

Endsley, Mica R. 2015c. "Final Reflections: Situation Awareness Models and Measures." *Journal of Cognitive Engineering and Decision Making* 9 (1): 101–111. doi:10.1177/1555343415573911.

Endsley, Mica R. 2019. "A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM." *Human Factors*, September, 0018720819875376. doi:10.1177/0018720819875376.

Endsley, Mica R. 2020. "The Divergence of Objective and Subjective Situation Awareness: A Meta-Analysis." *Journal of Cognitive Engineering and Decision Making* 14 (1): 34–53. doi:10/ggqfzd.

Endsley, Mica R., Cheryl A. Bolstad, Debra G. Jones, and Jennifer M. Riley. 2003. "Situation Awareness Oriented Design: From User's Cognitive Requirements to Creating Effective Supporting Technologies." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 47: 268–272.

Endsley, Mica R., and Erik S. Connors. 2008. "Situation Awareness: State of the Art." In *IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century*: 1–4. doi: 10.1109/PES.2008.4596937

Endsley, Mica R., and Debra G. Jones. 2011. *Designing for Situation Awareness: An Approach to User-Centered Design*. Second Edition. New York: CRC Press.

Ferguson, Christopher J. 2015. “Do Angry Birds Make for Angry Children? A Meta-Analysis of Video Game Influences on Children’s and Adolescents’ Aggression, Mental Health, Prosocial Behavior, and Academic Performance.” *Perspectives on Psychological Science* 10 (5): 646–666. doi: 10.1177/1745691615592234

Ferguson, Christopher J., and Moritz Heene. 2012. “A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science’s Aversion to the Null.” *Perspectives on Psychological Science* 7 (6): 555–561. doi:10.1177/1745691612459059.

Fisher, Aaron J., John D. Medaglia, and Bertus F. Jeronimus. 2018. “Lack of Group-to-Individual Generalizability Is a Threat to Human Subjects Research.” *Proceedings of the National Academy of Sciences* 115 (27): 1–10. doi:10.1073/pnas.1711978115.

Flach, John M. 1995. “Situation Awareness: Proceed with Caution.” *Human Factors* 37 (1): 149–157. doi:10.1518/001872095779049480.

Friese, Malte, and Julius Frankenbach. 2020. “p-Hacking and Publication Bias Interact to Distort Meta-Analytic Effect Size Estimates.” *Psychological Methods* 25 (4): 456–471. doi:10/gg22vx.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima. 2012. “Why We (Usually) Don’t Have to Worry About Multiple Comparisons.” *Journal of Research on Educational Effectiveness* 5 (2): 189–211. doi:10.1080/19345747.2011.618213.

Gelman, Andrew, and Eric Loken. 2013. “The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No ‘Fishing Expedition’ or ‘p-Hacking’ and the Research Hypothesis Was Posited Ahead of Time.” http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Glass, Gene V. 2015. “Meta-Analysis at Middle Age: A Personal History.” *Research Synthesis Methods* 6 (3): 221–231. doi:10.1002/jrsm.1133.

Goodhue, Dale. L. and Thompson, Roland L. (1995). "Task-technology fit and individual performance." *MIS quarterly 19* (2): 213-236. doi: 10.2307/249689

Gurevitch, Jessica, Julia Koricheva, Shinichi Nakagawa, and Gavin Stewart. 2018. “Meta-Analysis and the Science of Research Synthesis.” *Nature* 555 (7695): 175–182. doi:10.1038/nature25753.

Gusenbauer, Michael. 2019. “Google Scholar to Overshadow Them All? Comparing the Sizes of 12 Academic Search Engines and Bibliographic Databases.” *Scientometrics* 118 (1): 177–214. doi:10/cxdz.

Hancock, Peter A. and Gerald Matthews (2019). "Workload and performance: Associations, insensitivities, and dissociations." *Human Factors 61*(3), 374-392. doi: 10.1177/0018720818809590

Harzing, A. W. 2020. “Publish or Perish 7.” *Harzing.Com*. https://harzing.com/resources/publish-or-perish.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. “The Extent and Consequences of P-Hacking in Science.” *PLOS Biology* 13 (3): e1002106. doi:10.1371/journal.pbio.1002106.

Hedges, Larry V., Elizabeth Tipton, and Matthew C. Johnson. 2010. “Robust Variance Estimation in Meta-Regression with Dependent Effect Size Estimates.” *Research Synthesis Methods* 1 (1): 39–65. doi:10.1002/jrsm.5.

Higgins, Julian P T, Simon G Thompson, Jonathan J Deeks, and Douglas G Altman. 2003. “Measuring Inconsistency in Meta-Analyses.” *British Medical Journal* 327 (7414): 557–560. doi:10.1136/bmj.327.7414.557

Hutchins, Edwin. 1995. *Cognition in the Wild*. MIT press.

Imbens, Guido W., and Michal Kolesar. 2016. “Robust Standard Errors in Small Samples: Some Practical Advice.” *Review of Economics and Statistics* 98 (4): 701–712. doi:10.1162/REST_a_00552.

IntHout, Joanna, John PA Ioannidis, and George F Borm. 2014. “The Hartung-Knapp-Sidik-Jonkman Method for Random Effects Meta-Analysis Is Straightforward and Considerably Outperforms the Standard DerSimonian-Laird Method.” *BMC Medical Research Methodology* 14 (1). doi:10.1186/1471-2288-14-25.

Ioannidis, John P. A., Marcus R. Munafò, Paolo Fusar-Poli, Brian A. Nosek, and Sean P. David. 2014. “Publication and Other Reporting Biases in Cognitive Sciences: Detection, Prevalence, and Prevention.” *Trends in Cognitive Sciences* 18 (5): 235–241. doi:10.1016/j.tics.2014.02.010.

Jipp, Meike, and Phillip L. Ackerman. 2016. “The Impact of Higher Levels of Automation on Performance and Situation Awareness: A Function of Information-Processing Ability and Working-Memory Capacity.” *Journal of Cognitive Engineering and Decision Making* 10 (2): 138–166. doi:10.1177/1555343416637517.

Jones, Debra G. 2015. “A Practical Perspective on the Utility of Situation Awareness.” *Journal of Cognitive Engineering and Decision Making* 9 (1): 98–100. doi:10.1177/1555343414554804.

Lankton, P. 2019. “Endsley’s Model of SA.” *Wikipedia*, retrieved June 24, 2019 from: https://en.wikipedia.org/wiki/Situation_awareness#/media/File:Endsley-SA-model.jpg.

Kenny, David A., and Charles M. Judd. 1986. “Consequences of Violating the Independence Assumption in Analysis of Variance.” *Psychological Bulletin* 99 (3): 422–431. doi:10.1037/0033-2909.99.3.422.

Knapp, Guido, and Joachim Hartung. 2003. “Improved Tests for a Random Effects Meta-Regression with a Single Covariate.” *Statistics in Medicine* 22 (17): 2693–2710.

Konstantopoulos, Spyros. 2011. “Fixed Effects and Variance Components Estimation in Three-Level Meta-Analysis.” *Research Synthesis Methods* 2 (1): 61–76. doi: 10.1002/jrsm.35

Kriegeskorte, Nikolaus, W Kyle Simmons, Patrick S F Bellgowan, and Chris I Baker. 2009. “Circular Analysis in Systems Neuroscience: The Dangers of Double Dipping.” *Nature Neuroscience* 12 (5): 535–540. doi:10.1038/nn.2303.

Kühberger, Anton, Astrid Fritz, and Thomas Scherndl. 2014. “Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size.” *PLoS ONE* 9 (9): 1–8. doi:10.1371/journal.pone.0105825.

Lichacz, Frederick M. J. 2017. “The Missing Cognitive Link in Situation Awareness Research.” In *Engineering Psychology and Cognitive Ergonomics*, edited by Don Harris, 307–314. Routledge. doi:10.4324/9781315094472-36.

Loft, Shayne, Lisa Jooste, Yanqi Ryan Li, Timothy Ballard, Samuel Huf, Ottmar V. Lipp, and Troy A. W. Visser. 2018. “Using Situation Awareness and Workload to Predict Performance in Submarine Track Management: A Multilevel Approach.” *Human Factors* 60 (7): 978–991. doi:10.1177/0018720818784803.

Mann, C. 1990. “Meta-Analysis in the Breech.” *Science* 249 (4968): 476–480. doi:10/djb9wr.

Mann, C. 1994. “Can Meta-Analysis Make Policy?” *Science* 266 (5187): 960–962. doi:10.1126/science.7973676.

Mathur, Maya B., and Tyler J. VanderWeele. 2019. "New Metrics for Meta-Analyses of Heterogeneous Effects: Metrics for Meta-Analyses." *Statistics in Medicine* 38 (8): 1336–1342. doi:10.1002/sim.8057.

Mathur, Maya B., and Tyler J. VanderWeele. 2020a. "Robust Metrics and Sensitivity Analyses for Meta-Analyses of Heterogeneous Effects." *Epidemiology* 31 (3): 356–358. doi:10/ggzkwg.

Mathur, Maya B., and Tyler J. VanderWeele. 2020b. "Meta-Regression Methods to Characterize Evidence Strength Using Meaningful-Effect Percentages Conditional on Study Characteristics." *OSF Preprints*. doi:10.31219/osf.io/bmtdq.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement." *PLOS Medicine* 6 (7): e1000097. doi:10.1371/journal.pmed.1000097.

Molenaar, P.C., and C.G. Campbell. 2009. "The New Person-Specific Paradigm in Psychology." *Current Directions in Psychological Science* 18: 112–117. doi:10.1111/j.1467-8721.2009.01619.x.

Nielsen, Jakob, and Jonathan Levy. 1994. "Measuring Usability: Preference vs. Performance." *Commun. ACM* 37 (4): 66–75. doi:10.1145/175276.175282.

O'Brien, K. S., and D. O'Hare. 2007. "Situational Awareness Ability and Cognitive Skills Training in a Complex Real-World Task." *Ergonomics* 50 (7): 1064–1091. doi:10.1080/00140130701276640.

O'Hagan, A.D., J. Issartel, A. Wall, F. Dunne, P. Boylan, J. Groeneweg, M. Herring, M. Campbell, and G. Warrington. 2019. "'Flying on Empty'–Effects of Sleep Deprivation on Pilot Performance." *Biological Rhythm Research*. doi:10.1080/09291016.2019.1581481.

Ouzzani, Mourad, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. "Rayyan—a Web and Mobile App for Systematic Reviews." *Systematic Reviews* 5 (1): 1–10. doi:10/gfkdzd.

Özcan, Orçun Orkan, and Murat Perit Çakır. 2013. "Exploring the Effects of Working Memory Capacity, Attention, and Expertise on Situation Awareness in a Flight Simulation Environment." In *17th International Symposium on Aviation Psychology*, 98–103. https://corescholar.libraries.wright.edu/isap_2013/94/.

Parasuraman, Raja, Thomas B. Sheridan, and Christopher D. Wickens. 2008. "Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs." *Journal of Cognitive Engineering and Decision Making* 2 (2): 140–160. doi:10.1518/155534308X284417.

Perrow, Charles. 2011. *Normal Accidents: Living with High Risk Technologies*. Princeton, NJ: Princeton University Press.

Pustejovsky, James E., and Elizabeth Tipton. 2018. "Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models." *Journal of Business & Economic Statistics* 36 (4): 672–683. doi:10.1080/07350015.2016.1247004.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria. https://www.R-project.org/.

Reason, James. 1990. *Human Error*. Cambridge University Press.

Rodgers, Melissa A., and James E. Pustejovsky. 2020. "Evaluating Meta-Analytic Methods to Detect Selective Reporting in the Presence of Dependent Effect Sizes." *Psychological Methods*. doi:10/ghd56b.

Rosenthal, Robert. 1979. “The File Drawer Problem and Tolerance for Null Results.” *Psychological Bulletin* 86 (3): 638–641. doi:10.1037/0033-2909.86.3.638.

Rosenthal, Robert, and M. Robin DiMatteo. 2001. “Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews.” *Annual Review of Psychology* 52 (1): 59–82. doi:10.1146/annurev.psych.52.1.59.

Rousseau, Robert, Sébastien Tremblay, Simon Banbury, Richard Breton, and Adel Guitouni. 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. doi:10.1080/14639220903010076.

Salmon, Paul M., Neville A. Stanton, Guy Walker, and Damian Green. 2006. “Situation Awareness Measurement: A Review of Applicability for C4i Environments.” *Applied Ergonomics* 37 (2): 225–238. doi:10.1016/j.apergo.2005.02.001.

Salmon, Paul M., Neville A. Stanton, Guy H. Walker, Chris Baber, Daniel P. Jenkins, Richard McMaster, and Mark S. Young. 2008. “What Really Is Going on? Review of Situation Awareness Models for Individuals and Teams.” *Theoretical Issues in Ergonomics Science* 9 (4): 297–323. doi:10.1080/14639220701561775.

Sarter, Nadine B., and David D. Woods. 1991. “Situation Awareness: A Critical but Ill-Defined Phenomenon.” *The International Journal of Aviation Psychology* 1 (1): 45–57. doi:10.1207/s15327108ijap0101_4.

Sauro, Jeff, and James R Lewis. 2009. “Correlations among Prototypical Usability Metrics: Evidence for the Construct of Usability.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1609–1618. doi:10.1145/1518701.1518947.

Scammacca, Nancy, Greg Roberts, and Karla K. Stuebing. 2014. “Meta-Analysis With Complex Research Designs: Dealing With Dependence From Multiple Measures and Multiple Group Comparisons.” *Review of Educational Research* 84 (3): 328–364. doi:10.3102/0034654313500826.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11): 1359–1366. doi:10.1177/0956797611417632.

Sisk, Victoria F., Alexander P. Burgoyne, Jingze Sun, Jennifer L. Butler, and Brooke N. Macnamara. 2018. “To What Extent and Under Which Circumstances Are Growth Mind-Sets Important to Academic Achievement? Two Meta-Analyses.” *Psychological Science* 29 (4): 549–571. doi:10.1177/0956797617739704.

Smith, Gregory T. 2005. “On Construct Validity: Issues of Method and Measurement.” *Psychological Assessment* 17 (4): 396–408. doi:10.1037/1040-3590.17.4.396.

Stanton, Neville A., Peter RG Chambers, and J. Piggott. 2001. “Situational Awareness and Safety.” *Safety Science* 39 (3): 189–204. doi:https://doi.org/10.1016/S0925-7535(01)00010-8.

Stanton, Neville A., Paul M. Salmon, and Guy H. Walker. 2015. “Let the Reader Decide: A Paradigm Shift for Situation Awareness in Sociotechnical Systems.” *Journal of Cognitive Engineering and Decision Making* 9 (1): 44–50. doi:10.1177/1555343414552297.

Strauss, Milton E., and Gregory T. Smith. 2009. “Construct Validity: Advances in Theory and Methodology.” *Annual Review of Clinical Psychology* 5 (1): 1–25. doi:10.1146/annurev.clinpsy.032408.153639.

