Ergonomics, 2014
Vol. 57, No. 7, 959–972, http://dx.doi.org/10.1080/00140139.2014.910612

Taylor & Francis
Taylor & Francis Group

Inter-rater reliability of query/probe-based techniques for measuring situation awareness

Nathan Lau[a], Greg A. Jamieson[b]* and Gyrd Skraaning Jr.[c]

[a]*Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA;* [b]*Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, ON, M5S 3G8, Canada;* [c]*Industrial Psychology, OECD Halden Reactor Project, Halden, Øsfold, Norway*

(*Received 9 July 2013; accepted 21 March 2014*)

Query- or probe-based situation awareness (SA) measures sometimes rely on process experts to evaluate operator actions and system states when used in representative settings. This introduces variability of human judgement into the measurements that require inter-rater reliability assessment. However, the literature neglects inter-rater reliability of query/probe-based SA measures. We recruited process experts to provide reference keys to SA queries in trials of a full-scope nuclear power plant simulator experiment to investigate the inter-rater reliability of a query-based SA measure. The query-based SA measure demonstrated only 'moderate' inter-rater reliability even though the queries were seemingly direct. The level of agreement was significantly different across pairs of experts who had different levels of exposure to the experiment. The results caution that inter-rater reliability of query/probe-based techniques for measuring SA cannot be assumed in representative settings. Knowledge about the experiment as well as the domain is critical to forming reliable expert judgements.

Practitioner Summary: When the responses of domain experts are treated as the correct answers to the queries or probes of SA measures used in representative or industrial settings, practitioners should take caution in assuming (or otherwise assess) inter-rater reliability of the situation awareness measures.

Keywords: subject-matter-experts; performance assessment; ratings; full-scope simulator; experiments; process control; nuclear

Introduction

Situation awareness (SA) is an essential part of human performance in many domains (see e.g. Jeannot 2000; Rousseau, Tremblay, and Breton 2004; Stanton 2010). Numerous empirical studies employ SA measures to understand cognitive work, to evaluate designs and to assess operator performance in complex systems. Researchers have adopted a multitude of techniques for measuring SA (Salmon et al. 2006), including querying/probing (e.g. Endsley 1995), eye-tracking (e.g. Drøivoldsmo et al. 1998), subjective self-rating (e.g. Taylor 1990) and expert rating (e.g. Neal et al. 1998).

The most commonly adopted SA measures employ the query- or probe-based technique, which assesses declarative knowledge about the situation by directly questioning the operators (or participants in the experiment). Query/probe-based measures have contributed significantly to the empirical foundation of SA (see Fracker 1991; Pew 2000; Jeannot, Kelly, and Thompson 2003; Salmon et al. 2006 for reviews of SA measures). Thus, the validity and reliability of these measures deserve serious attention. However, the literature on individual SA measures or techniques rarely goes beyond initial evaluation based on a subset of psychometric properties. Even prominent SA measures cannot be considered fully validated as some psychometric properties are only judged on secondary results of one or two empirical studies (e.g. Endsley 2000). This is consistent with the caution raised by Stanton and Young (1999, 2003) that validation evidence for human factors methods is generally lacking. The psychometric validation of query/probe-based SA measures requires further empirical research (also see, Salmon et al. 2009). In particular, the reliability of query/probe-based SA measures requires additional attention given the dominant focus on validity in the literature.

The triadic nature of inter-rater reliability

One dimension of measurement reliability is inter-rater reliability and agreement (see e.g. Cohen 1960; Shrout and Lane 2012; Mitchell 1979), which refer to the degree of covariation and agreement between raters on a set of targets (e.g. system states) using a particular scale (e.g. a Likert scale), respectively.[1] Inter-rater reliability assessment is a sound psychometric practice whenever multiple judges provide ratings to form the basis for measurement (e.g. Murphy and Davidshofer 1998).

*Corresponding author. Email: jamieson@mie.utoronto.ca

For work settings where measurement or judgement criteria are difficult to define (e.g. medicine), inter-rater reliability is a prevalent topic of discussion for researchers and practitioners.

Inter-rater reliability has a triadic nature comprising (i) the raters, (ii) the rating instrument/measure and (iii) the target/situation. First, *the raters* differ from one another in background knowledge, interpretation of the rating instruments and judgement of available information. For example, Eckes (2008) conducted an empirical study on writing assessment, identifying six rater types through two-mode clustering analyses. The similarities between raters are estimated by the intra-class correlation coefficient – ρ (Maxwell and Delaney 1990). Training raters on the targets and the measurement instruments often increases ρ, thereby improving inter-rater reliability (e.g. Bernardin and Buckley 1981; Stuhlmann et al. 1999; Miller et al. 2003). Training may adopt a psychophysical approach, such as applying constrained scaling to calibrate the raters on specifics scales (Boring 2003; Boring and West 2008), if manipulation on physical or well-defined dimensions is available and relevant for the judgement. Identifying these well-defined dimensions for calibration in representative settings can be challenging. Some research has examined training for reliability in detail. For instance, Lievens (2001) showed that schema-driven training led to higher inter-rater reliability than data-driven training. Based on their study of observational SA assessment in a representative nuclear process control setting, Patrick et al. (2006) assert that the judges' familiarity with the experimental apparatus (i.e. the full-scope simulator) and scenarios is as critical to the study outcome as the judges' knowledge of the rating instruments.

Second, the measurement *instruments* are developed to guide and quantify the judgement of the raters. Methodological details such as the procedures and formats of the instrument can affect the rating variability. Conway, Jako, and Goodman (1995) conducted a meta-analysis of interview rating studies for personnel selection, discovering an interaction effect that standardising questions and responses improved inter-rater reliability more for interviews conducted by individual raters than for interviews by all raters simultaneously. Jonsson and Svingby (2007) reviewed the literature in education, concluding that topic-specific and analytical scoring rubrics could improve inter-rater reliability.

Third, the *target/situation* dictates the level of cognitive demand on the raters. Viswesvaran, Ones, and Schmidt (1996) and Conway and Huffcutt (1997) conducted meta-analyses on job performance rating studies, discovering that performance categories (e.g. communication competence vs. productivity) and job complexities could lead to significantly different inter-rater reliability. Similar findings on inter-rater reliability differences across performance categories also appear in the education literature (e.g. Ramos, Schafer, and Tracz 2003). Though often manipulable in controlled research settings, the nature of targets associated with low inter-rater reliability, such as diagnosis of complex medical cases, is not changeable in practice. In representative experimental settings, researchers face the challenge of maintaining strict control over scenario developments as participants introduce their own dynamics (Patrick et al. 2006). Thus, some measurements with low inter-rater reliability are simply interpreted with caution as researchers try to advance training and measurement techniques related to the targets of interest. In other cases, measurements are improved through consensus of multiple raters (Stemler 2004). That said, some targets are intrinsically more difficult to judge than others – irrespective of the quality of the raters and instruments employed.

In summary, the degree of inter-rater reliability can be attributed to the characteristics of the raters, measurement methods and the targets.

