Exceptional performing screener project: methods and findings from the field study

Gerald D. Gibb[1], Karen D. Gibb[1], J. Michael Barrientos[2], Jesse D. Flint[1], Kelly Hale[1], & Darren Wilson[2]
[1]Design Interactive and [2]Department of Homeland Security
USA

Abstract

Earlier work (Gibb, et al., 2007) discussed methodological approaches for improving threat detection skills of the U.S. airport security workforce. This paper presents the unique techniques developed and findings from an extensive field data collection effort examining differences between exceptional performers (EPs; ≥97th percentile) in the screening workforce and other screeners. The study was conducted using operational X-ray machines and sixty test bags with 119 security officers from nine U.S. airports. The empirical methods used to identify EP participants and threat stimuli are described along with the cognitive task analytical techniques used for capturing the high level skills of the exceptional performers. The findings indicated that EPs use subtle cues, distinctive techniques and approaches, refined image analysis methods and recognition of anomalies that have led to the successful detection of simulated threats (IEDs [Improvised Explosive Device], guns, knives, and other weapons). The transfer of acquired knowledge, expertise, and techniques of EPs to the workforce at large was possible because the information was transformable into discrete training elements. An illustration of key concepts is included. Recently a prototype training program based on this analysis was developed and piloted.

Introduction

The U.S. Transportation Security Administration (TSA) and Department of Homeland Security (DHS) are working to improve aviation security. The performance of the Transportation Security Officer (TSO) workforce is paramount to meeting that mission. While major advances have been made in developing and deploying advanced technologies improving detection of threats targeting the air transportation system, enhancing the performance of the workforce has been more challenging. One key performance area is the capability of the workforce to effectively use X-ray technology to detect threat objects. Since X-ray technology is the principle method to screen carry-on belongings at U.S. airport checkpoints it is critical that workforce performance is maximized. This investigation focused on identifying what effective threat detection practices are employed at the individual operator level that contribute to threat detection. The two key facets were to empirically identify the subset of the workforce with exceptional threat detection

In D. de Waard, K. Brookhuis, R. Wiczorek, F. di Nocera, R. Brouwer, P. Barham, C. Weikert, A. Kluge, W. Gerbino, & A. Toffetti (Eds.) (2014). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2013 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

skills and use established methodologies to capture their skills, abilities, and techniques.

Over the past decade several field studies and investigations (Barrientos, et al., 2011) have demonstrated there are workforce personnel that consistently excel above their peers relative to Threat Image Projection (TIP) performance measures. TIP is a comprehensive, objective and quantitative system that assesses threat detection in the operational environment.

Research methods

Apparatus and participants

There were three major elements to successfully execute this study: a) selection of "discriminator" test stimuli (threat images), b) identification of personnel from three performance tiers, and c) design and crafting of test bags. The test stimuli was a limited threat library that included only discriminator TIP images as determined by an item analysis (Barrientos, et al., 2011). Discriminator TIP images are threats that exhibit among the lowest overall detection rates but EP screeners detect at significantly higher levels than the workforce. As there were two different Threat Image Ready X-ray Systems (TRX) deployed across U.S. security checkpoints, analytical work indicated two separate threat image sets were appropriate. The Smiths Detection test Threat Image Library contained 32 images whilst the Rapiscan test Threat Image Library included 44 images.

In determining how EP personnel differed from the workforce three study groups were established – EP (Exceptional Performers) who were defined as at or above the 95^{th} percentile; mid-tier performance where the mean performance is within the 45^{th} to 55^{th} percentile; and lowest-tier where the mean performance is at or below the 10^{th} percentile, over a six-month window based on TIP performance (Barrientos, et al.; 2011). The percentile bands represented a rank ordering and do not imply substandard or inadequate performance as officers are evaluated annually and meet accepted standards. EP Screener groups were identified for each of the Smiths Detection and Rapiscan TRX operator populations; one each representing large domestic or international airports and one representing regional airports. The distinction between large and small sites is based on the operational TIP:Bag ratios set on the TRX systems and not passenger throughput or physical size of the airport. Two mid-tier and two low-tier groups, one each for the Smiths Detection and Rapiscan TRX operators were also identified. Airport sites were selected on the basis of having the highest number of TSOs within a defined tier.

There were 60 test bags created using rollaboards, shoulder bags, duffles, and backpacks; representing moderately cluttered, typical passenger carry-on items. No liquids, gels, pastes, or aerosols greater than the limit of 3.3 ounces were included. No prohibited items (e.g. knives, flammables) were used. The bags contained sufficient light organic clutter (e.g. clothing, plush dolls) to mitigate the contents from shifting during transport.