Strybel, TZ, K Minakata, J Nguyen, R Pierce, and K-P L Vul 2009. “Optimizing Online Situation Awareness Probes in Air Traffic Management Tasks.” *Symposium on Human Interface Human Interface*: 845–854. doi:10.1007/978-3-642-02559-4_91.

Tenney, Y. J., and R. W. Pew. 2006. “Situation Awareness Catches on: What? So What? Now What?” *Reviews of Human Factors and Ergonomics* 2 (1): 1–34. doi:10.1177/1557234X0600200102.

Tipton, Elizabeth. 2015. “Small Sample Adjustments for Robust Variance Estimation with Meta-Regression.” *Psychological Methods* 20 (3): 375–393. doi:10.1037/met0000011.

Tremblay, Sébastien. 2017. *A Cognitive Approach to Situation Awareness: Theory and Application*. Routledge.

van Winsen, Roel, and Sidney W. A. Dekker. 2015. “SA Anno 1995: A Commitment to the 17th Century.” *Journal of Cognitive Engineering and Decision Making* 9 (1): 51–54. doi:10.1177/1555343414557035.

Vidulich, Michael A. 2000. “Sensitivity of Situation Awareness Metrics in User Interfaces.” In *Situation Awareness Analysis and Measurement*, Mica R. Endsley and Daniel J. Garland (Eds), 203–223. CRC Press.

Viechtbauer, Wolfgang. 2010. “Conducting Meta-Analyses in R with the *metafor* package.” *Journal of Statistical Software* 36 (3): 1–48. doi:10/gckfpj.

Vosgerau, Joachim, Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. 2019. “99% Impossible: A Valid, or Falsifiable, Internal Meta-Analysis.” *Journal of Experimental Psychology: General* 148 (9): 1628–1639. doi:10.1037/xge0000663.

Vul, Edward, Christine Harris, Piotr Winkielman, and Harold Pashler. 2009. “Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition.” *Perspectives on Psychological Science* 4 (3): 274–290. doi:https://doi.org/10.1111/j.1745-6924.2009.01125.x.

Walshe, Nuala C., Clare M. Crowley, Sinéad O’Brien, John P. Browne, and Josephine M. Hegarty. 2019. “Educational Interventions to Enhance Situation Awareness: A Systematic Review and Meta-Analysis.” *Simulation in Healthcare* 14 (6): 398–408. doi:10/ghdjh5.

Wicherts, J.M., C.S. Veldkamp, H.E.M. Augusteijn, M. Bakker, R.C.M. van Aert, and M.A.L.M. van Assen. 2016. “Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking.” *Frontiers in Psychology* 7: 1–12. doi:10.3389/fpsyg.2016.01832

Wickens, C. D. 2008. “Situation Awareness: Review of Mica Endsley’s 1995 Articles on Situation Awareness Theory and Measurement.” *Human Factors* 50 (3): 397–403. doi:10.1518/001872008X288420.

Preprint version 7: Updates to supplemental material, May 2022

- There were three errors for calculations with detailed effects only, Appendix F:

    i. Fixed the last column for detailed effects in *Table F5*: R object = *table.ns* and file = *results/TabF5_ghosts_no_ghosts.csv*

    ii. Fixed meta-analytic mean for detailed effects in *Figure F4*: R object = *plot.detailed*, file = *results/FigF4_detailed.pdf*)

    iii. Fixed the prop of detailed effects below in *Table F6*: files = *results/props.det.RDS* and *results/TabF6_props_detailed.csv*

- Refactored proportion estimates using the clustering (*cluster.name* = .) now implemented in *prop_stronger()*. Huge improvements to runtime, so bootstrapping is always run now. Proportions are similar to the previous implementation but new confidence intervals are wider.

- OSF version: All results generated at run-time now saved in */results*. This is now consistent with the Code Ocean version.

Supplemental Online Material for The Validity of Situation Awareness for Performance: A meta-analysis

Appendix A: PRISMA Checklist

Table A1. PRISMA checklist. This table contains the location and a summary of information for each PRIMSA checklist item. Source of checklist: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 6(7): e1000097. doi:10.1371/journal.pmed1000097 For more information, visit: www.prisma-statement.org.

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	First page. This work was only identified as a meta-analysis in title. Both were stated in the Introduction.
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	First page. The abstract includes the background, objectives, method, results, implications, and conclusion. The systematic review and analyses were not pre-registered.
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Introduction of the paper.
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Introduction. General statement about outcome measures (associations among SA and performance measures). We considered the Problem and Outcomes, but not the other elements of PICOS, they were not applicable here because

Section/topic	#	Checklist item	Reported on page #
			this is not a systematic review/meta-analysis of clinical research. PICOS is Patient/Population/Problem, Intervention, Comparison, Outcomes, and Study Design.
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Methods. See online materials for the abstract review guide. We did not pre-register a protocol.
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Characteristics, eligibility, and rationale are described in the Introduction and Methods, also see Appendix C.
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Information sources described in Appendix C. See online materials (Systematic Review folder). Dates of coverage were anytime to Jan 2020. We contacted two authors from one paper requesting raw correlations (the paper only reported significant correlations following Bonferroni corrections). The authors did not have access to the results at that time. See online materials (Bakdash, Marusich, Cox et al. 2021a) for search dates.

Section/topic	#	Checklist item	Reported on page #
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Methods and online materials (includes exported references at every stage of the systematic review), also see Appendix C and Bakdash, Marusich, Cox et al. (2021a).
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Methods and Appendix C.
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Methods and Appendix C.
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Methods, Appendix C, and comments in the data files and code for judgment calls on included/excluded results. For example, in Gugerty (1997) we only included correlations by experiments 1-3 and did not include the correlations combining all three experiments.
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	Methods. We attempted to mitigate reporting bias by including ghost results. We also excluded detected overfit results.
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	Methods. The principal summary measure was a Pearson correlation coefficient. When possible, other effects were converted to a Pearson correlation.
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	Results and Appendix F for alternative models. Both meta-analytic inconsistency and heterogeneity were

Section/topic	#	Checklist item	Reported on page #
			estimated. Heterogeneity was also estimated for other models in Appendix F.
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	Methods. The only assessments of bias were ghost and overfit results.
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	Methods and Appendix F. All analyses in Appendix F were pre-planned *except* for non-expert vs. expert participants, objective vs. subjective SA, and the text search for confounds and other factors. These three analyses were performed following comments and suggestions from reviewers.
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Methods and Appendix C.
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Methods and Appendix D.
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Methods and Appendix F. We only performed aggregate analyses of the number of effects and papers with ghost results.
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Results and Appendix F. We do not present results by paper because of the ghost results and because a by-paper model cannot be fit with CRVE.

Section/topic	#	Checklist item	Reported on page #
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Results and Appendix F. In the paper, the forest plots include mean effects with confidence intervals and a prediction interval. We also show raincloud plots and analyze proportions of effects below three key thresholds.
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Methods and Appendix F (ghost results and overfitting)
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Appendix F: Multiple sensitivity analysis and other analyses. Results were generally robust across all analyses. Bakdash, Marusich, Cox et al. (2021a; 2021b) contain a few additional analyses not reported in the supplemental material.
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	Discussion and Conclusion.
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Methods, Discussion, Conclusion, and Appendix F.
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Discussion and Conclusion.
FUNDING			

Section/topic	#	Checklist item	Reported on page #
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	N/A. The authors have no conflict of interest for funding or other sources of support.

Appendix B: R Package References

Aust, Frederik, and Marius Barth. 2020. *papaja: Create APA Manuscripts with R Markdown*. Manual. https://github.com/crsh/papaja.

Bates, Douglas, and Martin Maechler. 2019. *Matrix: Sparse and Dense Matrix Classes and Methods*. Manual. https://CRAN.R-project.org/package=Matrix.

Boessenkool, Berry. 2020. *BerryFunctions: Function Collection Related to Plotting and Hydrology*. Manual. https://CRAN.R-project.org/package=berryFunctions.

Brueckl, Markus, and Florian Heuer. 2018. *IrrNA: Coefficients of Interrater Reliability - Generalized for Randomly Incomplete Datasets* (version 0.1.4). https://CRAN.R-project.org/package=irrNA.

Champely, Stephane. 2020. *Pwr: Basic Functions for Power Analysis*. Manual. https://CRAN.R-project.org/package=pwr.

Csárdi, Gábor. 2019. *Pkgconfig: Private Configuration for "r" Packages*. Manual. https://CRAN.R-project.org/package=pkgconfig.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press. http://statwww.epfl.ch/davison/BMA/.

Fisher, Zachary, Elizabeth Tipton, and Hou Zhipeng. 2017. *Robumeta: Robust Variance Meta-Regression*. Manual. https://CRAN.R-project.org/package=robumeta.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. 3rd ed. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Fox, John, Sanford Weisberg, and Brad Price. 2020. *CarData: Companion to Applied Regression Data Sets*. Manual. https://CRAN.R-project.org/package=carData.

Gamer, Matthias, Jim Lemon, and Ian Fellows Puspendra Singh. 2019. *Irr: Various Coefficients of Interrater Reliability and Agreement* (version 0.84.1). https://CRAN.R-project.org/package=irr.

Garnier, Simon, Noam Ross, Bob Rudis, Marco Sciaini, and Cédric Scherer. 2018. *Viridis: Default Color Maps from "Matplotlib"* (version 0.5.1). https://CRAN.R-project.org/package=viridis.

Gaslam, Brodie. 2020. *Fansi: ANSI Control Sequence Aware String Functions*. Manual. https://CRAN.R-project.org/package=fansi.

Genz, Alan, and Frank Bretz. 2009. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Heidelberg: Springer-Verlag.

Gerber, Florian. 2020. *OptimParallel: Parallel Version of the L-BFGS-B Optimization Method* (version 1.0-1). https://CRAN.R-project.org/package=optimParallel.

Grosser, Malte. 2019. *Snakecase: Convert Strings into Any Case*. Manual. https://CRAN.R-project.org/package=snakecase.

Hallman, Jeff. 2020. *Tis: Time Indexes and Time Indexed Series*. Manual. https://CRAN.R-project.org/package=tis.

Hankin, Robin K. S. 2006. "Special Functions in R: Introducing the Gsl Package." *R News* 6 (4).

Harrell Jr, Frank E, with contributions from Charles Dupont, and many others. 2020. *Hmisc: Harrell Miscellaneous*. Manual. https://CRAN.R-project.org/package=Hmisc.

Henry, Lionel, and Hadley Wickham. 2020a. *Purrr: Functional Programming Tools*. Manual. https://CRAN.R-project.org/package=purrr.

Henry, Lionel, and Hadley Wickham. 2020b. *Rlang: Functions for Base Types and Core r and "tidyverse" Features*. Manual. https://CRAN.R-project.org/package=rlang.

Henry, Lionel, Hadley Wickham, and Winston Chang. 2020. *Ggstance: Horizontal "ggplot2" Components*. Manual. https://CRAN.R-project.org/package=ggstance.

Hothorn, Torsten. 2019. *TH.Data: TH's Data Archive*. Manual. https://CRAN.R-project.org/package=TH.data.

Hothorn, Torsten, Frank Bretz, and Peter Westfall. 2008. "Simultaneous Inference in General Parametric Models." *Biometrical Journal* 50 (3): 346–363. doi:10/b5f2gc.

Hoyt, A. C. Del Re & W. T. 2014. *MAd: Meta-Analysis with Mean Differences*. Manual. https://cran.r-project.org/package=MAd.

J, Lemon. 2006. "Plotrix: A Package in the Red Light District of R." *R-News* 6 (4): 8–12.

Lüdecke, Daniel. 2019. *Esc: Effect Size Computation for Meta Analysis (Version 0.5.1)*. Manual. doi:10.5281/zenodo.1249218.

Mahmoudian, Mehrad. 2020. *Varhandle: Functions for Robust Variable Handling*. Manual. https://CRAN.R-project.org/package=varhandle.

Mathur, Maya B., Rui Wang, and Tyler J. VanderWeele. 2019. *MetaUtility: Utility Functions for Conducting and Interpreting Meta-Analyses*. Manual. https://CRAN.R-project.org/package=MetaUtility.

Müller, Kirill. 2018. *Bindrcpp: An "rcpp" Interface to Active Bindings*. Manual. https://CRAN.R-project.org/package=bindrcpp.

Müller, Kirill, and Hadley Wickham. 2020. *Tibble: Simple Data Frames*. Manual. https://CRAN.R-project.org/package=tibble.

Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes*. Manual. https://CRAN.R-project.org/package=RColorBrewer.

Novomestky, Frederick. 2012. *Matrixcalc: Collection of Functions for Matrix Calculations*. Manual. https://CRAN.R-project.org/package=matrixcalc.

Pustejovsky, James. 2020. *ClubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections*. Manual. https://CRAN.R-project.org/package=clubSandwich.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria. https://www.R-project.org/.

Revelle, William. 2020. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Manual. Evanston, Illinois. https://CRAN.R-project.org/package=psych.

Rinker, Tyler W., and Dason Kurkiewicz. 2018. *pacman: Package Management for R*. Manual. Buffalo, New York. http://github.com/trinker/pacman.

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with r*. New York: Springer. http://lmdvr.r-forge.r-project.org.

Schafer, Juliane, Rainer Opgen-Rhein, Verena Zuber, Miika Ahdesmaki, A. Pedro Duarte Silva, and Korbinian Strimmer. 2017. *Corpcor: Efficient Estimation of Covariance and (Partial) Correlation*. Manual. https://CRAN.R-project.org/package=corpcor.

Soetaert, Karline. 2018. *Shape: Functions for Plotting Graphical Shapes, Colors*. Manual. https://CRAN.R-project.org/package=shape.

Stephens, Jeremy, Kirill Simonov, Yihui Xie, Zhuoer Dong, Hadley Wickham, Jeffrey Horner, reikoch, Will Beasley, Brendan O'Connor, and Gregory R. Warnes. 2020. *Yaml: Methods to Convert r Data to YAML and Back*. Manual. https://CRAN.R-project.org/package=yaml.

Terry M. Therneau, and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Tierney, Luke, A. J. Rossini, Na Li, and H. Sevcikova. 2018. *Snow: Simple Network of Workstations*. Manual. https://CRAN.R-project.org/package=snow.

Ushey, Kevin. 2021. *Renv: Project Environments*. Manual. https://CRAN.R-project.org/package=renv.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. 4th ed. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the metafor Package." *Journal of Statistical Software* 36 (3): 1–48. doi:10/gckfpj.

Wei, Taiyun, and Viliam Simko. 2017. *R Package "Corrplot": Visualization of a Correlation Matrix*. Manual. https://github.com/taiyun/corrplot.

Westgate, Martin J. 2019. *Revtools: Tools to Support Evidence Synthesis* (version 0.4.1). https://CRAN.R-project.org/package=revtools.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations*. Manual. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley. 2020. *Forcats: Tools for Working with Categorical Variables (Factors)*. Manual. https://CRAN.R-project.org/package=forcats.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. doi:10/ggddkj.