Query/probe-based SA measures

The literature contains a number of query/probe-based SA measures, several of which target specific domains. The most prominent is the Situation Awareness Global Assessment Technique (SAGAT; Endsley 1988, 2000). Other query/probe-based measures (some of which are proposed as SAGAT variants) include Situation Awareness Control Room Inventory (SACRI; Hogg et al. 1995) for process control; Situation Present Assessment Method (SPAM; Durso et al. 1998), Situation Awareness Verification and ANalysis Tool (SAVANT; Willems and Heiney 2001), Situation Awareness bei Lotsen der Streckenflugkontrolle im kontext von Automatisierung[2] (SALSA; Hauss and Eyferth 2003) and Situation Awareness for Solutions for Human-Automation (SASHA; Jeannot, Kelly, and Thompson 2003) for air traffic control; QUantitative Analysis of Situation Awareness (QUASA; McGuinness 2004) for command and control and Analog SAGAT (ASAGAT; Gatsoulis, Gurvinder, and Dehghani-Sanij 2010) for tele-robotic control. The degree of psychometric evaluation varies across these query/probe-based SA measures, and only secondary empirical data on inter-rater reliability are available for SAGAT, SACRI and SALSA.

Query/probe-based measures share the feature of eliciting and assessing declarative knowledge of the operators (or participants) about the situation by direct questioning either during a pause (i.e. query) or in real-time (i.e. probe) during system (or simulator) operations. For example, QUASA asks operators to answer true or false to statements such as, 'The Commander of the [. . .] Air Force has recently resigned over corruption charges' (McGuinness 2004). The appeal of the technique lies in the clear or 'direct' relationship between the measurements and the SA notion, minimising the need for inferences in the interpretation (see e.g. Endsley 1995).

Query/probe-based measures diverge in their methodological details in the preparation, data collection and analysis phases of a study. The *preparation phase* involves creating a set of questions and defining the time for administering them. Some measures prescribe task analysis to help generate the questions (e.g. SAGAT) while others rely on process experts to create an inventory of questions (e.g. SPAM, SAVANT, SACRI). Some measures prescribe relatively standardised questions and responses to focus on specific domains or SA dimensions (e.g. metacognition in QUASA) whereas others prescribe no restrictions to allow a broad range of domain applications (e.g. SAGAT). Some measures simply augment existing measures. For example, ASAGAT employs SAGAT with continuous (as opposed to categorical) response options and is advocated for tele-robotic control. Finally, SA measures offer differing guidance on query timing. Some measures recommend random selection to minimise cuing (e.g. SAGAT, SPAM) while others recommend strategic selection to account for scenario events (e.g. SASHA).

The *data collection phase* involves administering the queries/probes. For this phase, the biggest difference between measures is whether the questions are administered in real time (i.e. probes) or during a pause (i.e. queries) of a scenario trial. The sensitivity and validity of both methods are a matter of ongoing debate (cf., Durso, Bleckley, and Dattel 2006; Jones and Endsley 2004).

The *data analysis phase* involves processing experimental data to arrive at a final score. Once again, query/probe-based SA measures differ in how this is done. Most measures recommend using the simulator log data to form the basis of correct answers, though expert judgement is permitted (e.g. SAGAT). However, some measures rely solely on expert judgements (e.g. SASHA). The calculation for the final scores is typically percentage correct, although a few measures employ signal detection theory (e.g. SACRI, QUASA). The scoring methods can constrain the response alternatives at the data collection phase. For instance, SACRI and QUASA have a relatively standardised format for their response alternatives so that signal detection theory can be employed.

To date, the development of query/probe-based SA measures have emphasised improving sensitivity and validity or satisfying unique requirements of a domain. To our knowledge, there is no literature specifically examining inter-rater reliability of any query/probe-based SA measure in detail, despite the importance indicated by psychometric research. This may be because early psychometric research on SA measures mainly involved controlled experiments with limited system or scenario complexity, thereby alleviating the need for expert judgements. However, inter-rater reliability becomes relevant when experts are recruited to develop and score query/probe-based SA measures in representative settings.

In these settings, process experts are often needed to interpret system states and operator actions relative to queries/probes. As complexity increases the interpretation of these indicators, actions and effects becomes challenging. Multiple indicators must be interpreted relative to operating contexts and several sets of control actions can be equally satisfactory for a given operating situation. For this reason, experts often provide critical support in the preparation and data collection phases of representative experiments. For query/probe-based SA measures, researchers and practitioners typically rely on experts (i) to formulate a set of relevant queries or probes to assess operator knowledge about the situation and (ii) to determine the correct responses for the event-dependent operating conditions created by operator responses to scenario events in individual experimental trials. This reliance is exemplified by the SAVANT (Willems and Heiney 2002) and SACRI (Hogg et al. 1995) methods, developed for use with full-scope simulators in air traffic control and nuclear process control, respectively.

The only empirical treatments of inter-rater reliability of query/probe-based measures are the secondary results of (i) a nuclear simulator experiment employing SACRI (Hogg et al. 1995), (ii) an air-traffic control experiment employing SALSA (Hauss and Eyferth 2003) and (iii) a flight simulator experiment employing an adapted version of SAGAT (Prince et al. 2007). In the first study, SACRI demonstrated an inter-rater reliability of 0.93 in comparing 468 query responses about directional changes of process parameters (e.g. increasing or decreasing) between two raters. There are two limitations to this result. First, Hogg et al. did not specify the inter-rater reliability statistic. The reliability could be inflated if the statistic (e.g. Pearson moment product correlation) does not account for chance agreements (see e.g. Cohen 1960). Second, the responses do not appear to depend on cognitive judgements or information processing necessary to obtain the SA that the queries were intended to elicit. Specifically, the raters provided response keys after the experimental trials based on criteria that were pre-defined for viewing graphs rather than impacts on the nuclear process, potentially translating a demanding information-processing task to a perceptual judgement task.

In the second study, Hauss and Eyferth (2003) examined the degree of agreement between experts in identifying relevance of aircraft parameters for the SALSA measure and reported an inter-rater reliability statistic of 0.73 based on 15 air traffic situations. The authors omitted detailed descriptions of the experts and statistics, both of which are needed to interpret this result.

In the third study, Prince et al. (2007) examined the degree of agreement between two raters scoring 41 crews of participants on 48 queries in a flight simulator experiment using a heavily adapted SAGAT measurement method.

The correct responses in this study were provided by a third expert and applied by the two raters, who assigned a score between one and five to individual participant answers. The adopted measurement method was a mixture of rating scale- and query/probe-based techniques. Prince et al. reported a Pearson product-moment correlation of 0.88. The Pearson product-moment correlation is not an ideal inter-rater reliability statistic because it does not account for chance agreement (e.g. Shrout and Lane 2012) between raters. The authors also omitted detailed descriptions of the experts and rating environment, both of which are needed to interpret this result.

To our knowledge, these three studies provide the only empirical evidence on inter-rater reliability of query/probe-based SA measures.[3] Although the statistics reported appear to be good on the surface, these studies present significant limitations to drawing generalisable conclusions about inter-rater reliability. This represents a severe gap in the literature. Additional empirical evidence is necessary to provide clearer references for estimating inter-rater reliability of query/probe-based SA measures and to inform methods for training raters, refining measures and developing scenarios/targets.

Overview

To address the paucity of inter-rater reliability assessment, we recruited process experts to provide reference keys based on data collected from a full-scope nuclear power plant simulator experiment at the OECD Halden Reactor Project (HRP). In the experiment, licensed operators completed scenarios designed to study the effects of new control room technology on human performance. Process experts provided reference keys to the SA queries based on their observation of operator control actions and process displays during each scenario trial. The full scope-simulator experiment provided a representative nuclear process control environment that required operators to draw on their domain expertise to monitor the process and to answer the SA queries.