Wickham, Hadley, and Jennifer Bryan. 2020. *Usethis: Automate Package and Project Setup*. Manual. https://CRAN.R-project.org/package=usethis.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. Manual. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Lionel Henry. 2020. *Tidyr: Tidy Messy Data*. Manual. https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Jim Hester, and Winston Chang. 2020. *Devtools: Tools to Make Developing r Packages Easier*. Manual. https://CRAN.R-project.org/package=devtools.

Wickham, Hadley, Jim Hester, and Romain Francois. 2018. *Readr: Read Rectangular Text Data*. Manual. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization*. Manual. https://CRAN.R-project.org/package=scales.

Wilke, Claus O. 2020. *Cowplot: Streamlined Plot Theme and Plot Annotations for "Ggplot2"* (version 1.1.0). https://CRAN.R-project.org/package=cowplot.

Wright, Kevin. 2019. *Pals: Color Palettes, Colormaps, and Tools to Evaluate Them* (version 1.6). https://CRAN.R-project.org/package=pals.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC. https://yihui.org/knitr/.

Xie, Yihui. 2016. *Bookdown: Authoring Books and Technical Documents with R Markdown*. Boca Raton, Florida: Chapman and Hall/CRC. https://github.com/rstudio/bookdown.

Xie, Yihui, J.J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman and Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Zeileis, Achim, and Yves Croissant. 2010. "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software* 34 (1): 1–13. doi:10/gd3vrb.

Appendix C: Additional Information about the Systematic Review and Coding

Databases and search terms used

Table C1. Details about systematic review searches. Note the specific indices available can vary for some databases depending on the institutional subscription. The first two authors discovered this inadvertently after getting a different number of items for identical searches from what they thought were the same databases.

Database	Number of Documents	Search Terms Used	Indices	Comments (filenames are italicized)
SCOPUS via UT Dallas	956	((ALL ("situation* awareness") AND (performance OR decision*)) AND (((association OR regression OR link OR correlation OR validity))) AND ((participant OR user OR cognitive OR cognition OR behavior OR behavioral OR "human factors" OR psychology OR experiment OR study))) AND ((spam OR sagat OR sabars OR sart OR mars OR sars OR "subject matter expert" OR observer OR "*confidence" OR cars)) AND TITLE-ABS-KEY-AUTH ("situation* awareness") AND (LIMIT-TO (LANGUAGE , "English"))	N/A	*scopus.ris* Manually fixed errors when this file was imported into Zotero.

Database	Number of Documents	Search Terms Used	Indices	Comments (filenames are italicized)
SAGE	360	for [All "situation* awareness"] AND [[All "performance"] OR [All "decision*"]] AND [[All "correlation"] OR [All "regression"] OR [All "link"] OR [All "association"] OR [All "validity"]] AND [Abstract "situation* awareness"]	N/A	*SAGE_abstract_p1.ris* *SAGE_abstract_p2.ris* *SAGE_abstract_p3.ris* *SAGE_abstract_p4.ris*
	149	for [All "situation* awareness"] AND [[All "performance"] OR [All "decision*"]] AND [[All "correlation"] OR [All "regression"] OR [All "link"] OR [All "association"] OR [All "validity"]] AND [[All "spam"] OR [All "sagat"]]	N/A	*SAGE_no_abs_p1.ris* *SAGE_no_abs_p2.ris* SAGE was primarily used to search HFES conference proceeding papers.
PsycINFO	711	TX "situation* awareness" AND (TX(decision* OR performance))	N/A	*PsycINFO Part2.ris*
	119	AB("situation* awareness") AND (TX(association OR regression OR link OR correlation OR validity))	N/A	*PsycINFO Part1.ris*

Database	Number of Documents	Search Terms Used	Indices	Comments (filenames are italicized)

Web of Science via UT Arlington	517	(ALL=("situation* awareness") AND ALL=("performance" OR "decision*") AND ALL=("association" OR "regression" OR "link" OR "correlation" OR "validity")) AND LANGUAGE: (English)	Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
			Timespan=All years
				WoS_savedrecs_517_Part1_full_record.bib
				WoS_savedrecs_517_Part2_full_record.bib

	281	(ALL=("situation* awareness") AND ALL=("performance" OR "decision*") AND ALL=("SPAM" OR "SAGAT" OR "SABARS" OR "SART" OR "MARS" OR "subject matter expert" OR "observer" OR "confidence" OR "CARS")) AND LANGUAGE: (English)	Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC
			Timespan=All years
				WoS_savedrecs_281_full_record.bib

Database	Number of Documents	Search Terms Used	Indices	Comments (filenames are italicized)
	46	CF=("Applied Human Factors and Ergonomics") AND ALL=("situation* awareness") AND ALL=("performance" OR "decision*")		*WoS_savedrecs_46_AHFE.bib* Relevant AHFE papers that should have appeared in the two previous two searches did not. This may have been a bug at the time the searches were run.
Google Scholar via Publish or Perish	300	"situation awareness" AND (decision OR performance) AND (regression OR correlation)		*GS search 1 PoPCites.bib*
	300	"situation awareness" AND (SPAM OR SAGAT OR MARS OR SABARS OR SART)		*GS search 2 PoPCites.bib* These two searches were performed while connected to a Virtual Private Network (VPN) to avoid customized search recommendations.
DTIC	789	("situation* awareness" AND "performance" AND "decision*") AND (SPAM OR SAGAT OR SART OR SABARS OR "mission awareness rating scale" OR "situation awareness rating scales") AND (association OR regression OR link OR correlation OR validity)		*DTIC_search.csv* Publicly available documents only (distribution unlimited).

An earlier preprint of this work (Bakdash et al. 2020, version 2) and Endsley (2019) were used to check and refine the coverage of the searches for the current systematic review. In total, these searches produced 5,312 documents for potential inclusion in the meta-analysis. An additional 2 documents were obtained from citation tracing, for a total of 5,314. An initial review for duplicates reduced this number to 3,128 documents, which then underwent abstract review by three or more raters.

Abstract review

Raters coded each abstract as ‘exclude, ‘include’, or ‘maybe,' based on several rules of thumb. For example, abstracts indicating non-empirical work (e.g. literature review, simulated data), analysis of exclusively team-level data, or with no mention of SA in the title or abstract were excluded. Abstracts that explicitly mentioned quantitative measurement of both SA and performance were generally included. Those abstracts that mentioned SA and performance with no statement that both were quantitatively analyzed, and those that were otherwise difficult to code, were typically coded as ‘maybe.’

Full-text review

The full-text for each of the 628 papers that passed abstract screening was reviewed by two or more raters using the following codes:

0: exclude (see Table C2 below for detailed exclusion reasons)
1: include – new, a paper not found in the earlier preprint (Bakdash et al. 2020)
2: include – previously included (a paper included in the earlier preprint)
3: not sure, need to discuss
4: exclude – overfit model (results obtained from treating repeated measures data as independent)

Disagreements among raters, or papers coded as ‘**3**,’ were resolved in discussion on a separate consensus sheet. All such disagreements were able to be resolved. In one instance, a single document with conference proceedings contained one paper that met inclusion criteria and multiple others that did not; in this case, the entire document was coded as an include.

A total of 77 documents were included in the final meta-analysis. Of those documents, 32 were included in the previous preprint and 45 were new additions. The total number of documents included here was substantially higher than both the earlier preprint (Bakdash et al. 2020) of this paper and Endsley (2019): respectively, approximately double and 70% more papers. This shows the two prior systematic reviews did not adequately cover the relevant literature. Table C2 describes the detailed exclusion reasons for the remaining 551 documents (note that some documents met multiple exclusion criteria; for simplicity, only the first reason was used to construct the table).

Table C2. Detailed Exclusion Reasons. Number of papers for each exclusion reason and an example for each one.

Exclusion Reason	Literature Search: Number of Papers	Example
1) No association for SA and performance reported	290	Separate analyses of SA and performance (Marusich et al. 2016)
2) No measure of task performance in an experiment	67	Scores of elite fighter pilots on a test of stress, situational awareness, and cognitive ability (e.g., working memory capacity; O'Hare 1997)
3) SA is a label for performance or no measure of SA	56	Assessing SA using simulated flight performance (Andre, Wickens, and Moorman 1991); note this paper was published before clear distinctions were made between SA and performance
4) Theoretical, conceptual, or narrative review; simulated data	49	Narrative literature review of SA and other human factors constructs (Parasuraman, Sheridan, and Wickens 2008)
5) Teams	23	SA and performance in teams of pilots in simulated air combat (Endsley 1990)
6) Thesis, dissertation, book chapter, or book	2	Book chapter on expertise and SA (Endsley 2006)
7) Could not be found	1	The abstract suggested possible inclusion, but the full-text document could not be found through inter-library loan (Montano, McDermid, and Cairns 2011)
8) Statistical		
a) Overfitting	26	Repeated measures of SA and performance by participant were incorrectly modeled as independent. For example, Strybel et al. (2008) reports a sample size of $N = 13$ but correlations with more than 50 degrees of freedom

Exclusion Reason	**Literature Search: Number of Papers**	**Example**
b) Could not calculate correlation from statistics	6	No inferential statistics reported (Georg et al. 2018)
9) Not written in English	8	Critical portions of the results were written in a foreign language, impeding raters' ability to identify and code relevant effects (Jung and Myung 2008)
10) Duplicate	23	Paper used identical data/results to another paper found in the review (Zhang, Kaber, and Hsiang 2008)
Total	551	

SA assessment methods

In addition to coding SA Measures in each paper, we also coded the SA Assessment Method that each paper used, see Table C3 below. These methods are categorical labels for how SA was assessed. They include the same data (papers and their effects) as the breakdown by SA measure in Table 2 in the paper. As was the case for Table 2 in the paper, because some papers used multiple SA measures they also had multiple assessment methods (also see Table D1). Thus, a single paper can also be represented in Table C3 below multiple times.

Table C3. Assessment methods for SA measures. This includes the typical SA measure for each assessment method and the median sample size, number of papers, and number of effects for each assessment method.

Assessment Method	**Typical SA Measure**	**Median Sample Size (*N*)**	**Number of Papers**	**Number of Effects (*k*)**
Freeze probe	SAGAT, General Knowledge, Other	20	35	171
Observer rating	SABARS, MARS, SARS	15.5	5	24
Post-trial probe	Explicit Recall, General Knowledge, SAGAT	24	16	122

Assessment Method	**Typical SA Measure**	**Median Sample Size (*N*)**	**Number of Papers**	**Number of Effects (*k*)**
Post-trial self-rating	SART, MARS, Other	23	30	155
Real-time probe	SPAM, General Knowledge, Other	21	19	206

Appendix D: References and Summary for Papers included in the Meta-analysis

References for 77 papers included in the meta-analysis are shown below. Unindexed papers (i.e., papers found only in Google Scholar or DTIC) are denoted by *; this includes technical reports and conference papers. Note two technical reports (Bowden and Loft 2013; Durso, Hackworth, and Truitt 1999) were also published as journal articles (Loft et al. 2015; Durso et al. 1998), respectively.

Adams-White, Jade E., Jacqueline M. Wheatcroft, and Michael Jump. 2018. "Measuring Decision Accuracy and Confidence of Mock Air Defence Operators." *Journal of Applied Research in Memory and Cognition* 7 (1): 60–69. doi:10/ggp3ff.

Albina, A.R. 2019. "Assessing the Impact of a GIS for Improving Novice Crisis Decision-Making." *25th Americas Conference on Information Systems*. doi:http://doi.org/10.1007/s13398-014-0173-7.2.

Barros, PG De, RW Lindeman, and ... 2011. "Enhancing robot teleoperator situation awareness and performance using vibro-tactile and graphical feedback." *2011 IEEE Symposium on 3D User Interfaces (3DUI)*. doi:10/bcg4c9.

*Bowden, Vanessa K., and Shayne Loft. 2013. *Situation Awareness Measurement Techniques for Submarine Track Management*. https://apps.dtic.mil/dtic/tr/fulltext/u2/a580215.pdf.

Cha, Jackie S., Nicholas E. Anton, Tomoko Mizota, Julie M. Hennings, Megan A. Rendina, Katie Stanton-Maxey, Hadley E. Ritter, Dimitrios Stefanidis, and Denny Yu. 2019. "Use of Non-Technical Skills Can Predict Medical Student Performance in Acute Care Simulated Scenarios." *The American Journal of Surgery* 217 (2): 323–328. doi:10/ggqr3j.

Clark, Hallie, Anne Collins McLaughlin, and Jing Feng. 2017. "Situational Awareness and Time to Takeover: Exploring an Alternative Method to Measure Engagement with High-Level Automation." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61 (1): 1452–1456. NA. doi:10/gfkwdn.

Cooper, S, T McConnell-Henry, R Cant, and ... 2011. "Managing deteriorating patients: registered nurses' performance in a simulated setting." *The Open Nursing* 5: 120–126. doi:10/fzchfz.

Cooper, Simon, Leigh Kinsman, Penny Buykx, Tracy McConnell-Henry, Ruth Endacott, and Julie Scholes. 2010. "Managing the Deteriorating Patient in a Simulated Environment: Nursing Students Knowledge, Skill and Situation Awareness." *Journal of Clinical Nursing* 19 (15–16): 2309–2318. doi:10/cxx2n7.

Crooks, C. L., Chang-Ya Hu, and Robert P. Mahan. 2001. "Cue Utilization and Situation Awareness during a Simulated Experience." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 45 (22): 1563–1567. doi:10/fx39vt.

Cummings, M. L., and Stephanie Guerlain. 2007. "Developing Operator Capacity Estimates for Supervisory Control of Autonomous Vehicles." *Human Factors* 49 (1): 1–15. doi:10/b68h7z.

de Winter, J. C. F., Y. B. Eisma, C. D. D. Cabrall, P. A. Hancock, and N. A. Stanton. 2018. "Situation awareness based on eye movements in relation to the task environment." *Cognition, Technology & Work* 21 (1): 99–111. doi:10/ggqfw6.

Durso, F. T., M. Kathryn Bleckley, and Andrew R. Dattel. 2006. "Does Situation Awareness Add to the Validity of Cognitive Tests?" *Human Factors* 48 (4): 721–733. doi:10/bqsh4q.

*Durso, F. T., C. A. Hackworth, T. R. Truitt, J. Crutchfield, D. Niklic, and C. A. Manning. 1999. *Situation Awareness as a Predictor of Performance in En Route Air Traffic Controllers.*

Federal Aviation Administration, Office of Aviation Medicine. 1999-03375-001. https://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/1990s/media/AM99-03.pdf

Entin, Eileen B. 2000. “An Exploratory Investigation of Relationships between Situation Awareness and Performance in an Attack Helicopter Domain.” In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 44 (1):113–115. doi:10/fzdx8p.

*Fracker. 1991. “Measures of Situation Awareness: An Experimental Evaluation.” https://apps.dtic.mil/dtic/tr/fulltext/u2/a262732.pdf

Gatsoulis, Y, and Gurvinder S. Virk. 2007. “Performance metrics for improving human robot interaction.” *Advances In Climbing And Walking Robots*” 716–725 doi:10/d3t67r.

Gatsoulis, Yiannis, Gurvinder S. Virk, and Abbas A. Dehghani-Sanij. 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/cv4vkz.

Gregoriades, A, and A Sutcliffe. 2018. “Simulation-based evaluation of an in-vehicle smart situation awareness enhancement system.” *Ergonomics*: 61 (7): 947–965. doi:10/ggqgt4.

Grigoleit, Tristan, Hector Silva, Mary Ann Burress, Dan Chiappe, SM Cetiner, P Fechtelkotter, and M Legatt. 2017. “Toward a Descriptive Measure of Situation Awareness in Petrochemical Refining.” In *Advances in Human Factors in Energy: Oil, Gas, Nuclear and Electric Power Industries* 495: 3–14. doi:10/ggqrxp.

Gugerty, Leo J. 1997. “Situation Awareness during Driving: Explicit and Implicit Knowledge in Dynamic Spatial Memory.” *Journal of Experimental Psychology: Applied* 3 (1): 42–66. doi:10/bhfsfj.

Gugliotta, A., P. Ventsislavova, P. Garcia-Fernandez, E. Pea-Suarez, E. Eisman, D. Crundall, and C. Castro. 2017. “Are Situation Awareness and Decision-Making in Driving Totally Conscious Processes? Results of a Hazard Prediction Task.” *Transportation Research Part F: Traffic Psychology and Behaviour* 44: 168–179. doi:10/ggqf3z.

Gutzwiller, Robert S., and Benjamin A. Clegg. 2013. “The Role of Working Memory in Levels of Situation Awareness.” *Journal of Cognitive Engineering and Decision Making* 7 (2): 141–154. doi:10/ggqhj6.

Hogan, Michael P., David E. Pace, Joanne Hapgood, and Darrell C. Boone. 2006. “Use of Human Patient Simulation and the Situation Awareness Global Assessment Technique in Practical Trauma Skills Assessment.” *The Journal of Trauma* 61 (5): 1047–1052. doi:10/c37j3g.

Jannat, Mafruhatul, David S. Hurwitz, Christopher Monsere, and Kenneth H. Funk. 2018. “The Role of Driver’s Situational Awareness on Right-Hook Bicycle-Motor Vehicle Crashes.” *Safety Science* 110 (December): 92–101. doi:10/gd4vnf.

Jeon, M., B.N. Walker, and T.M. Gable. 2014. “Anger Effects on Driver Situation Awareness and Driving Performance.” *Presence: Teleoperators and Virtual Environments* 23 (1): 71–89. doi:10/f55d8x.

Jipp, Meike, and Phillip L. Ackerman. 2016. “The Impact of Higher Levels of Automation on Performance and Situation Awareness: A Function of Information-Processing Ability and Working-Memory Capacity.” *Journal of Cognitive Engineering and Decision Making* 10 (2): 138–166. doi:10/ggqhj5.

Johnson, Vanessa, Robert J. Pleban, and Jennifer S. Tucker. 2009. “Investigating the Effects of Desktop Computer Simulation Training on Situation Awareness (SA) and Adaptive

Decision-Making Skills." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 53 (18): 1196–1200. doi:10/fzr872.

Jung, D., S. Jo, and R. Myung. 2008. "A Study of Relationships between Situation Awareness and Presence That Affect Performance on a Handheld Game Console." In *ACM Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology*: 240–243. doi:10/ddr5nz.

Kaber, D., Y. Zhang, S. Jin, P. Mosaly, and M. Garner. 2012. "Effects of Hazard Exposure and Roadway Complexity on Young and Older Driver Situation Awareness and Performance." *Transportation Research Part F: Traffic Psychology and Behaviour* 15 (5): 600–611. doi:10/f37vxc.

Kaber, David B., and Mica R. Endsley. 2004. "The Effects of Level of Automation and Adaptive Automation on Human Performance, Situation Awareness and Workload in a Dynamic Control Task." *Theoretical Issues in Ergonomics Science* 5 (2): 113–153. doi:10/dnmhc2.

Kass, Steven J., Lisa A. VanWormer, William L. Mikulas, Shauna Legan, and David Bumgarner. 2011. "Effects of Mindfulness Training on Simulated Driving: Preliminary Results." *Mindfulness* 2 (4): 236–241. doi:10/ckz28f.

Kraemer, Jan, and Heinz-Martin Süß. 2015. "Real Time Validation of Online Situation Awareness Questionnaires in Simulated Approach Air Traffic Control." *Procedia Manufacturing* 3: 3152–3159. doi:10/ggqrxs.

Lafond, Daniel, Michel B. DuCharme, Jean-Franois Gagnon, and Sbastien Tremblay. 2012. "Support requirements for cognitive readiness in complex operations." *Journal of Cognitive Engineering and Decision Making* 6 (4): 393–426. doi:10/ggqf7p.

*Laptaned, U. 2006. "Situation awareness in virtual environments: A theoretical model and investigation with different interface designs." In *Proceedings of the 9th IASTED International Conference on Computers and Advanced Technology in Education* 167–174. http://eprints.utcc.ac.th/962/.

Lehtonen, Esko, Heidi Sahlberg, Emilia Rovamo, and Heikki Summala. 2017. "Learning game for training child bicyclists situation awareness." *Accident Analysis and Prevention* 105: 72–83. doi:10/gbm285.

Lin, C.J., T.-L. Hsieh, and S.-F. Lin. 2013. "Development of Staffing Evaluation Principle for Advanced Main Control Room and the Effect on Situation Awareness and Mental Workload." *Nuclear Engineering and Design* 265: 137–144. doi:10/f5qd2w.

Lo, Julia C., Emdzad Sehic, Karel A. Brookhuis, and Sebastiaan A. Meijer. 2016. "Explicit or Implicit Situation Awareness? Measuring the Situation Awareness of Train Traffic Controllers." *Transportation Research Part F: Traffic Psychology and Behaviour* 43 (November): 325–338. doi:10/f3t2hz.

Loft, Shayne, Lisa Jooste, Yanqi Ryan Li, Timothy Ballard, Samuel Huf, Ottmar V. Lipp, and Troy A. W. Visser. 2018. "Using Situation Awareness and Workload to Predict Performance in Submarine Track Management: A Multilevel Approach." *Human Factors* 60 (7): 978–991. doi:10/ggqf5q.

Loft, Shayne, Daniel B. Morrell, and Samuel Huf. 2013. "Using the Situation Present Assessment Method to Measure Situation Awareness in Simulated Submarine Track Management." *International Journal of Human Factors and Ergonomics* 2 (1): 33. doi:10/ggx42m.

Loft, Shayne, Daniel B. Morrell, Kate Ponton, Janelle Braithwaite, Vanessa Bowden, and Samuel Huf. 2016. "The Impact of Uncertain Contact Location on Situation Awareness and

Performance in Simulated Submarine Track Management." *Human Factors* 58 (7): 1052–1068. doi:10/f88wbj.

Lukosch, Heide, Daan Groen, Shalini Kurapati, Roland Klemke, and Alexander Verbraeck. 2016. "The Role of Awareness for Complex Planning Task Performance: A Microgaming Study." *International Journal of Game-Based Learning* 6 (2): 15–28. doi:10/ggqf5j.

*Matthews, Michael D., and Scott A. Beal. 2002. *Assessing Situation Awareness in Field Training Exercises*. http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA408560.

McDermott, Patricia L., and Alia Fisher. 2013. "Methodologies for Assessing Situation Awareness of Unmanned System Operators." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57: 167–171. doi:10/ggqrxx.

Miles, J.D., and T.Z. Strybel. 2017. "Measuring Situation Awareness of Student Air Traffic Controllers with Online Probe Queries: Are We Asking the Right Questions?" *International Journal of Human-Computer Interaction* 33: 55–65. doi:10/ggqf93.

Mogford, RH. 1997. "Mental models and situation awareness in air traffic control." *The International Journal of Aviation Psychology* 7 (4): 331–341. doi:10/bcpnmv.

Nickel, Courtney, Carolyn Knight, Aaron Langille, and Alison Godwin. 2019. "How Much Practice Is Required to Reduce Performance Variability in a Virtual Reality Mining Simulator?" *Safety* 5 (18): 2–11. doi:10/ggqf3m.

O'Brien, K. S., and D. O'Hare. 2007. "Situational Awareness Ability and Cognitive Skills Training in a Complex Real-World Task." *Ergonomics* 50 (7): 1064–1091. doi:10/d36bgj.

O'Hagan, A.D., J. Issartel, A. Wall, F. Dunne, P. Boylan, J. Groeneweg, M. Herring, M. Campbell, and G. Warrington. 2019. "Flying on Empty Effects of Sleep Deprivation on Pilot Performance." *Biological Rhythm Research* 51 (7). 1133–1154. doi:10/ggqk52.

Onal, E., J. Schaffer, J. O'Donovan, L. Marusich, M.S. Yu, C. Gonzalez, and T. Höllerer. 2014. "Decision-Making in Abstract Trust Games: A User Interface Perspective." In *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*: 21–27. doi:10/ggqrx7.

Paletta, Lucas, Amir Dini, Cornelia Murko, Saeed Yahyanejad, Michael Schwarz, Gerald Lodron, Stefan Ladstätter, Gerhard Paar, and Rosemarie Velik. 2017. "Towards Real-Time Probabilistic Evaluation of Situation Awareness from Human Gaze in Human-Robot Interaction." In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, 247–248. doi:10/ggwkjh.

*Pierce, Russell S., Thomas Z. Strybel, and Kim-Phuong L. Vu. 2008. "Comparing Situation Awareness Measurement Techniques in a Low Fidelity Air Traffic Control Simulation." In *Proceedings of the 26th International Congress of the Aeronautical Sciences (ICAS), Anchorage, AS*. http://icas.org/ICAS_ARCHIVE/ICAS2008/PAPERS/579.PDF

*Pleban, Robert J. 2009. *Training Situation Awareness and Adaptive Decision-Making Skills Using a Desktop Computer Simulation*. https://apps.dtic.mil/dtic/tr/fulltext/u2/a494799.pdf.

Puuska, Samir, Lauri Rummukainen, Jussi Timonen, Lauri Lääperi, Markus Klemetti, Lauri Oksama, and Jouko Vankka. 2018. "Nationwide Critical Infrastructure Monitoring Using a Common Operating Picture Framework." *International Journal of Critical Infrastructure Protection* 20 (March): 28–47. doi:10/gdfq73.

Riley, Jennifer M., David B. Kaber, and John V. Draper. 2004. "Situation Awareness and Attention Allocation Measures for Quantifying Telepresence Experiences in

Teleoperation." *Human Factors and Ergonomics in Manufacturing* 14 (1): 51–67. doi:10/bj8s83.

Riley, Jennifer M., and Laura D. Strater. 2006. "Effects of Robot Control Mode on Situation Awareness and Performance in a Navigation Task." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50 (3): 540–544. doi:10/fzs7kv.

Rogers, Meghan, Yu Zhang, David Kaber, Yulan Liang, Shruti Gangakhedkar, and D Harris. 2011. "The Effects of Visual and Cognitive Distraction on Driver Situation Awareness." In *International Conference on Engineering Psychology and Cognitive Ergonomics* 6781: 186–195. doi:10/d6wmjb.

Rose, J. A., C. Bearman, J. Dorrian, and NA Stanton. 2013. "An Evaluation of the Low-Event Task Subjective Situation Awareness (LETSSA) Technique." In *Advances in Human Factors and Ergonomics Series*, 690–703. ISBN: 978-1-4398-7124-9

Rose, J., C. Bearman, and J. Dorrian. 2018. "The Low-Event Task Subjective Situation Awareness (LETSSA) Technique: Development and Evaluation of a New Subjective Measure of Situation Awareness." *Applied Ergonomics* 68: 273–282. doi:10/ggqhkw.

Salmon, Paul M., Neville A. Stanton, Guy H. Walker, Daniel Jenkins, Darshna Ladva, Laura Rafferty, and Mark Young. 2009. "Measuring Situation Awareness in Complex Systems: Comparison of Measures Study." *International Journal of Industrial Ergonomics*, *Selected papers from ECCE 2007, the 25th Anniversary Conference of the European Conference on Cognitive Ergonomics*, 39 (3): 490–500. doi:10/d5wjgh.

Saus, Evelyn-Rose, Bjrn Helge Johnsen, Jarle Eid, and Julian F. Thayer. 2012. "Who benefits from simulator training: Personality and heart rate variability in relation to situation awareness during navigation training." *Computers in Human Behavior* 28: 1262–1268. doi:10/gf36g7.

Saus, Evelyn-Rose, Bjørn H. Johnson, Jarle Eid, Per K. Riisem, Rune Anderson, and Julian F. Thayer. 2006. "The Effect of Brief Situational Awareness Training in a Police Shooting Simulator: An Experimental Study." *Military Psychology* 18(Suppl.): S3–S21. doi:10/c6pv5m.

Schuster, D., J.R. Keebler, J. Zuniga, and F. Jentsch. 2012. "Individual Differences in SA Measurement and Performance in Human-Robot Teaming." In *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA*:187–190.. doi:10/ggqrzs.

*Selcon, S. J., and R. M. Taylor. 1990. "Evaluation of the Situational Awareness Rating Technique (SART) as a Tool for Aircrew Systems Design." In *AGARD, Situational Awareness in Aerospace Operations 8 p(SEE N 90-28972 23-53).* https://apps.dtic.mil/dtic/tr/fulltext/u2/a223939.pdf.

*Sollenberger, Randy L., and Earl S. Stein. 1995a. "A Simulation Study of Air Traffic Controller Situational Awareness." https://apps.dtic.mil/dtic/tr/fulltext/u2/a522540.pdf.

*Sollenberger, Randy L., and Earl S. Stein. 1995b. *The Effects of Structured Arrival and Departure Procedures on TRACON Air Traffic Controller Memory and Situational Awareness.* DOT/FAA/CT-TN95/27. FEDERAL AVIATION ADMINISTRATION TECHNICAL CENTER. https://apps.dtic.mil/docs/citations/ADA303800.

*Stanners, Melinda, and Han T. French. 2005. *An Empirical Study of the Relationship between Situation Awareness and Decision Making*. DSTO-TR-1687. DEFENCE SCIENCE AND TECHNOLOGY ORGANIZATION EDINBURGH (AUSTRALIA) LAND OPERATIONS DIV. https://apps.dtic.mil/docs/citations/ADA434593.

*Strater, Laura D., Mica R. Endsley, Robert J. Pleban, and Michael D. Matthews. 2001. *Measures of Platoon Leader Situation Awareness in Virtual Decision-Making Exercises*. DTIC Document. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA390238.

Strybel, Thomas Z., Kim-Phuong L. Vu, Jerome Kraft, and Katsumi Minakata. 2008. "Assessing the Situation Awareness of Pilots Engaged in Self Spacing." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52: 11–15. NA. doi:10/fzs5gq.

Sulistyawati, Ketut, Yoon Ping Chui, and D Harris. 2009. "Confidence Bias in Situation Awareness." In , 5639:317–325. SPRINGER-VERLAG BERLIN. doi:10/bxpfgj.