Method

We describe the study in five parts. The first part describes the raters – the process experts – whose qualifications are particularly relevant to interpreting information about the targets of the rating task. The second part describes the rating instrument – the SA measure – which exemplifies key features of query/probe-based measures. The third part describes the rating targets – the parameter behaviours in the experimental trials – the interpretation of which impose substantial cognitive demands given the complexity of the representative environment. The fourth and fifth parts describe the environment and procedure for the experts performing the rating task, respectively.

Raters

Three process experts participated in this study. (To avoid confusion between the two sets of participants, *process expert* refers to the rater in the inter-rater reliability study, and *operator* refers to the participants in the full-scope simulator study.) The process experts provided written consent to allow their ratings and individual descriptions to be included in reports of the study, despite the high likelihood of being identified by others in their organisation from those descriptions. Given the small number of raters, their characteristics were important for the interpretation of results.

Process expert A had supported HRP research activities for two years leading up to the experiment without any experience of providing reference keys to any query/probed-based measures previously. He worked as a control room operator for over a decade and helped develop the test simulator. He designed the scenarios for the experiment and identified the process parameters for the SA queries (see 'Rating targets – full-scope simulator experimental trials' section). He also advised HRP scientists on the simulated nuclear processes and operations.

Process expert B was a recently retired shift supervisor at the plant that the test simulator replicates. He had worked as a shift supervisor for over three decades and participated in about ten HRP simulator studies prior to this experiment. He was an operator participant in the pilot trials of the experiment.

Process expert C had supported HRP research activities for over a decade leading up to the experiment with prior experience of providing the reference keys to the queries in earlier experiments. He had previously worked as a control room operator for over a decade and was a major contributor to the development of the simulator. He co-developed several experimental manipulations, programmed the scenarios and advised other HRP scientists on the simulator nuclear processes and operations.

Based on their experience, all three process experts were deemed qualified to support experiment preparation and collect human performance measurements. The process experts had different levels of experience with the simulator and with the experiment; both of which may have influenced their answers to the SA queries because knowledge of the simulator and experiment can improve judgement on the effects of scenario events and operator actions on the nuclear process and thus enhance the inter-rater reliability.

Rating instrument – SA measure

The query/probe-based SA measure was an adaptation of SACRI, operationalising SA as *accurate detection of meaningful changes in relevant parameters of process plants*. During the preparation phase, a process expert determined relevant process parameters (i.e. sensor readings in the plant) that represented the operating context and process events in the scenario. Parameters were deemed *relevant* if the process expert concluded that the behaviours of the parameters must be known to the operators to complete the scenarios successfully. The awareness of these parameter behaviours in the recent past was elicited through administration of queries of the form specified in Figure 1. An example of such queries would be: 'Recently, the total feedwater flow 312KA031 has?' The response alternatives were – (i) increased, (ii) stayed the same and (iii) decreased – forming a theoretically comprehensive set. The operators (including the experts) were instructed to disregard random fluctuations or noise in the nuclear process, and to select the *most recent behaviour* of the critical parameters in the queries for each scenario period. The operators had full discretion to define the time window that constitutes 'recent' because this is considered part of monitoring industrial plants. This measure standardises the queries except for selection of parameters (which were scenario-dependent) and fully standardises the response alternatives.

During data collection, a process expert froze the simulator while the operators answered the queries without any access to process displays. The operators' answers are labelled *responses*. The process expert simultaneously answered the queries with access to the process displays and a-priori knowledge of the process faults. The process expert's answers are labelled *reference keys*. Final scores were calculated as the proportion correct (or matches) between the responses and reference keys.

This measure is sufficiently illustrative of query/probe-based SA measures (also see the 'Query/probe-based SA measures' section). Similar to other query/probe-based measures, this measure elicits the operator's declarative knowledge by directly questioning participants during pauses in the scenarios. The queries are highly standardised to elicit operator awareness of specific parameter behaviours, differing only in terms of parameter selection. Further, the response alternatives are fully standardised and fixed for all queries. The form of the queries and responses in this measure does not violate any constraints prescribed by any query/probe-based SA measure. Similarly, the query administration process is similar to many existing measures. The reliance on process experts for correct answers is an accepted practice (Endsley 2000). In essence, it is conceivable that other query/probe-based measures could administer similar questions for assessing SA. The empirical results here would therefore serve as a reasonable reference for estimating inter-rater reliability for query/probe-based SA measures in representative settings.

Rating targets – full-scope simulator experimental trials

The SA measure was employed in a full-scope nuclear power plant simulator experiment, recruiting nine crews of licensed operators. The targets of the rating task for the process experts were the parameter behaviours in relation to the scenario events and operator control actions.

The experiment employed the HAlden Man-machine laboratory BOiling water reactor (HAMBO; Karlsson et al. 2001; Øwre et al. 2002), simulating a 1200 MW facility. The crews, each composed of three operators with different roles, were to operate the plant simulator for eight scenarios as if they were on duty. Each scenario introduced a unique set of process faults (e.g. a stuck valve) at different times to challenge the crews in operating the simulator. Consequently, the parameter behaviours varied according to both scenario events and operator control actions. Every scenario trial was divided into two periods by a freeze for administering queries (Figure 2).

Rating environment

The inter-rater reliability study was conducted in three parts. Part I collected data from two process experts who provided reference keys to SA queries from the observation gallery during the scenario trials of two crews. Part II collected data from

Query Structure:

Recently, the [parameter, code] has:

Response Alternatives:

Increased/Stayed the same/Decreased

Figure 1. Query and response format of the query-based SA measure.

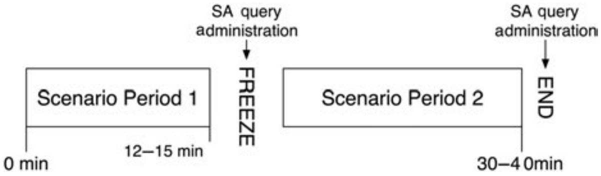

Figure 2. Overview of the scenario structure.

one additional process expert who provided reference keys to the same queries after the experiment based on graphs of the parameters generated from the simulator log for the trials of the same two crews. Due to differences in rating environment (and procedure) between Parts I and II, Part III collected data from the same three process experts following the same protocol as in Part II to verify the results from Parts I and II.

Part I – real-time data collection

Part I was conducted in the observation gallery (Figure 3) in parallel with the data collection. Process experts A and C operated the simulator and implemented the scenarios. Process expert A acted as field operator, electrician, plant manager and other roles as necessary by responding to phone calls from the simulator control room. Process expert A also provided the reference answers to the SA queries and rated operators on other performance scales (also see, Lau, Jamieson, and Skraaning 2012). Process expert C was the designated technical support person for simulator operations but did not provide reference keys during Part I due to his high workload.

Process expert B provided reference keys to the SA queries for the inter-rater reliability study. Process expert B occupied a workstation (Figure 4) with (i) two displays for process information, (ii) earbuds for audio feed from the control room, (iii) a laptop for inputting scores/ratings, (iv) a pair of binoculars for observing operators, (v) descriptions of the scenarios, (vi) pen and paper for note taking and (vii) a pair of ear muffs for muting any discussion on operator performance in the observation gallery. An opaque, black curtain was installed between process experts A and B.

Part II – post-experiment data collection

Part II was conducted after the data collection for the simulator experiment in a data analysis laboratory, which had a large workstation equipped with a computer and two large LCD monitors to collect data. A human factors researcher was also present in case the process expert C had questions while providing reference keys to the SA queries. The researcher was seated in another desk at a distance, facing away from process expert C.

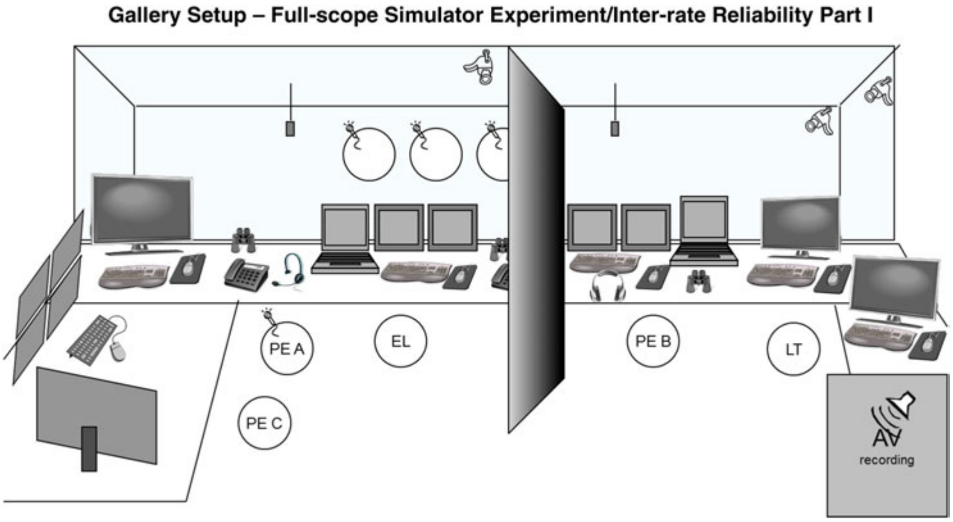
Gallery Setup – Full-scope Simulator Experiment/Inter-rate Reliability Part I
PE A
EL
PE C
PE B
LT
AV
recording

Figure 3. Observation gallery of the full-scope simulator experiment/Part I of the inter-rater reliability study. PE A, B and C, process expert A, B and C (PE C did not answer SA queries in Part I of the study); EL, experimental leader; LT, laboratory technician.

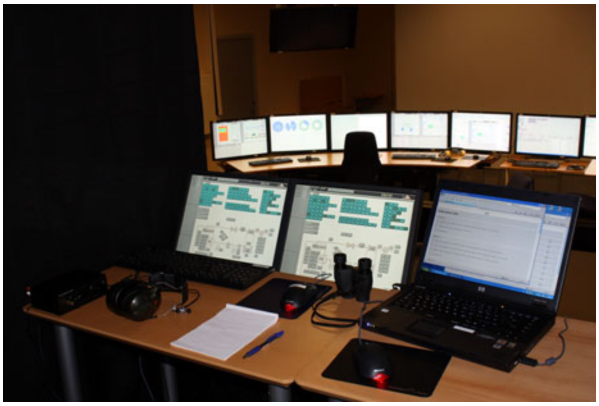

Figure 4. Workstation for PE B in Part I.

Part III – post-experiment result verification

Part III was conducted under the same conditions as in Part II. Each process expert provided a second set of reference keys individually.

Procedure

Part I – real-time data collection

The process experts were instructed to refrain from discussing the SA measure or operator performance. Process expert A had received instruction on the SA measure in the preparation phase of the full-scope simulator experiment and was verbally introduced to the inter-rater reliability concept and study.

Process expert B (in Part I) answered the SA queries for sixteen trials of two crews in the simulator experiment. His instructions were given in two sessions. The first session consisted of a PowerPoint presentation providing an overview of the inter-rater reliability concept and study. The second session began by reviewing the procedure for answering the SA queries with explanations and examples of the SA queries. The second session ended by familiarising process expert B with the apparatus and data collection tools in the observation gallery (Figures 3 and 4).

Process expert B was seated at his workstation and answered SA queries by entering his response with mouse and keyboard into the computerised questionnaire system, simultaneously and in the same manner as process expert A. Process expert B was new to the role of performance rater in an experimental setting; therefore, we decided a priori to treat the first two trials as practice sessions, with scores omitted from the data analysis. In total, process experts A and B answered 350 SA queries, 44 of which were considered practice.

Part II – post-experiment data collection

Part II was conducted four weeks after the simulator experiment. Process expert C was given a brief introduction about the inter-rater reliability study and was verbally instructed to review trend graphs of process parameters and answer the SA queries in the order of crew, scenario and period. Process expert C was familiar with the SA queries because he provided reference keys to similar SA queries in prior HRP experiments.

To support his ratings with contextual information, process expert C was provided with the scenario descriptions and task performance items (Skraaning Jr et al. 2010). Videos of the experimental trials were also available but process expert C did not choose to view any. In total, process expert C answered 328 SA queries in two sessions on the same day, with a lunch break in between. Excluding practice queries, Parts I and II shared 290 SA queries, which formed the final data-set.

The graphs were generated by a java applet, compiled into PowerPoint files and presented on one LCD monitor in the order of crew, scenario and period. The computerised questionnaire system presented the queries and collected the responses on a separate LCD monitor.

Part III – post-experiment result verification

Part III was conducted ten months after the simulator experiment. Process experts A, B and C followed the same procedure as in Part II to provide a second set of reference keys for the 225 queries pertaining to one crew in the simulator experiment.

Table 1. Descriptive statistics between process experts for Parts I and II.

	B (Part I)		C (Part II)	
Process expert	Agreement proportion	Marginal proportion	Agreement proportion	Marginal proportion
A (Part I)	0.676	0.385	0.783	0.428
B (Part I)	–		0.614	0.414

Notes: Agreement proportion refers to the proportion of reference keys matched between two process experts. Marginal proportion refers to the proportion expected from chance association given the data collected.

Analysis and results

Cohen's Kappa, κ (Cohen 1960) was applied to assess the agreement between process experts. κ is a correlation statistic that adjusts for chance association in a nominal scale and treats raters as exchangeable for generalisation to the rater population (see, Fleiss and Cohen 1973; Shrout and Fleiss 1979). Table 1 presents the proportion of agreement and chance agreement between the process experts that were used to calculate the κ statistics for data collected from Parts I and II. Table 2 presents the proportion and chance agreement for data collected from Part III.

Tables 3 and 4 present the κs between the three process experts based on their reference keys from Parts I and II containing 290 queries, and from Part III containing 225 SA queries, respectively. Figure 5 graphs the inter-rater reliability κ results. For the 225 queries overlapping across the three parts of the study, *intra*-rater reliability was computed for process experts A ($\kappa = 0.543$, $\sigma = 0.049$), B ($\kappa = 0.572$, $\sigma = 0.048$), and C ($\kappa = 0.604$, $\sigma = 0.050$).