Sulistyawati, Ketut, Christopher D. Wickens, and Yoon Ping Chui. 2011. "Prediction in Situation Awareness: Confidence Bias and Underlying Cognitive Abilities." *International Journal of Aviation Psychology* 21 (2): 153–174. doi:10/b4mh79.

*Taylor, R. M., S. J. Selcon, and A. D. Swinden. 1995. ""Measurement of Situational Awareness and Performance- A Unitary SART Index Predicts Performance on a Simulated ATC Task." In *Human Factors in Aviation Operations: Proceedings of the 21st Conference for Aviation Psychology (EAAP) Volume 3*, edited by R. Fuller, N. Johnston, and N. McDonald: 275–280.

*Valentine, Nick, Alexander Wearing, and Mary Omodei. 2007. *Resource Utilisation and Situational Awareness in a Computer Simulated Decision Task: A Pilot Study*. https://apps.dtic.mil/docs/citations/ADA473106.

*Venturino, Michael, William L. Hamilton, and Stephen R. Dvorchak. 1990. "Performance-Based Measures of Merit for Tactical Situation Awareness." *AGARD, Situational Awareness in Aerospace Operations 5 p(SEE N 90-28972 23-53).* https://apps.dtic.mil/dtic/tr/fulltext/u2/a223939.pdf

Visser, Troy A. W., Angela D. Bender, Vanessa K. Bowden, Stephanie C. Black, Jayden Greenwell-Barnden, Shayne Loft, and Ottmar V. Lipp. 2019. "Individual Differences in Higher-Level Cognitive Abilities Do Not Predict Overconfidence in Complex Task Performance." *Consciousness and Cognition* 74. doi:10/ggqhk9.

Wijayanto, T, S Wibirama, ZZ Maryoto, and ... 2016. "Effects of morning-night differences and sleep deprivation on situation awareness and driving performance." In *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. doi:10/ggqr2f.

Wojtusch, J., D. Taubert, T. Graber, and K. Nergaard. 2019. "Evaluation of Human Factors for Assessing Human-Robot Interaction in Delayed Teleoperation." In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*: 3787–3792. doi:10/ggqr2g.

Yang, Chih-Wei, Tsung-Ling Hsieh, Shiau-Feng Lin, Chiuhsiang Joe Lin, and Hui-Ming Teng. 2011. "Operators' Signal-Detection Performance in Video Display Unit Monitoring Tasks of the Main Control Room." *Safety Science* 49 (10): 1309–1313. doi:10/b8t7bf.

Table D1 shows summary information for included papers.

Table D1. Summary information for included literature by paper/experiment. This table includes the author and year of each included paper/experiment and additional summary information (sample size, sample size from statistics, number of effects, number of SA measures, and SA measure type(s).

Author	**Year**	**Median *N***	**Median *N* from Statistics**	**Number of Repeated Measures (*k*)**	**Number of SA Measures**	**SA Measure Type(s)**
Adams-White	2018	60	60	1	1	SART
Albina	2019	28	28	2	1	SART
Barros	2011	27	27	3	1	General knowledge
Bowden	2013	171	171	32	3	SPAM, SART, SAGAT
Cha	2019	41	41	1	1	General knowledge
Clark	2017	50	32	1	1	SAGAT
Cooper	2010	51	51	1	1	SAGAT
Cooper	2011	35	35	2	1	General knowledge
Crooks	2001	165	165	4	2	SAGAT, SART
Cummings and Guerlain	2007	42	42	2	2	SPAM, SAGAT
de Winter	2018	86	86	1	1	General knowledge
Durso	2006	88	84	6	1	SPAM
Durso	1998	12	12	2	2	SPAM, SAGAT
Entin	2000	24	24	18	2	SAGAT, General knowledge
Fracker	1991	32	32	6	1	Other
Gatsoulis et al.	2010	30	30	5	2	Other, SAGAT

Author	Year	Median *N*	Median *N* from Statistics	Number of Repeated Measures (*k*)	Number of SA Measures	SA Measure Type(s)
Gatsoulis	2007	16	16	1	1	General knowledge
Gregoriades	2018	17	17	3	1	SAGAT
Grigoleit	2017	11	11	48	2	SAGAT, SPAM
Gugerty (exp 1)	1997	34	34	5	2	Explicit Recall, Other
Gugerty (exp 2)	1997	79	79	5	2	Explicit Recall, Other
Gugerty (exp 3)	1997	77	77	5	2	Explicit Recall, Other
Gugliotta	2017	121	121	3	1	General knowledge
Gutzwiller	2013	99	99	10	1	Other
Hogan	2006	16	16	1	1	SAGAT
Jannat	2018	51	51	2	1	SAGAT
Jeon	2014	35	30	11	1	General knowledge
Jipp	2016	57	57	3	1	SAGAT
Johnson	2009	35	35	6	1	MARS
Jung	2008	30	30	1	1	SAGAT
Kaber	2004	24	24	12	1	SAGAT
Kaber	2012	20	20	1	1	SPAM
Kass et al.	2011	16	16	4	1	SAGAT
Kraemer	2015	57	57	6	2	SAGAT, SPAM
Lafond (exp 1)	2012	10	10	2	1	General knowledge
Lafond (exp 2)	2012	40	40	1	1	General knowledge

Author	Year	Median *N*	Median *N* from Statistics	Number of Repeated Measures (*k*)	Number of SA Measures	SA Measure Type(s)
Laptaned	2006	60	60	2	2	SAGAT, SART
Lehtonen	2017	80	80	6	1	Explicit Recall
Lin	2013	60	60	4	1	SART
Lo et al. (exp 1)	2016	9	9	12	2	MARS, SAGAT
Lo et al. (exp 2)	2016	20	20	15	1	SAGAT
Loft	2016	50	50	20	1	SPAM
Loft	2013	55	55	21	2	SPAM, SART
Loft	2018	59	59	6	1	SAGAT
Lukosch	2016	142	107	1	1	SART
Matthews	2002	15.5	15.5	8	2	MARS, SABARS
McDermott	2013	39	39	6	1	Other
Miles	2017	84	84	14	1	SPAM
Mogford	1997	37	34	6	1	General knowledge
Nickel	2019	20	18	2	1	General knowledge
O'Brien	2007	28	18	15	1	SAGAT
OHagan	2019	7	7	1	1	General knowledge
Onal	2014	95	95	1	1	SAGAT
Paletta	2017	12	12	3	1	SAGAT
Pierce	2008	21	21	49	3	SART, SAGAT, SPAM
Pleban	2009	35	35	12	1	MARS

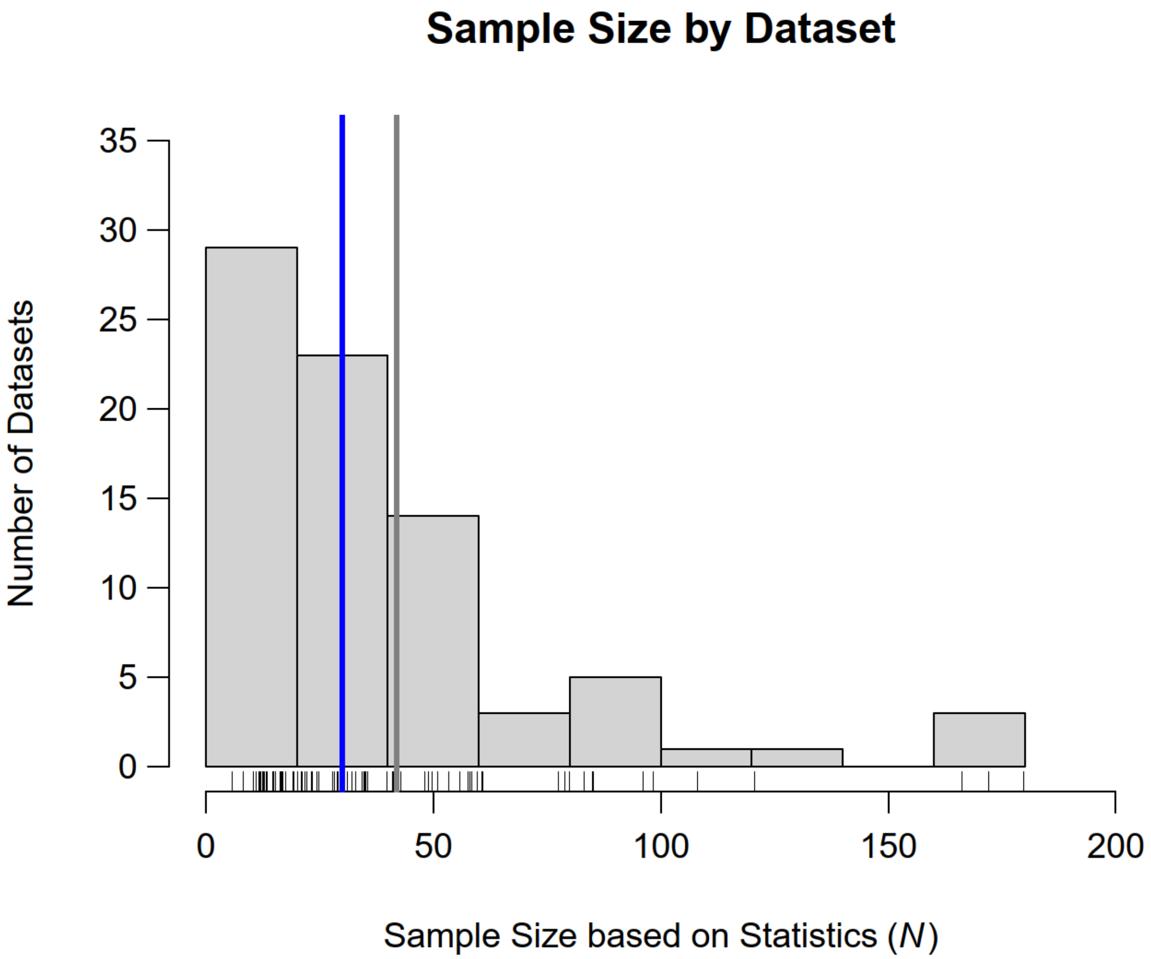
Author	Year	Median *N*	Median *N* from Statistics	Number of Repeated Measures (*k*)	Number of SA Measures	SA Measure Type(s)
Puuska	2018	13	13	16	2	SAGAT, SART
Riley	2004	24	24	4	1	SAGAT
Riley	2006	20	20	1	1	SAGAT
Rogers	2011	20	20	64	1	General knowledge
Rose	2013	23	23	4	2	SAGAT, Other
Rose	2018	26	26	1	1	Other
Salmon	2009	20	20	8	2	SART, SAGAT
Saus	2006	40	40	32	2	SARS, SABARS
Saus	2012	36	36	2	1	SARS
Schuster	2012	53	53	2	2	SPAM, SART
Selcon	1990	12	12	3	1	SART
Sollenberger	1995a	16	16	58	2	Explicit Recall, Direct-SR
Sollenberger (same data as above ref)	1995b	16	16	58	2	Explicit Recall, Direct-SR
Stanners	2005	24	24	2	1	SAGAT
Strater	2001	14	14	21	3	SABARS, SAGAT, Other
Strybel	2008	13	13	2	1	SPAM
Sulistyawati	2010	16	16	5	2	SAGAT, Other
Sulistyawati and Chui (same data as above ref)	2010	16	16	5	2	SAGAT, Other
Taylor	1995	12	12	6	1	SART

Author	**Year**	**Median *N***	**Median *N* from Statistics**	**Number of Repeated Measures (*k*)**	**Number of SA Measures**	**SA Measure Type(s)**
Valentine	2007	16	14	9	2	SAGAT, Other
Venturino	1990	16	16	4	2	Direct-SR, Other
Visser	2019	180	180	4	2	SPAM, SART
Wijayanto	2016	12	12	3	1	SAGAT
Wojtusch	2019	28	28	3	1	SART
Yang et al.	2011	13	13	3	1	SART

Appendix E: Dataset Characteristics

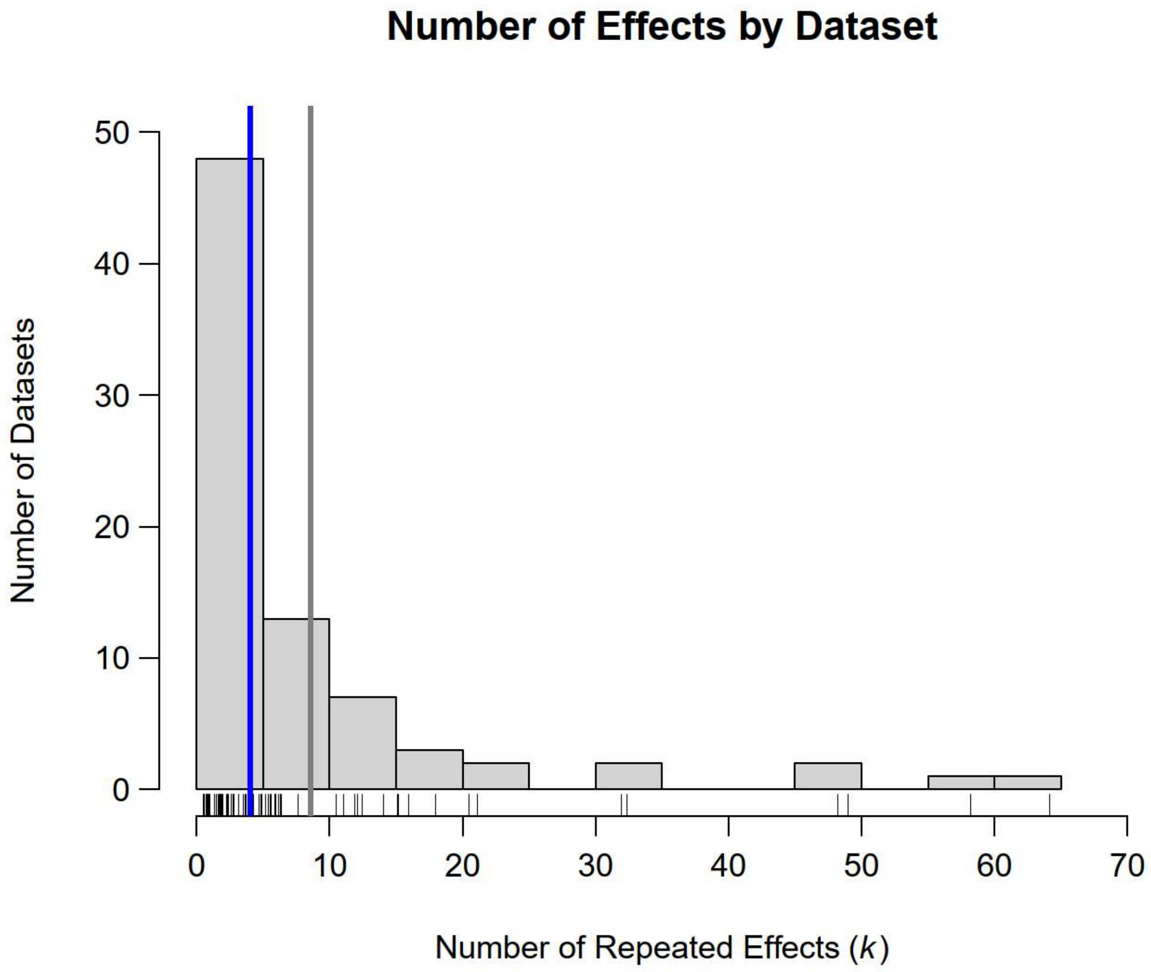
The median sample size from statistics for 'papers' (unique experiments or datasets) was *N* = 30.00 and the mean was *N* = 41.85: See Figure E1.