The three κs calculated using data from Parts I and II formed a range with lower and upper bounds at 'fair' and 'substantial' agreement, respectively. (For a common interpretation of κ values, see Landis and Koch 1977; Sims and Wright 2005; Shrout 1998). For experiments representative of the complexity of industrial settings, 'ideal' agreement between raters is unlikely and 'substantial' agreement would typically be considered acceptable. However, the observed κ values did not positively indicate 'substantial' agreement; thus, the results suggested that this SA measure could not assume inter-rater reliability. Given the difference between rating environment and procedure between Parts I and II, these reliability results needed verification. The κs calculated with data collected in Part III formed a range with lower and upper bounds at 'fair' and 'moderate' agreement, respectively, thereby verifying the inter-rater reliability results from Parts I and II.

The differences between the κs were unexpectedly large across rater pairs, prompting an examination of the effect of individual process expert on inter-rater reliability. The statistical significance of the process expert effect on the inter-rater reliability coefficients was tested with multiple comparisons following Holm's procedure for controlling type I error rate (Howell 2002). Table 5 indicates significant reliability coefficients for all κ pairs: $\kappa_{A,C} > \kappa_{A,B}$ ($Z = 2.39$, $p = 0.02$); $\kappa_{A,C} > \kappa_{B,C}$ ($Z = 4.32$, $p < 0.01$); and $\kappa_{A,B} > \kappa_{B,C}$ ($Z = 1.99$, $p = 0.05$) suggesting a significant effect of expert-pairs.

The reliability coefficients should therefore be interpreted with respect to the process expert characteristics (see 'Raters' section) and the experimental procedure (see 'Rating environment' and 'Procedure' sections).

Discussion

Inter-rater reliability of query/probe-based SA measures

This study is the first detailed empirical investigation of the inter-rater reliability of a query/probe-based SA measure. In keeping with the triadic nature of inter-rater reliability, we discuss the results in terms of targets, measures and raters.

The targets in this study were representative of the complexity associated with monitoring process plants (see e.g. Mumaw et al. 2000). This complexity may have led to variability in the interpretation of process indicators and constrained inter-rater reliability of the measure. Operator detection of parameter changes is likely a result of not only sensory registration of signals but also of information processing in relation to the operating context. The importance of operating

Table 2. Descriptive statistics between process experts for Part III.

	B (Part III)		C (Part III)	
Process expert	Agreement proportion	Marginal proportion	Agreement proportion	Marginal proportion
A (Part III)	0.689	0.381	0.680	0.399
B (Part III)	–		0.600	0.422

Table 3. κ statistics between process experts for Parts I and II.

	B (Part I)		C (Part II)	
Process expert	κ	σ	κ	σ
A (Part I)	0.473	0.045	0.620	0.042
B (Part I)	–		0.341	0.049

Table 4. κ statistics between process experts for Part III.

	B (Part III)		C (Part III)	
Process expert	κ	σ	κ	σ
A (Part III)	0.497	0.050	0.467	0.052
B (Part III)	–		0.308	0.056

context for the simulator was evidenced by process experts frequently referencing the scenario descriptions when providing reference keys to the queries in Parts I and II. Given such complexity, experts do not always agree.

As complexity is inherent to industrial systems (i.e. the targets), research attention must turn to improving the correspondence of the queries and response sets (i.e. the measure) to situations to minimise variability from rater misinterpretation. This SA measure with a relatively standardised set of queries and responses only produced κ results of ‘moderate’ reliability. Psychometric research from other domains suggests that inter-rater reliability often improves with standardisation of questions and responses (e.g. Conway, Jako, and Goodman 1995) and sometimes varies with different performance dimensions (e.g. Haddock et al. 1999; Jonsson and Svingby 2007; Viswesvaran, Ones, and Schmidt 1996). Guidance on formulating queries to elicit SA should therefore account for system and operational characteristics to maximise compatibility with information processing of the operator in coping with the situation. Further, the query/probe-based technique poses a unique challenge to studying psychometrics because queries or probes typically vary from one scenario trial to another. This stands in contrast to static questionnaires or rating scales. That is, all query/probe-based SA measures, which adopt the combination of the SA notion and the query/probe-based technique, intrinsically lead to questions customised for each scenario (i.e. the situation). Customising questions to scenarios is important for sensitivity and construct validity; however, the variation in questions and responses would require multiple studies or meta-analyses to assess any single psychometric dimension. Given the substantial variation in query and response formats of SA

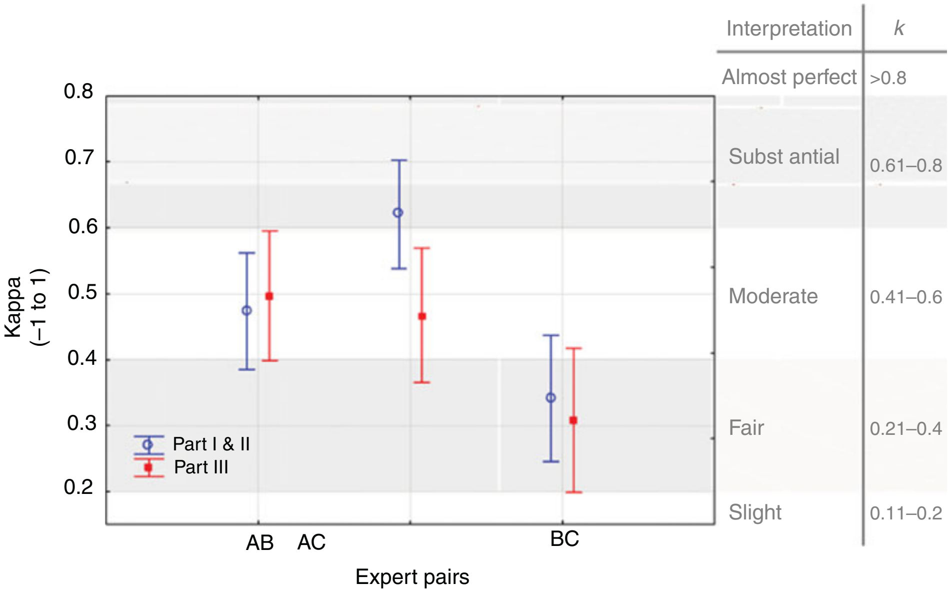

Figure 5. Mean and 95% confidence intervals of κs for Parts I and II, and III of the study.

Table 5. Multiple comparisons of κs following the Holm's multistage procedure.

	AC		BC	
Process expert pair	*Z*	*p*	*Z*	*p*
AB	2.39	0.017*	1.99	0.047*
AC	–		4.32	0.000*

*Significant effect following Holm's procedure to control for type I error-rate.

measurements across studies and limited psychometric data on those measurements, practitioners cannot assume inter-rater reliability of query/probe-based SA measures, especially in complex or representative settings. This highlights the need for further *guidance and evaluation* on how to formulate and standardise queries/probes and responses that consistently map specific types of information (e.g. parameter changes, diagnosis of process faults, fault severity). In essence, the results of this study suggest that inter-rater reliability should be an explicit consideration in the development and evaluation of any query/probe-based SA measure that relies on experts.

The raters in this empirical study were three process experts with extensive experience in the nuclear power generation domain. The results illustrated the influence that process expert characteristics and data collection procedures had on inter-rater reliability (see Table 6). Process experts A and C demonstrated the greatest agreement; process experts B and C demonstrated the least agreement; process experts A and B demonstrated an intermediate level of agreement (Figure 5). The most likely explanation can be found in the differences in the degree of involvement of the process experts in the preparation of the experiment and data collection procedures (i.e. Part I vs. Part II). Process experts A and C, both HRP resident experts, were heavily involved in preparing the full-scope simulator experiment and, hence, possessed deep knowledge of the situations. In contrast, process expert B was only exposed to the full-scope simulator experiment as an operator (participant) in the pilot trials. Thus, even though process expert A answered the queries in real time (Part I) while process expert C answered them after the experiment (Part II), they demonstrated the greatest agreement. Process experts A and B, on the other hand, shared the same data collection procedure, resulting in the second highest agreement. Process experts B and C demonstrated the lowest level of agreement as they shared neither preparation of the full-scope simulator experiment nor data collection procedure. Given the circumstances, the variation of inter-rater reliability between experts is not surprising. The clearest implication of these results may be that knowledge of the experiment and scenarios is probably more important than the overall data collection procedure or environment. These results also raise the issue of rater selection and training for representative experiments, an issue that has not been clearly addressed by inter-rater reliability research in general. That research tends to focus on general knowledge, selecting raters according to experience (e.g. years or cases practiced for a particular disorder) or formal credentials (e.g. residents or medical doctors) and training raters about the measures to align their interpretation of the scales. This particular study suggests that researchers should attend to the raters' detailed understanding of the scenarios challenging the operators (i.e. the targets). This resonates with the finding from an observational SA study by Patrick et al. (2006), which was also conducted in a full-scope nuclear power plant simulator. They concluded that rater effectiveness depends heavily on the level of effort invested in designing scenarios, exploring multiple resolutions to process faults, and familiarising raters with the various relevant secondary operator tasks. Further research is necessary to understand rater training and selection for query/probe-based SA measures in representative settings.

Finally, the paradoxical as well as the triadic nature of inter-rater reliability might continually challenge the development of SA measures relying on expert judgement. Inter-rater reliability is likely to be *highest but least relevant* on well-defined, easily agreed-upon, topics and in controlled settings where optimum solutions are computable. Inter-rater reliability is likely to be *lowest but most relevant* on ill-defined topics requiring expert judgements, and in complex settings

Table 6. Defining characteristics of process experts.

Expert A	Expert B	Expert C
15+ years operating experience	15+ years operating experience	15+ years operating experience
2 years as process expert in research	N/A, participated in 10 experiments	10+ years as process expert in research
Developed simulator	N/A	Developed simulator
Developed scenarios for experiment	Participated in pilot experiment	Developed scenarios for experiment
Developed measures	Reviewed descriptions on the experiment and measures	Developed measures

where trade-offs and contexts are critical. In other words, inter-rater reliability is expected to be lower, yet more important, in representative than in controlled settings. Reliance on experts tends to increase with experimental representativeness, thereby elevating the complexity of the queries and risk of disagreement between experts. Kraemer (1992) noted a similar paradox for medical treatment decisions in practice from the point of view of employing consensus of multiple raters, stating that extremely high inter-rater reliability (i.e. $\kappa > 0.8$) suggests single rating to be sufficient whereas moderate inter-rater reliability (i.e. $0.2 < \kappa < 0.8$) could benefit from consensus of multiple raters. Basically, treatment decisions for well-understood medical cases are likely to have high inter-rater reliability requiring minimal attention, whereas treatment decisions for complex medical cases are likely to have moderate inter-rater reliability, requiring multi-rater assessments.

This paradox emphasises the importance of examining inter-rater reliability of query/probe-based SA measures for representative experiments. Given that SA aims to reflect a cognitive human performance dimension in complex systems, some aspects of SA measurements are likely to be complex and ill defined. This leads to reliance on experts and raises concerns for inter-rater reliability. The lack of research attention given to the inter-rater reliability of query/probe-based measures of SA stands in stark contrast to psychometric research in other settings where researchers periodically discuss means to improve upon as well as to evaluate measurements from an inter-rater reliability perspective. Empirical research should experimentally investigate different principles or methods to improve inter-rater reliability of query/probe-based SA measures in addition to providing psychometric indices of existing measures. The empirical results of this study can hopefully prompt further psychometric research for advancing SA measurement techniques and measures.

Limitations

This inter-rater reliability study has limitations with respect to its triadic nature – rater, method and target. The limitation related to rater is the recruitment of only three process experts. This limitation is moderated statistically by using Cohen's κ, which treats raters as exchangeable, permitting generalisation to rater population. Publications on inter-rater reliability based on two to three raters are not uncommon (e.g. Rosenzweig et al. 1999; Ramos, Schafer, and Tracz 2003; Devitt et al. 1997; Saxton, Belanger, and Becker 2012; Patrick et al. 2006; Prince et al. 2007). Nevertheless, increasing the number of raters would incrementally improve the representativeness of, and our confidence in, the inter-rater reliability statistics. Representative studies are frequently confronted with the trade off between limited access to and validity gained with domain expertise in professional operators.

The limitation related to method is the confounding assessment of rater judgements between interpreting 'recently' (i.e. the time window for judgement) and parameter behaviours. That is, the source of reliability (or lack thereof) may stem from the challenge in determining the time window that was relevant for the queries given the scenario, behaviours that the parameter were exhibiting during that time window, or both. Additional empirical experiments are necessary to identify the attributes of this measure impacting the inter-rater reliability.

The limitation related to target is the variation in the rating procedure between Part I (real-time ratings by experts A and B) and Part II (post experiment ratings by expert C). In contrast, all raters in typical inter-rater reliability studies are given exactly the same tasks and situated in the same environment. Nevertheless, the conclusions from Part I and II are robust given similar results from Part III, in which all three experts followed the same protocol.

Though ideal, controlled conditions with many raters are unrealistic in representative experiments involving process experts whose availability is scarce. It is often impractical to recruit multiple process experts to perform ratings and conduct the identical tasks during data collection for the sake of providing reliability data. Process experts may be responsible for different preparation tasks from one experiment to another. To be practically useful, SA measures should be sufficiently robust to these uncontrolled properties commonly found in representative settings.

Future work

Future work is necessary to investigate how specific methodological details of query/probe-based SA measures impact their psychometric properties (see Shrout and Lane 2012). For instance, the forms of the query and response can be experimentally manipulated to seek the best match for eliciting specific types of situation knowledge.

Advances in improving and collecting inter-rater reliability measurements are necessary to overcome the sparse access to experts/operators and paucity of empirical data. Besides collecting ratings from a few experts during scenario trials, representative experiments should include other means to improve and collect reliability data as part of the method. For example, upon completion of all scenario trials, participant crews can provide ratings based on video recordings and simulator logs of other participant crews, thereby increasing available data. In addition, calibration techniques from psychophysics may be adapted and integrated into some query/probe-based measures.

Research must provide additional empirical data on other forms of reliability – internal consistency, test–retest and parallel form reliability – that appear largely as secondary results in the literature. Advances in psychometric methodology might be necessary to examine internal consistency reliability, of which the research focuses on a static set of questions rather than queries/probes that vary from one trial to another.