Figure E1. Sample size by dataset. The x-axis is the sample size based on statistics for each dataset and the y-axis the number of datasets. Each tick mark above the x-axis represents the sample size for each dataset. The vertical blue line is the median sample size and the dark grey line is the mean.



Most datasets (64 out of 79) assessed situation awareness (SA) and performance more than once (Figure E2). The median number of assessments of SA and performance per dataset was *k* = 4.00 and the mean was *k* = 8.58.

Figure E2. Number of effects by dataset. The x-axis is the number of repeated measures for each dataset, and the y-axis shows the number of datasets. Each tick mark above the x-axis represents the number of repeated measures for each dataset. The vertical blue line is the median sample size and the dark grey line is the mean.

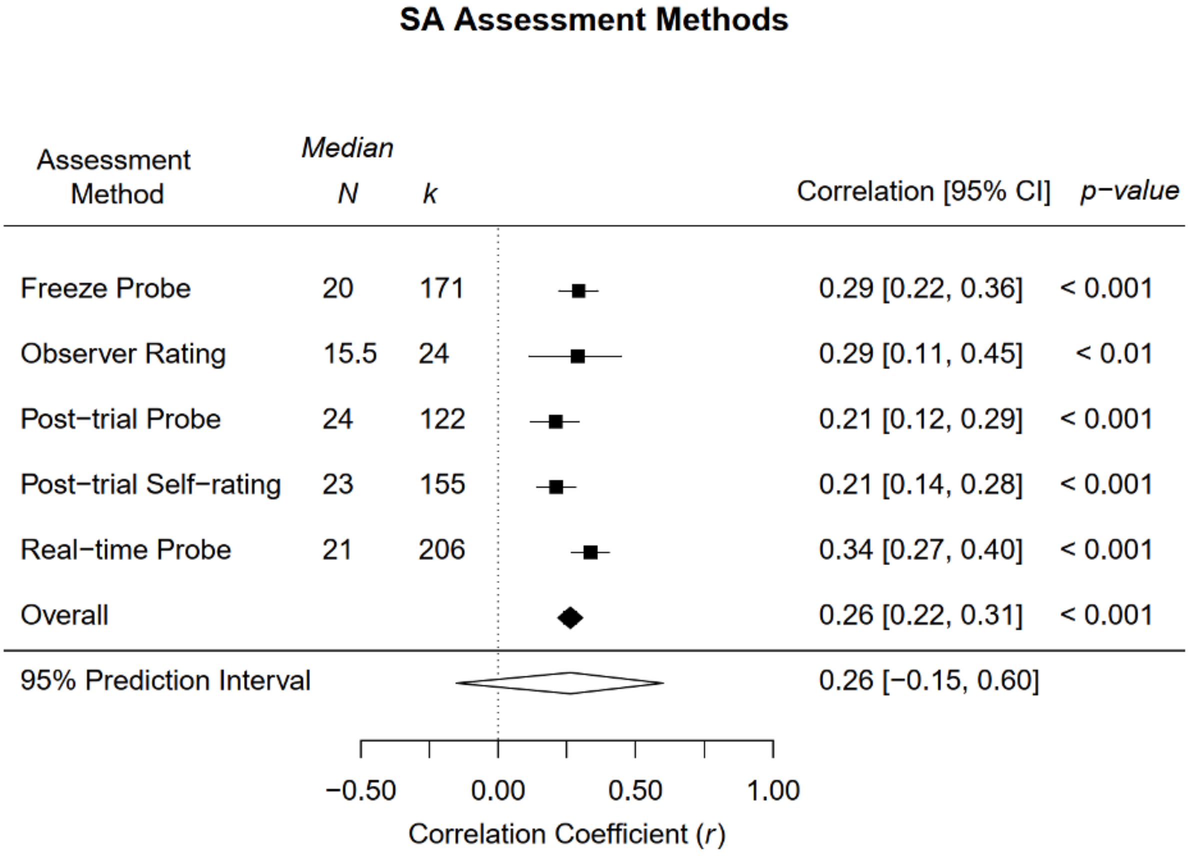


Appendix F: Additional Analyses

SA assessment methods

Figure F1 shows a forest plot of the results of a meta-analysis using SA assessment methods (i.e., how SA was assessed). This model uses the same data as the SA measures meta-analytic model presented previously in the paper, but instead has SA assessment methods as a moderator instead of labels for SA measures. All five SA assessment methods had mean effects significantly greater than zero, approximately small to medium in magnitude.

Figure F1. Forest plot of SA assessment methods and performance. Symbols are the same as Figure 4 in the paper.



Meta-Analytic Means: Post-Hoc Comparisons

We also performed exploratory post-hoc comparisons among meta-analytic means, both by SA measure and SA assessment methods (Tables F1 and F2, respectively), using Tukey's Honestly Significant Difference with a False Discovery Rate adjustment. There were a large number of comparisons with 45 by SA measure and 10 by SA assessment.

There were some significant differences among meta-analytic means by SA measure and one for assessment method. However, there were no discernible patterns that particular SA measures, or assessment methods, consistently had significantly higher or lower means in these comparisons. Consequently, the results do not clearly show specific SA measures or assessment methods that regularly have statistically higher mean correlations with performance than other ones.

Table F1. SA measures post-hoc comparisons. The estimate is the difference between the two effect sizes, SE is the pooled standard error, and the Z-value and p-value are the inferential results.

Comparison	**Estimate**	**SE**	**Z-value**	***p*-value**
Explicit Recall - Direct-SR	-0.15	0.01	-17.07	< 0.001
General Know. - Direct-SR	-0.16	0.07	-2.28	0.12
MARS - Direct-SR	-0.19	0.17	-1.09	0.54
SABARS - Direct-SR	-0.07	0.09	-0.79	0.62
SAGAT - Direct-SR	-0.08	0.07	-1.23	0.51
SARS - Direct-SR	-0.04	0.09	-0.49	0.74
SART - Direct-SR	-0.22	0.07	-3.30	< 0.01
SPAM - Direct-SR	-0.09	0.07	-1.32	0.49
Other - Direct-SR	0.06	0.05	1.20	0.51
General Know. - Explicit Recall	-0.01	0.08	-0.18	0.90
MARS - Explicit Recall	-0.04	0.17	-0.23	0.90
SABARS - Explicit Recall	0.07	0.10	0.77	0.62
SAGAT - Explicit Recall	0.07	0.07	0.91	0.58
SARS - Explicit Recall	0.10	0.09	1.14	0.52
SART - Explicit Recall	-0.07	0.07	-1.00	0.55
SPAM - Explicit Recall	0.06	0.07	0.81	0.62
Other - Explicit Recall	0.20	0.05	4.30	< 0.001
MARS - General Know.	-0.03	0.17	-0.15	0.90
SABARS - General Know.	0.09	0.09	1.02	0.55
SAGAT - General Know.	0.08	0.05	1.52	0.44
SARS - General Know.	0.12	0.08	1.46	0.44
SART - General Know.	-0.06	0.05	-1.06	0.54
SPAM - General Know.	0.07	0.06	1.28	0.50
Other - General Know.	0.22	0.10	2.24	0.12

Comparison	Estimate	SE	Z-value	*p*-value
SABARS - MARS	0.11	0.16	0.70	0.66
SAGAT - MARS	0.11	0.17	0.62	0.70
SARS - MARS	0.14	0.16	0.89	0.58
SART - MARS	-0.03	0.17	-0.19	0.90
SPAM - MARS	0.10	0.17	0.58	0.70
Other - MARS	0.24	0.18	1.32	0.49
SAGAT - SABARS	-0.01	0.08	-0.12	0.91
SARS - SABARS	0.03	0.02	1.90	0.26
SART - SABARS	-0.15	0.08	-1.81	0.29
SPAM - SABARS	-0.02	0.08	-0.20	0.90
Other - SABARS	0.13	0.11	1.14	0.52
SARS - SAGAT	0.04	0.07	0.54	0.72
SART - SAGAT	-0.14	0.04	-3.62	< 0.01
SPAM - SAGAT	-0.01	0.04	-0.17	0.90
Other - SAGAT	0.14	0.10	1.45	0.44
SART - SARS	-0.18	0.08	-2.35	0.12
SPAM - SARS	-0.05	0.08	-0.61	0.70
Other - SARS	0.10	0.11	0.91	0.58
SPAM - SART	0.13	0.03	4.68	< 0.001
Other - SART	0.28	0.09	2.91	0.03
Other - SPAM	0.15	0.10	1.52	0.44

Table F2. SA assessment techniques post-hoc comparisons. The estimate is the difference between the two effect sizes, SE is the pooled standard error, and the Z-value and p-value are the inferential results.

Comparison	Estimate	SE	Z-value	*p*-value
Observer Rating - Freeze Probe	-0.01	0.09	-0.06	0.95
Post-trial Probe - Freeze Probe	-0.09	0.06	-1.54	0.31
Post-trial Self-rating - Freeze Probe	-0.09	0.04	-1.97	0.16
Real-time Probe - Freeze Probe	0.05	0.05	0.97	0.55
Post-trial Probe - Observer Rating	-0.09	0.1	-0.87	0.55
Post-trial Self-rating - Observer Rating	-0.08	0.08	-1.04	0.55
Real-time Probe - Observer Rating	0.05	0.09	0.61	0.68
Post-trial Self-rating - Post-trial Probe	0.01	0.06	0.09	0.95
Real-time Probe - Post-trial Probe	0.14	0.05	2.72	0.03
Real-time Probe - Post-trial Self-rating	0.13	0.03	4.68	< 0.001

Sensitivity Analyses: CRVE Estimator, Correlation in Sampling Error, and Alternative Calculations for Ghost Results

To evaluate robustness of meta-analytic results under different assumptions, we conducted three sensitivity analyses. In general, the results appear to be robust: under varying assumptions results did not meaningfully change.

First, we assessed an alternative CRVE estimator, for the SA measures meta-analytic model, the CR2 estimator using the *R* package *clubSandwich* (Pustejovsky 2017); see Table F4. The CR2 estimator is recommended for unbalanced moderators (which we have here, see Table 2 in the paper; e.g., there were 170 effect sizes for SAGAT but only 30 effect sizes for MARS) and small sample sizes (Pustejovsky and Tipton 2018); also see Appendix E.

With the CR2 estimator, SA measures sometimes had slightly larger *p*-values than the previously presented model using the CR1p estimator; although, critically, significance for the overall mean did not change. The two different calculations for CR2 Satterthwaite and Saddlepoint *p*-values produced similar results. This is important because the CR2 estimator had insufficient degrees of freedom (< 4) for some SA measures, so the Satterthwaite *p*-values should be interpreted with caution.

Table F3. CRVE Adjustments and *p*-values for the overall effect size. We compare no CRVE adjustment to CR1p and two CR2 adjustment methods. This sensitivity analysis uses the same data reported in the main paper, it includes the draw of ghost results.

CRVE Adjustment:	**None**	**CR1p**	**CR2**		
SA Measure	***p*-value**	***p*-value**	**Satterthwaite *p*-value**	**Saddlepoint *p*-value**	**Degrees of Freedom**
Direct-SR	< 0.001	< 0.001	0.03	< 0.01	1.67
Explicit Recall	< 0.01	< 0.001	0.01	< 0.01	5.36
General Know.	< 0.001	< 0.001	< 0.01	< 0.001	8.38
MARS	0.05	0.25	0.36	0.36	3.10
SABARS	< 0.01	< 0.001	< 0.05	0.04	2.68
SAGAT	< 0.001	< 0.001	< 0.001	< 0.001	33.16
SARS	< 0.01	< 0.001	< 0.05	0.03	2.24
SART	< 0.001	< 0.001	< 0.001	< 0.001	14.47
SPAM	< 0.001	< 0.001	< 0.001	< 0.001	12.54
Other	< 0.001	< 0.001	< 0.001	< 0.001	12.13
Overall	< 0.001	< 0.001	< 0.001	< 0.001	70.18

Second, we compared different fixed values for the sampling error correlation (Table F4). This sensitivity analysis follows the recommendation by Hedges, Tipton, and Johnson (2010). The analysis shows the correlation in sampling error had minimal impact on estimates for the overall effect size and total heterogeneity. In addition, we estimated two key meta-analytic parameters (the overall effect size and total heterogeneity) using both *metafor* and also the *robumeta R* package (Fisher and Tipton 2015). The two packages use both different estimation methods and different weights, but produced fairly similar results here.

Table F4. Overall meta-analytic effects sizes for different sampling error correlations. For brevity, we do not present confidence intervals here. The overall effect size heterogeneity ($\hat{\tau}$) is a standard deviation expressed as an *r* value. *Estimated parameters for *rho* = 1.00 could not be calculated using *metafor*, so rho = 0.99 was used instead. This sensitivity analysis uses the same data reported in the main paper, it includes the draw of ghost results.

		Sampling Error Correlation					
Model	**Estimated Parameter**	***rho* = 0.00**	***rho* = 0.20**	***rho* = 0.40**	***rho* = 0.60**	***rho* = 0.80**	***rho* = 1.00***
metafor	Overall effect (*r)*	0.26	0.26	0.26	0.26	0.26	0.25
robumeta	Overall effect (*r*)	0.28	0.28	0.28	0.28	0.28	0.28
metafor	Total heterogeneity ($\hat{\tau}$)	0.21	0.20	0.02	0.21	0.23	0.26
robumeta	Total heterogeneity ($\hat{\tau}$)	0.22	0.22	0.22	0.22	0.22	0.22

Third, we compared meta-analytic models for different assumptions to calculate ghost results, including no ghosts (detailed only), see Table F5. The assumptions were as follows:

(1) Draw p-values from the distribution of non-significant effects that were reported in detail (this is the method described in the main paper).

(2) Assign fixed p-values: Assign non-significant p-values using *p* = 0.10, 0.50, or 0.99 (note this produces nonzero effect sizes, with the magnitude dependent on the sample size of the corresponding studies). Fixed *p*-values of 0.10 and 0.99 are implausible but provide useful boundary conditions

(3) Sign of correlation:

(a) Ratio: probabilistically choose signs using the same ratio of positive (74.79%) and negative (25.21%) signs from the non-significant results that were reported in detail (note the probabilities are unlikely to be representative, they are likely biased toward effect sizes with positive signs by selective reporting and other forms of reporting bias), or

(b) Positive: assume all r value have a positive sign (higher SA is always associated with better performance), this again is an unrealistic assumption but provides a boundary condition

(4) No ghost results: Only include effects reported in detail.