Conclusion

This study is the first explicit examination of inter-rater reliability of query/probe-based SA measures. The results for the SA measure evaluated in this representative study alert researchers to the inter-rater reliability issue of query/probe-based techniques for measuring SA. This particular query/probe-based SA measure demonstrated 'moderate' inter-rater reliability even though the queries were seemingly direct and simple. The lack of agreement on the 'correct' answers to queries and probes means that process experts may introduce unwanted variability in the data, thereby leading to inconsistency of SA measurements and experimental results. Thus, analysis and interpretation of data collected from query/probe-based SA measures must attend to inter-rater reliability. However, the literature neglects, at least by omission, inter-rater reliability of query/probe-based SA measures. There is virtually no discussion on constructing queries and probes about operations of complex systems that ensure consistent interpretation across professionals. Consequently, the interpretation of experimental results may be biased. This empirical study highlights the paucity of empirical assessment on the inter-rater reliability of query/probe-based measures that should be addressed in future studies to build a strong empirical foundation for SA research.

Acknowledgements

Thanks to Professor Neville Stanton, Justin Hollands, Mark Chignell and Birsen Donmez for providing valuable feedback on this work. We thank Andreas Bye, of the Halden Reactor Project, for his effort in making this study possible.

Funding

This research was supported through a grant from the Natural Science and Engineering Research Council of Canada and internal funding of the OECD Halden Reactor Project. We are indebted to Maren H. Rø-Eitrheim for supporting the recruitment of the process expert and gallery set-up for data collection.

Notes

1. Indices based on judges assigning similar rank ordering for the targets are referred to as inter-rater reliability, whereas indices based on judges assigning identical rating level of the targets are referred to as inter-rater agreement. The reliability statistics are typically used for measurements in interval and ratio scales, whereas agreement statistics are typically used for measurements in nominal scales. Common usage often does not distinguish between inter-rater reliability and agreement, but it is important to identify the type of measurement scales to select the appropriate statistics, which assume certain scale properties.
2. Translated 'Measuring Situation Awareness of Area Controllers within the Context of Automation' from German.
3. Empirical evidence is also limited for SA measures adopting techniques other than the query/probe-based technique. The authors are aware of only three other empirical studies that include inter-rater reliability indices – each for a different rating scale (Patrick et al. 2006; Waag and Houck 1994; Vidulich and Hughes 1991).

References

Bernardin, H. J., and M. R. Buckley. 1981. "Strategies in Rater Training." *The Academy of Management Review* 6: 205–212.

Boring, R. L. 2003. "Improving Human Scaling Reliability." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 47: 1820–1824.

Boring, R. L., and R. L. West. 2008. "Constrained scaling in psychometric magnitude mapping." *Proceedings of the 24th Annual Meeting of the International Society for Psychophysics* 24: 297–302.

Cohen, J. 1960. "The coefficient of agreement for nominal scales." *Educational and Psychological Measurement* 20: 37–46.

Conway, J. M., and A. I. Huffcutt. 1997. "Psychometric Properties of Multisource Performance Ratings: A Meta-Analysis of Subordinate, Supervisor, Peer, and Self-Ratings." *Human Performance* 10: 331–360.

Conway, J. M., R. A. Jako, and D. F. Goodman. 1995. "A Meta-Analysis of Interrater and Internal Consistency Reliability of Selection Interviews." *Journal of Applied Psychology* 80: 565–579.

Devitt, J., M. Kurrek, M. Cohen, K. Fish, P. Fish, P. Murphy, and J.-P. Szalai. 1997. "Testing the Raters: Inter-Rater Reliability of Standardized Anaesthesia Simulator Performance." *Canadian Journal of Anesthesia [Journal canadien d'anesthésie]* 44: 924–928.

Drøivoldsmo, A., G. Skraaning Jr., M. Sverrbo, J. Dalen, T. Grimstad, and G. Andresen. 1998. *Continuous Measures of Situation Awareness and Workload*. Halden: OECD Halden Reactor Project.

Durso, F. T., M. K. Bleckley, and A. R. Dattel. 2006. "Does Situation Awareness Add to the Validity of Cognitive Tests?" *Human Factors* 48: 721–733.

Durso, F. T., T. R. Truitt, C. A. Hackworth, J. M. Crutchfield, and C. A. Manning. 1998. “En Route Operational Errors and Situational Awareness.” *The International Journal of Aviation Psychology* 8: 177–194.

Eckes, T. 2008. “Rater Types in Writing Performance Assessments: A Classification Approach to Rater Variability.” *Language Testing* 25: 155–185.

Endsley, M. R. 1988. “Situation Awareness Global Assessment Technique (SAGAT).” In *National Aerospace, Electronics Conference (NAECON)*, 789–795. New York, NY: IEEE.

Endsley, M. R. 1995. “Measurement of Situation Awareness in Dynamic Systems.” *Human Factors* 37: 65–84.

Endsley, M. R. 2000. “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, 147–174. Mahwah, NJ: Lawrence Erlbaum Associates.

Fleiss, J. L., and J. Cohen. 1973. “The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability.” *Educational and Psychological Measurement* 33: 613–619.

Fracker, M. L. 1991. *Measures of Situation Awareness Review and Future Directions*. Dayton, OH: Wright-Patterson Air Force Base.

Gatsoulis, Y., V. S. Gurvinder, and A. 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: 69–98.

Haddock, G., J. Mccarron, N. Tarrier, and E. B. Faragher. 1999. “Scales to Measure Dimensions of Hallucinations and Delusions: The Psychotic Symptom Rating Scales (PSYRATS).” *Psychological Medicine* 29: 879–889.

Hauss, Y., and K. Eyferth. 2003. “Securing Future ATM-Concepts’ Safety by Measuring Situation Awareness in ATC.” *Aerospace Science and Technology* 7: 417–427.

Hogg, D. N., K. Follesø, F. S. Volden, and B. Torralba. 1995. “Development of a Situation Awareness Measure to Evaluate Advanced Alarm Systems in Nuclear Power Plant Control Rooms.” *Ergonomics* 38: 2394–2413.

Howell, D. 2002. *Statistical Methods for Psychology*. Pacific Grove, CA: Duxbury/Thomson Learning.

Jeannot, E. 2000. *Situation Awareness, Synthesis of Literature Research*. Brussels: Eurocontrol.

Jeannot, E., C. Kelly, and D. Thompson. 2003. *The Development of Situation Awareness Measures in ATM Systems*. Brussels: Eurocontrol.

Jones, D. G., and M. R. Endsley. 2004. “Use of Real-Time Probes for Measuring Situation Awareness.” *The International Journal of Aviation Psychology* 14: 343–367.

Jonsson, A., and G. Svingby. 2007. “The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences.” *Educational Research Review* 2: 130–144.

Karlsson, T., H. Jokstad, B. D. Meyer, C. Nilhlwing, S. Norrman, E. K. Puska, P. Raussi, and O. Tiihonen. 2001. *OECD Halden Reactor Project: The HAMBO BWR Simulator of HAMMLAB*. Halden Institutt for Energiteknikk.

Kraemer, H. C. 1992. “How Many Raters? Toward the Most Reliable Diagnostic Consensus.” *Statistics in Medicine* 11: 317–331.

Landis, J. R., and G. G. Koch. 1977. “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33: 159–174.

Lau, N., G. A. Jamieson, and G. Skraaning Jr. 2012. “Inter-Rater Reliability of Expert-Based Performance Measures.” In *Proceedings of the 8th American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC & HMIT)*, 1974–1982. San Diego, CA: American Nuclear Society.

Lievens, F. 2001. “Assessor Training Strategies and Their Effects on Accuracy, Interrater Reliability, and Discriminant Validity.” *Journal of Applied Psychology* 86: 255–264.

Maxwell, S. E., and H. D. Delaney. 1990. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mcguinness, B. 2004. “Quantitative Analysis of Situational Awareness (QUASA): Applying Signal Detection Theory to True/False Probes and Self-Ratings.” Proceedings of the Ninth International Command and Control Research and Technology Symposium, San Diego, CA.

Miller, T. J., T. H. Mcglashan, J. L. Rosen, K. Cadenhead, J. Ventura, W. Mcfarlane, D. O. Perkins, G. D. Pearlson, and S. W. Woods. 2003. “Prodromal Assessment With the Structured Interview for Prodromal Syndromes and the Scale of Prodromal Symptoms: Predictive Validity, Interrater Reliability, and Training to Reliability.” *Schizophrenia Bulletin* 29: 703–715.

Mitchell, S. K. 1979. “Interobserver Agreement, Reliability, and Generalizability of Data Collected in Observational Studies.” *Psychological Bulletin* 86: 376–390.

Mumaw, R. J., E. M. Roth, K. J. Vicente, and C. M. Burns. 2000. “There is More to Monitoring a Nuclear Power Plant than Meets the Eye.” *Human Factors* 42: 36–55.

Murphy, K. R., and C. O. Davidshofer. 1998. *Psychological Testing: Principles and Applications*. Upper Saddle River, NJ: Prentice Hall.

Neal, A., M. A. Griffin, J. Paterson, and P. Bordia. 1998. “Development of Measures of Situation Awareness, Task Performance, and Contextual Performance in Air Traffic Control.” Fourth Australian Aviation Psychology Symposium, Sydney, Australia.

Øwre, F., J. Kvalem, T. Karlsson, and C. Nihlwing. 2002. “A New Integrated BWR Supervision and Control System.” In *Proceedings of IEEE 7th Conference on Human Factors and Power Plants*, 4-41–4-47, Scottsdale, AZ.

Patrick, J., N. James, A. Ahmed, and P. Halliday. 2006. “Observational Assessment of Situation Awareness, Team Differences and Training Implications.” *Ergonomics* 49: 393–417.

Pew, R. W. 2000. “The State of Situation Awareness Measurement: Heading Toward the Next Century.” In *Situation Awareness Analysis and Measurement*, edited by M. R. Endsley and D. J. Garland, 33–37. Mahwah, NJ: Lawrence Erlbaum Associates.

Prince, C., E. Ellis, M. T. Brannick, and E. Salas. 2007. “Measurement of Team Situation Awareness in Low Experience Level Aviators.” *The International Journal of Aviation Psychology* 17: 41–57.

Ramos, K. D., S. Schafer, and S. M. Tracz. 2003. “Validation of the Fresno Test of Competence in Evidence Based Medicine.” *BMJ* 326: 319–321.

Rosenzweig, S., T. P. Brigham, R. D. Snyder, G. Xu, and A. J. Mcdonald. 1999. “Assessing Emergency Medicine Resident Communication Skills Using Videotaped Patient Encounters: Gaps in Inter-Rater Reliability.” *The Journal of Emergency Medicine* 17: 355–361.

Rousseau, R., S. Tremblay, and R. Breton. 2004. "Defining and Modeling Situation Awareness: A Critical Review." In *A Cognitive Approach to Situation Awareness: Theory and Application*, edited by S. P. Banbury, and S. Tremblay, 3–21. Hampshire: Ashgate.

Salmon, P. M., N. A. Stanton, G. H. Walker, and D. Green. 2006. "Situation Awareness Measurement: A Review of Applicability for C4i Environments." *Applied Ergonomics* 37: 225–338.

Salmon, P. M., N. A. Stanton, G. H. Walker, D. Jenkins, D. Ladva, L. Rafferty, and M. Young. 2009. "Measuring Situation Awareness in Complex Systems: Comparison of Measures Study." *International Journal of Industrial Ergonomics* 39: 490–500.

Saxton, E., S. Belanger, and W. Becker. 2012. "The Critical Thinking Analytic Rubric (CTAR): Investigating Intra-Rater and Inter-Rater Reliability of a Scoring Mechanism for Critical Thinking Performance Assessments." *Assessing Writing* 17: 251–270.

Shrout, P. E. 1998. "Measurement Reliability and Agreement in Psychiatry." *Statistical Methods in Medical Research* 7: 301–317.

Shrout, P. E., and J. L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin* 86: 420–428.

Shrout, P. E., and S. P. Lane. 2012. "Reliability." In *APA Handbook of Research Methods in Psychology. Vol 1: Foundations, Planning, Measures, and Psychometrics*, edited by H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher, 643–660. Washington, DC: American Psychological Association.

Sims, J., and C. C. Wright. 2005. "The Kappa Statistics in Reliability Studies: Use, Interpretation, and Sample Size Requirements." *Physical Therapy* 85: 257–268.

Skraaning, G., Jr., M. H. R. Eitrheim, N. Lau, C. Nihlwing, L. Hurlen, and T. Karlsson. 2009. *Coping with Automation in Future Plants: Results from the 2009 HAMMLAB Experiment*. Halden: OECD Halden Reactor Project.

Stanton, N. A. 2010. "Situation Awareness: Where Have We Been, Where Are We Now and Where Are We Going?" *Theoretical Issues in Ergonomics Science* 11: 1–6.

Stanton, N. A., and M. S. Young. 1999. "What Price Ergonomics?" *Nature* 399: 197–198.

Stanton, N. A., and M. S. Young. 2003. "Giving Ergonomics Away? The Application of Ergonomics Methods by Novices." *Applied Ergonomics* 34: 479–490.

Stemler, S. E. 2004. "A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability." *Practical Assessment, Research & Evaluation*. Accessed October 10, 2012. http://PAREonline.net/getvn.asp?v=9&n=4

Stuhlmann, J., C. Daniel, A. Dellinger, R. Kenton, and T. Powers. 1999. "A Generalizability Study of the Effects of Training on Teachers' Abilities to Rate Children's Writing Using a Rubric." *Reading Psychology* 20: 107–127.

Taylor, R. M. 1990. "Situation Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design." In *Situational Awareness in Aerospace Operations*. Neuilly Sur Seine: NATO-AGARD.

Vidulich, M. A., and E. R. Hughes. 1991. "Testing a Subjective Metric of Situation Awareness." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 35: 1307–1311.

Viswesvaran, C., D. S. Ones, and F. L. Schmidt. 1996. "Comparative Analysis of the Reliability of Job Performance Ratings." *Journal of Applied Psychology* 81: 557–574.

Waag, W. L., and M. R. Houck. 1994. "Tools for Assessing Situational Awareness in an Operational Fighter Environment." *Aviation, Space, and Environmental Medicine* 65: A13–A19.

Willems, B. F., and M. Heiney. 2001. "Real-Time Assessment of Situation Awareness of Air Traffic Control Specialists on Operational Host Computer System and Display System Replacement Hardware." 4th USA/Europe Seminars on Air Traffic Management Research and Development, Santa Fe, NM, USA. Eurocontrol [no pagination].

Willems, B. F., and M. Heiney. 2002. *Decision Support Automation Research in the En Route Air Traffic Control Environment*. Atlantic City International Airport, NJ: FAA William J. Hughes Technical Center.