Most parameter estimates were more or less similar under different assumptions. There were minor variations in the magnitude of the overall effect. However, differences in magnitude were, at most, approximately a small effect size (*r* = 0.10) and parameter estimates have uncertainty. Heterogeneity estimates were close to comparable across comparisons. Note under all plausible assumptions the heterogeneity neared the estimates of the overall mean.

We again caution that calculating ghost results (for draw and ratio) using the proportion of positive to negative non-significant effects reported in detail may be too conservative. This is because the ghost results are *missing not at random* (note this is the worst possible type of missing data because there is a probabilistic relationship between the value of the data and its omission, see Little and Rubin 2019). Hence, it is plausible the true proportion of selectively omitted results has a substantially greater proportion of negative effects than the detailed non-significant effects.

Table F5. Overall meta-analytic effects sizes for different assumptions about ghost results and no ghosts. As mentioned in the text, some of these assumptions are implausible (e.g., all selectively omitted non-significant effects are positive and/or have a *p*-value of 0.10) and used as boundary conditions. For brevity, we do not present confidence intervals here.

			Ghost Results: Ratio			**Ghost Results: Positive**			**Detailed Effects Only (No Ghost Results)**
Model	**Estimated Parameter**	**Draw/Sample**	***p* = 0.99**	***p* = 0.50**	***p* = 0.10**	***p* = 0.99**	***p* = 0.50**	***p* = 0.10**	----
metafor	Overall effect (*r*)	0.26	0.24	0.25	0.27	0.24	0.27	0.32	~~0.24~~ 0.33
robumeta	Overall effect (*r*)	0.28	0.26	0.27	0.29	0.26	0.29	0.33	~~0.26~~ 0.35
metafor	Total heterogeneity ($\hat{\tau}$)	0.21	0.22	0.21	0.24	0.22	0.19	0.19	~~0.22~~ 0.24
robumeta	Total heterogeneity ($\hat{\tau}$)	0.22	0.23	0.22	0.24	0.23	0.20	0.19	~~0.23~~ 0.22

Visualizations of Proportions of Effects below Key Thresholds: SA Measures and Assessment Techniques

The proportions of effects are visually depicted using raincloud plots in Figures F2 and F3, for SA measures and assessment methods respectively, relative to three key thresholds. No inferential analyses of proportions are reported here because some subsamples had too many effects with high variance and/or over dispersed effects to produce stable calculations with the default calibration using the *prop_stronger* function in the *MetaUtility R* package (Mathur, Wang, and VanderWeele, 2019).

A visual, exploratory examination of Figures F2 and F3 suggests the meta-analytic means often measures overestimate (i.e., exceed 50%) the proportion of effects below. A clear exception is in Figure F2 where the Explicit Recall and SARS SA measures have most effects below their meta-analytic means. Note, for both Figures F2 and F3, the consistent patterns with the majority of effects below the typical cognitive psychology effect size and a large effect size.

Figure F2. Raincloud plot of individual effects by SA measure. Three different effect size thresholds depicted by vertical bars. The interpretation of the individual effects (rain) and the clouds (distribution of effects) is the same as Figure 5 in the paper.

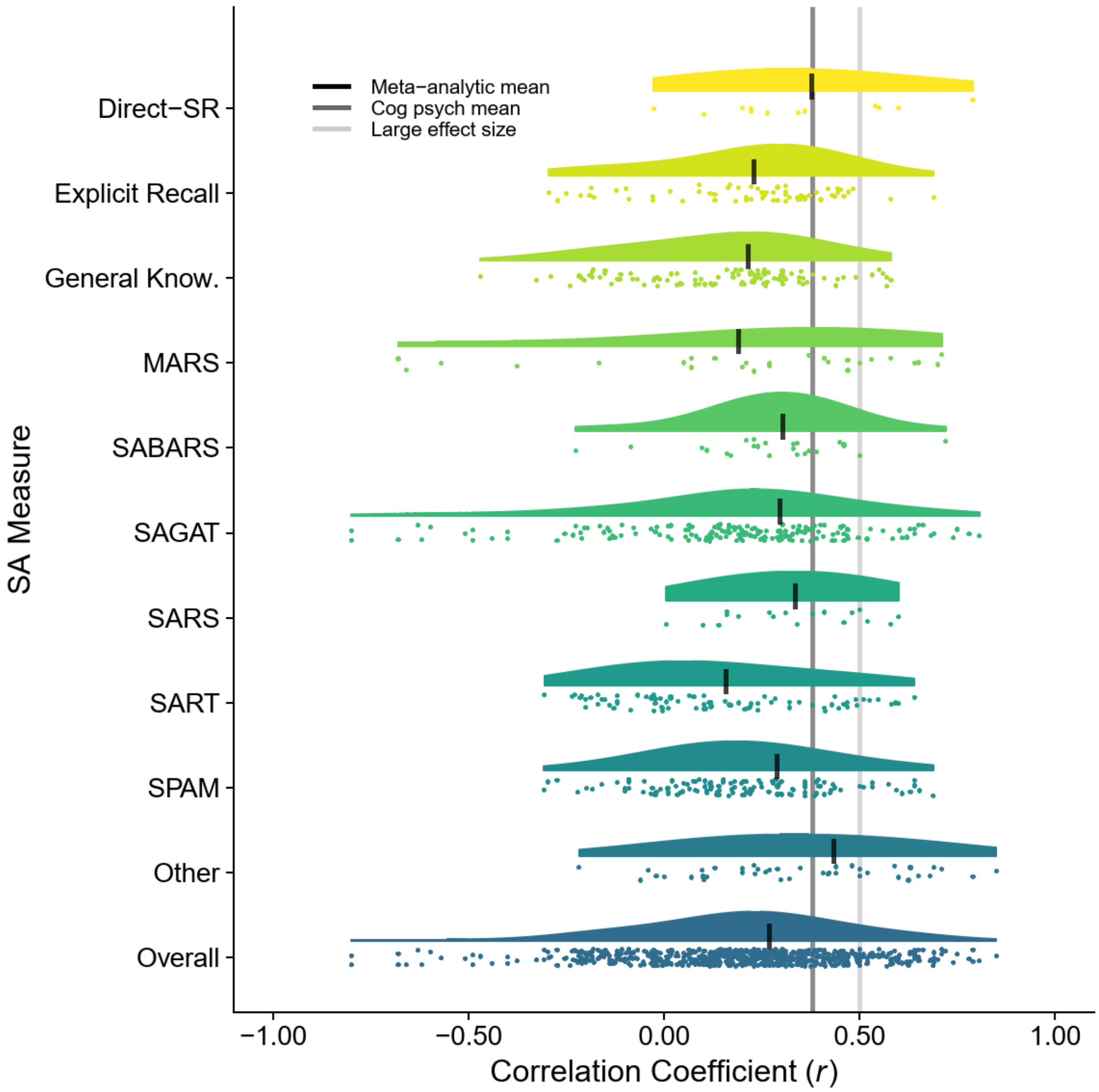
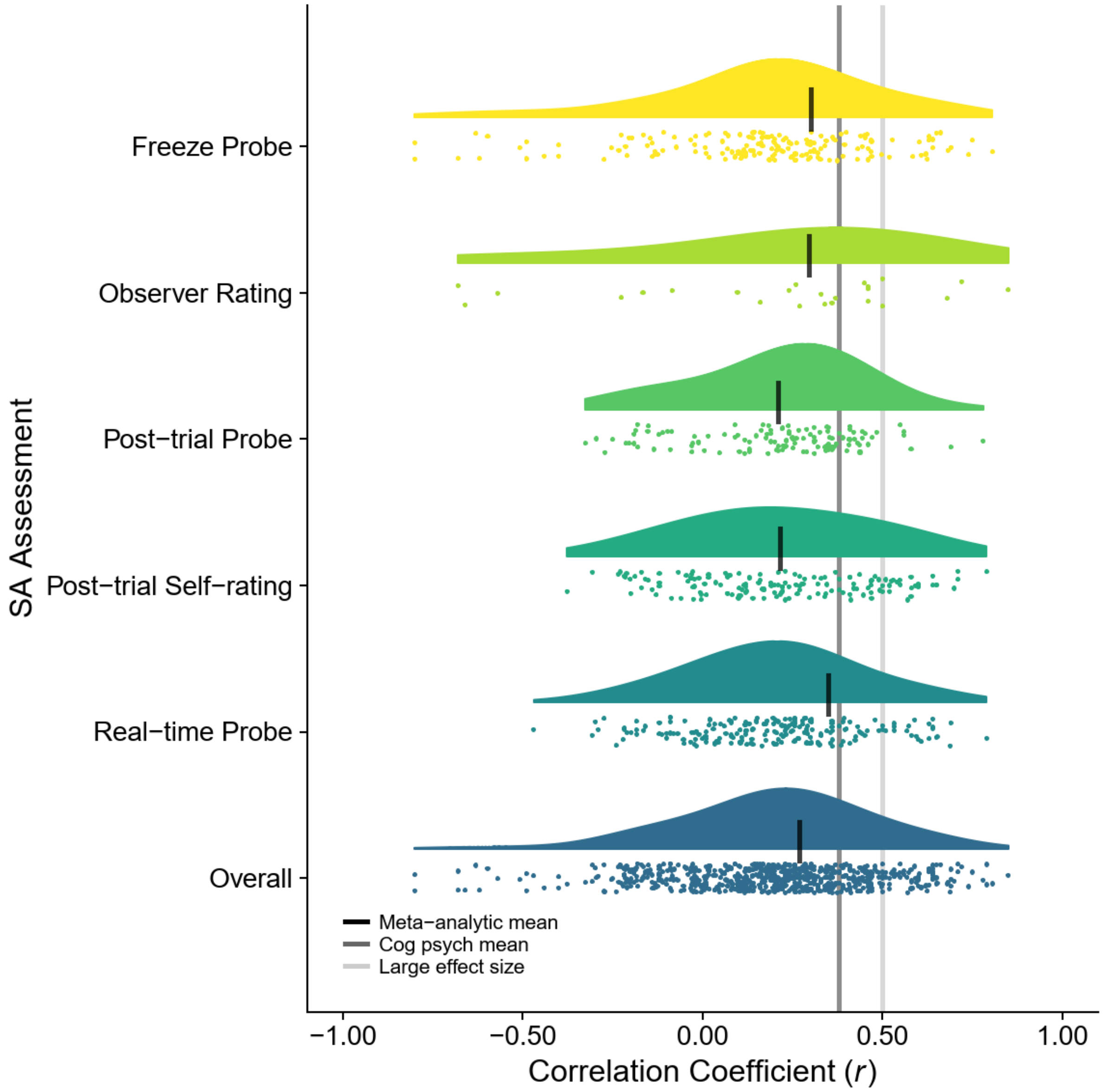


Figure F3. Raincloud plot of individual effects by SA assessment. Three different effect size thresholds depicted by vertical bars. The interpretation of the individual effects (rain) and the clouds (distribution of effects) is the same as Figure 5 in the paper.



Proportions for Detailed Effects Only

As a boundary condition, we also evaluated proportions for detailed effects only by omitting all ghost results (Figure F4 and Table F6). ~~Half~~ About two-thirds of detailed effects (~~50~~64%) were below their overall mean (r = ~~0.25~~0.33), ~~in contrast to~~ nearly identical ~~analyses of all~~ to all results where that~~e~~ overall mean (r = 0.26) ~~tended to~~ also overestimated (65%). For the second threshold, the proportion below for detailed effects (73%) appeared to be less than it was for all effects (86%). ~~However,~~ F~~f~~or the third threshold, the proportion below for detailed effects (90%) ~~neared the proportion below~~ was slightly lower than for all effects with ghosts (96%).

Results with detailed effects provide further evidence that SA has limited validity for performance. Even under the extreme condition of no ghost results, only 27% of detailed effects exceeded typical effect size in cognitive psychology ($r = 0.38$) with just 10% of detailed effects exceeding a large effect ($r = 0.50$).

Figure F4. Raincloud plot of detailed effects only. Three different effect size thresholds are depicted by vertical bars. The interpretation of the individual effects (rain) and the cloud (distribution of effects) is the same as Figured 3 and 5 in the paper.

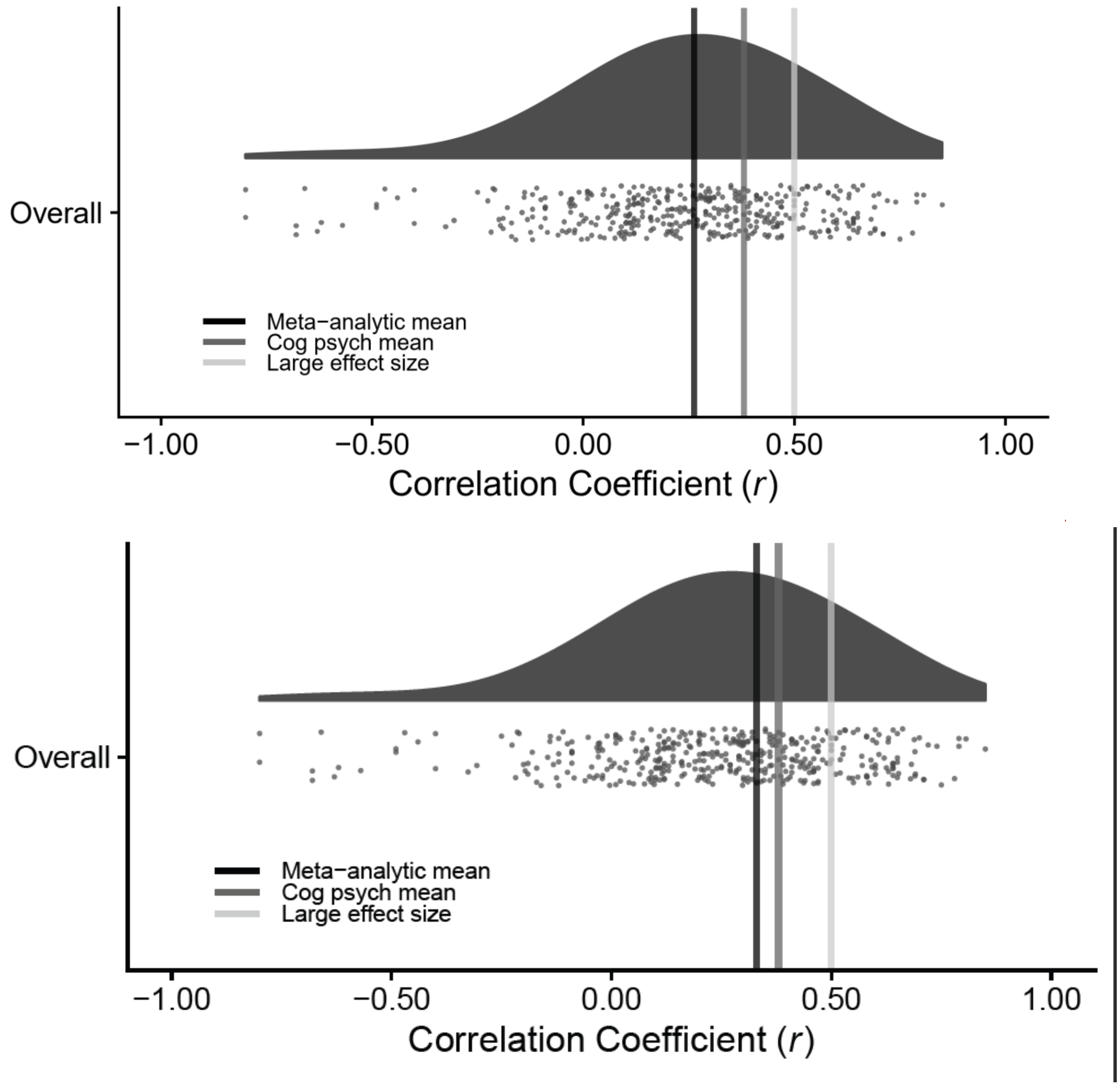


Table F6. Proportions of detailed effects only below three relevant thresholds.

Threshold	Proportion Below (%)	Lower 95% CI	Upper 95% CI
Overall Mean: $r = 0.25$	~~50.36~~ 64.17	~~49.83~~ 51	~~50.85~~ 76

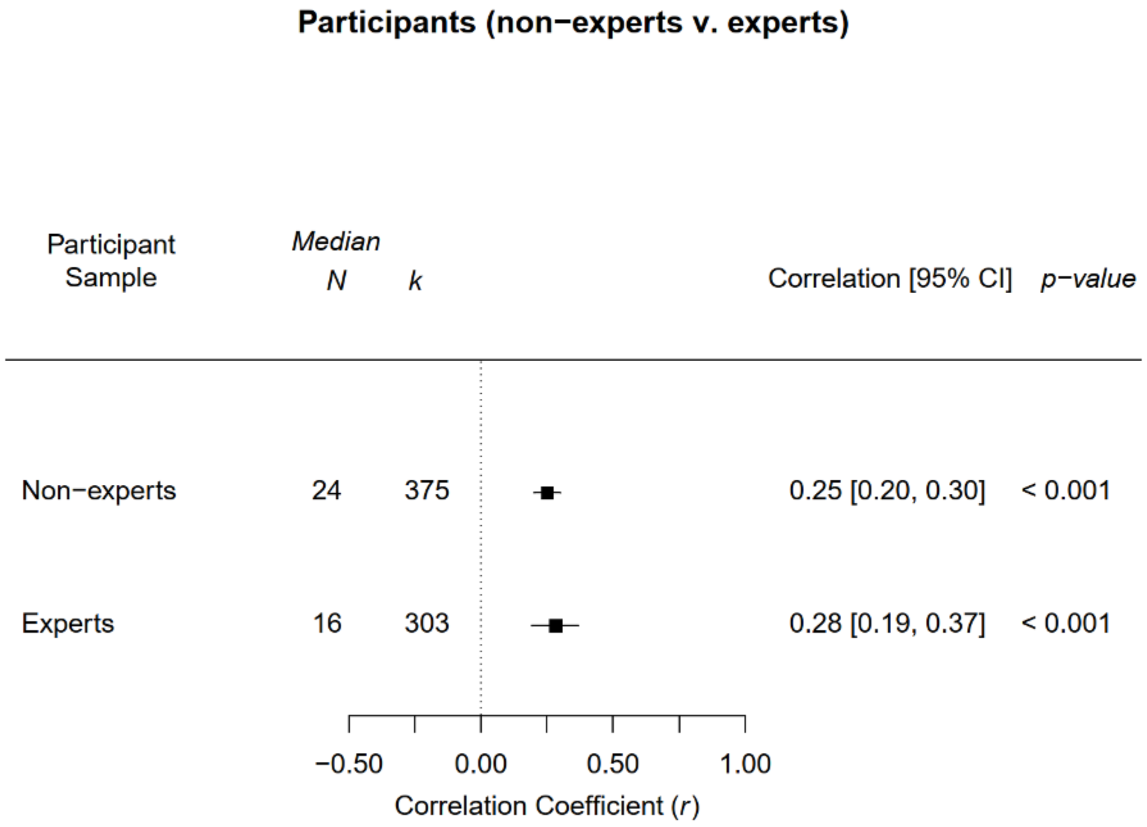
Typical Cognitive Psychology Effect: $r = 0.38$	~~72.70~~ 73.13	~~72.16~~ 61	~~73.24~~ 85
Large effect: $r = 0.50$	~~90.36~~ 90.30	~~89.99~~ 82	~~90.77~~ 96

Non-Expert vs. Expert Participants and Objective vs. Subjective SA

We ran two additional meta-analytic models (including ghost results): Non-expert vs. expert participants and objective vs. subjective SA. Non-experts were defined as not have specialized training and/or experience, they were typically undergraduate students. We operationalized expert participants as having specialized training and/or experience (e.g., air traffic controllers, health care professionals, medical students, nursing students, Soldiers in the military, etc). Examples of objective SA include General knowledge, SAGAT, and SPAM and examples of subjective SA include CARS, LETSSA, MARS, and SARS; see the data dictionary for details.

There was no significant difference in meta-analytic SA-performance associations for non-expert vs. expert participants ($p = 0.56$), see Figure F5. Meta-analytic effects size for each were in the upper end of the small range in terms of magnitude and both were similar to the overall effect of $r = 0.26$.

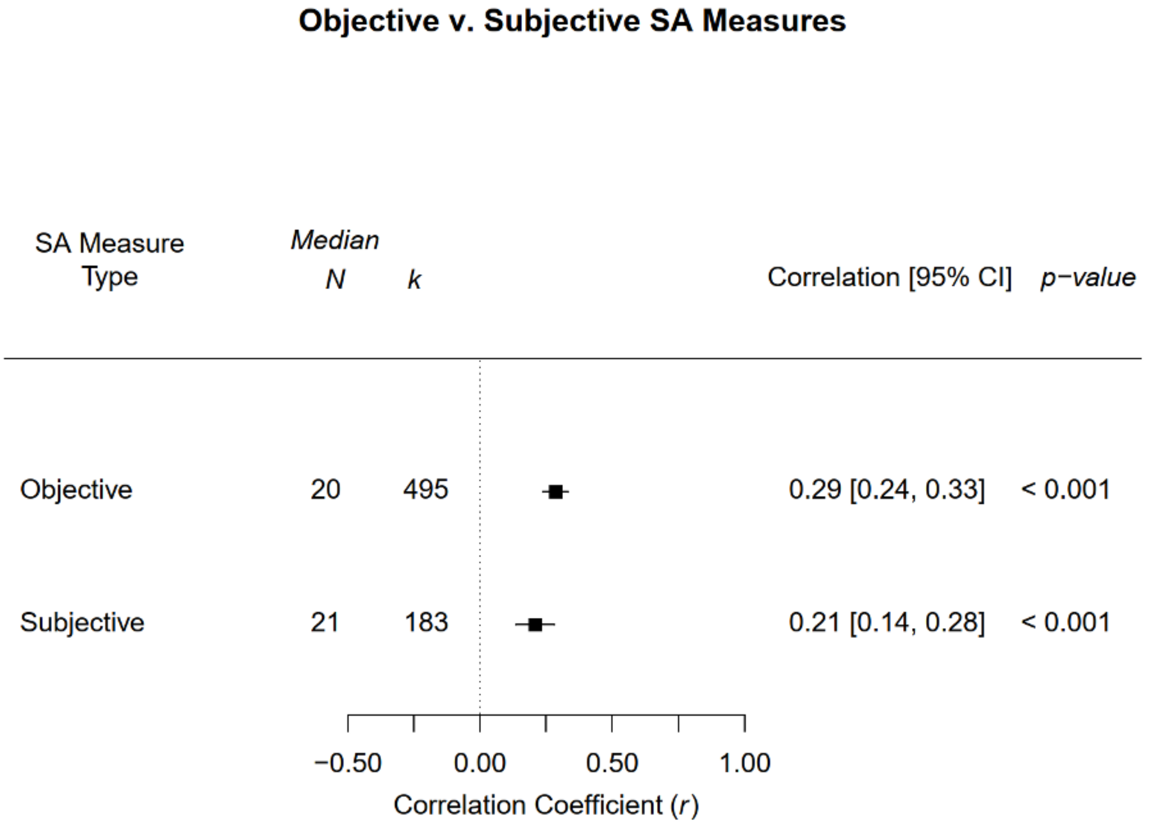
Figure F5. Forest plot of effects for non-experts vs. experts.



We found objective SA ($r = 0.29$) had a significantly stronger ($p = 0.02$) meta-analytic association with performance than subjective SA ($r = 0.21$), see Figure F6.

Nevertheless, the effect size for objective SA only neared a medium effect. Also, subjective SA was only 27% of the data (183 out of 678 effects). Thus, the lower pooled effect size for subjective SA had only a minor impact on the meta-analytic results.

Figure F6. Forest plot of effects for objective vs. subjective SA measures.



Confounds and Other Factors that May Influence Effects

One reviewer raised the point that: "Conclusions are based on correlations. There could be other factors influencing the correlations." We agree it is possible that results could be influenced by other factors, but we did not find strong evidence of confounds in the included papers and instead found partial evidence to the contrary. For restriction of range, we found a ceiling effect was reported for a total of three variables in two papers. Whereas two other papers explicitly assessed and found no evidence SA was statistically confounded with other variables.

To search for potential confounds in all 77 included papers, we used the following search terms: restriction of range, range restriction, floor, ceiling, confound, covariate, moderator, mediator, and hidden. In summary we found (for details see the "*PDF text searches.pdf*" document in Bakdash, Marusich, Cox et al. 2021a):

(1) *Restriction of range* or *range restriction*: No mention.
(2) *Floor*: One paper mentioned avoiding floor effects.
(3) *Ceiling*: Two papers mentioned finding ceiling effects for a total of three variables: One performance measure and two SA measures.

(4) *Confound*: Two papers found no statistical evidence for confounds impacting SA. Several papers mentioned avoiding or minimizing confounds in their experiment design. One paper reported working memory may be confounded with measuring SA. One paper mentioned potential confounds with self-report measures in the context of SA.
(5) *Covariate*: One paper used age as a covariate, another described why a variable was not used as a covariate.
(6) *Moderator* or *Mediator*: No relevant mention.
(7) *ANCOVA*: One paper reported using individual difference measures as covariates for predicting SA and performance.
(8) *Hidden*: No relevant mention.

While it is possible there could be hidden factors, there is a little evidence from the included empirical literature. Similarly, current SA theories do not clear specify there are other variables and/or confounding factors impacting the posited SA and performance link (see Figure 1 in the paper). Our conclusions are based on SA and performance as defined, measured, and reported in the included literature.

Another consideration is that confounds do not necessarily result in diminished effects; instead, confounds can produce inflated effects or may not have a meaningful impact (see Frank, 2000). For example, say a particular measure of SA also assesses some aspects of workload. In this case, when the SA and performance relationship is evaluated it will erroneously include some amount of workload. In such a case the incremental validity of SA, adjusted to remove the measurable effect of workload on performance, has to be lower than the unadjusted, raw correlation.

Appendix References (not cited in the paper or Appendix B or D)

Andre, A. D., C. D. Wickens, and L. Moorman. 1991. "Display Formatting Techniques for Improving Situation Awareness in the Aircraft Cockpit." *The International Journal of Aviation Psychology* 1 (3): 205–218.

Bakdash, Jonathan. Z., Laura R. Marusich, Katherine Cox (Gamble), M. N. Geuss, and E. G. Zaroukian. 2020. "The Validity of Situation Awareness for Performance: A Meta-Analysis (Version 2)." https://psyarxiv.com/kv7n3/.

Bowden, Vanessa K., and Shayne Loft. 2013. *Situation Awareness Measurement Techniques for Submarine Track Management*. https://apps.dtic.mil/dtic/tr/fulltext/u2/a580215.pdf.

Durso, F. T., C. A. Hackworth, and T. R. Truitt. 1999. "Situation Awareness as a Predictor of Performance in En Route Air Traffic Controllers." *FAA Office of Aviation Medicine Reports*, iii–11.

Durso, Francis T., Carla A. Hackworth, Todd R. Truitt, Jerry Crutchfield, Danko Nikolic, and Carol A. Manning. 1998. "Situation Awareness as a Predictor of Performance for En Route Air Traffic Controllers." *Air Traffic Control Quarterly* 6 (1). American Institute of Aeronautics and Astronautics: 1–20. doi:10/gf36hf.

Endsley, Mica R. 1990. "Predictive Utility of an Objective Measure of Situation Awareness." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 34: 41–45. http://pro.sagepub.com/content/34/1/41.short.

Endsley, Mica R. 2006. "Expertise and Situation Awareness." In *The Cambridge Handbook of Expertise and Expert Performance.*, edited by K. Anders Ericsson, Neil Charness, Paul J. Feltovich, Robert R. Hoffman (Eds), 633–651. New York, NY, US: Cambridge University Press. doi:10.1017/CBO9780511816796.036.

Frank, K. A. (2000). "Impact of a confounding variable on a regression coefficient." *Sociological Methods & Research 29* (2): 147-194. doi:10.1177/0049124100029002001

Georg, Jean-Michael, Johannes Feiler, Frank Diermeyer, and Markus Lienkamp. 2018. “Teleoperated Driving, a Key Technology for Automated Driving? Comparison of Actual Test Drives with a Head Mounted Display and Conventional Monitors.” In , NA:3403–3408. doi:10/ggqrxk.

Jung, DH, and RH Myung. 2008. “A study of relationships among situation awareness, presence, and performance on a handheld game console.” *Journal of the Ergonomics Society*. doi:10/bc2kxw.

Little, R. J., & Rubin, D. B. 2019. *Statistical Analysis with Missing Data*. Wiley.

Loft, S, V Bowden, J Braithwaite, DB Morrell, and ... 2015. “Situation Awareness Measures for Simulated Submarine Track Management.” *Human Factors 57* (2): 298-310. doi:10.1177/0018720814545515.

Marusich, Laura R., Jonathan Z. Bakdash, Emrah Onal, Michael S. Yu, James Schaffer, John O'Donovan, Tobias Höllerer, Norbou Buchler, and Cleotilde Gonzalez. 2016. “Effects of Information Availability on Command-and-Control Decision Making: Performance, Trust, and Situation Awareness.” *Human Factors* 58 (2): 301–321. doi:10.1177/0018720815619515.

Montano, Giuseppe, John McDermid, and Paul Cairns. 2011. “Automated Decision Support On-Board Modern Aircraft: A Cognitive Engineering Approach.” *Cognitive Technology*, Selected Papers from the 10th Bi-annual International Conference on Naturalistic Decision Making, 16 (2): 20–32.

O’Hare, David. 1997. “Cognitive ability determinants of elite pilot performance.” *Human Factors 39* (4): 540–552. doi:10/cmqdwn.

Zhang, Tao, David B. Kaber, and Simon M. Hsiang. 2008. “Characterization of Mental Models in a Virtual Reality-Based Multitasking Scenario.” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52: 388–392. doi:10/fxh7xv.