Prior to finalizing the test bag set each bag was screened by the test team and the level of clutter adjusted to provide a challenging detection environment. Test bags were 'themed' in order to closely mimic typical passenger baggage. For example, "female bags" contained stereotypical gender-specific clothing, curling irons, shoes, and jewellery whereas a child's bag would contain toys, smaller footwear. This was intentional to examine if TSOs were sensitive to relationships of articles within bags.

Specialized software and projection parameters were developed to support obtaining sufficient data for the protocol analysis (Ericsson & Simon, 1986) on the operational checkpoint TRX units. Therefore operational TIP system settings were not used as few threat stimuli would be projected across the test bag set for each participant. This innovative approach enabled participants to screen the bags using the system they normally worked. The following TRX settings and parameters were used:

R-V-R (Ratio-Variance-Random): This setting guides the frequency of TIP projections. It was set to 2-1-50 for the study. The setting generates one TIP for every two bags screened. The setting also provides for fifty percent of the TIPs as randomly projected. These settings allowed a TIP projection on every other bag screened, however, due to the random nature of the presentations, there are instances where no TIP is projected for 3, 4 or 5 bags and in other cases where a TIP is projected for every bag for 3, 4, 5 or more bags.

SDT (Secondary Decision Time): This parameter was set to 45 seconds. The SDT is the time the X-ray conveyor belt is stopped and the TIP button is selected before the TIP is considered a "Miss" and the feedback message is displayed. The maximum was selected to allow as much time as possible for the TSO to convey the cognitive, perceptual, decision, and analysis processes used screening each bag and for adequate data capture.

In some cases the TSO was talking through the review of the bag and "missed" the threat even though they had detected the threat and were talking about it, including positive identification of the location and the distinguishing characteristics when the 45 second time limit expired. In cases where there was no TIP presented and therefore no threat in the bag, the TSO continued verbalization for a time greater than 45 seconds until prompted by the facilitator to move the belt for the next bag.

Threat Category Distribution: These settings determine the projection distribution among the four threat categories. The settings differed between the Smiths Detection and Rapiscan systems to approximate the distribution of threats by threat category of the image libraries used.

Operator Logon ID and Password: A unique operator logon (e.g. 77777777) and User Name (EP Screener) was established to maintain anonymity.

Knowledge elicitation methods

Cognitive engineering, cognitive task analysis (CTA), and knowledge elicitation methodologies have advanced substantially (Hoffman, et al., 1995; Woods, 1993;

Salas & Klein, 2001; Someren, et al., 1994) since Klein, et al.'s ground-breaking study (1986). A history on the early stages of naturalistic decision making (an application of these methodologies specific to problem analysis and decision processes) is found in Lipshitz, et al. (2001). Hoffman, et al. (1998) in their examination of CTA cite numerous examples of successful applications in critical care, systems analysis, medicine, and military operations. These approaches demonstrated that highly refined skills and knowledge acquired by experts through experience can be identified and elicited. Moreover, the techniques are superior to conventional interviews, surveys, task analyses, and observational methods when the desired information is based on cognitive processes as opposed to behavioural practices (Dubois, 2002).

Approaches that have attempted to garner knowledge that relied on experts recalling past events, through a multitude of techniques, have often been met with failure to develop information in sufficient detail to develop training, build assessment tools, or establish performance criteria. Attaining vital detail requires eliciting implicit knowledge and transforming it into explicit information by exposing experts to the very tasks that require the use of the knowledge (Dubois, et al. 1995; Johnson & Johnson, 1987). Dubois (1995) maintains that verbalizations alone, if done outside the context of performing the actual job or tasks, are unlikely to generate the type of data desired. Consequently the most comprehensive and accurate data regarding cognitive processes is best attained as the expert is performing the task or a reasonable simulation of the same.

General methodology

All participants were briefed, completed background inventories, guided through a Think Aloud technique (Someren, et al., 1994) warm-up exercise, and performed the X-ray image screening task individually by scanning test bags one at a time. The flow of test bags was controlled so that only one bag image would display at a time and presented in the same sequence to each participant.

Screening each test bag consisted of two different stages characterized by the type of interaction between test facilitators and participant. The first stage was the onset of the bag image scrolling onto the display and terminated when the TIP system generated a feedback message. If no feedback was generated (Clear Bag), the facilitator urged the TSO to continue to the next bag. If the TSO pressed the TIP button before 45-seconds, indicating a threat was possibly detected, a feedback message appeared and this terminated the initial stage. During this stage the facilitator's primary function was to encourage the participant's verbalization of cognitive, perceptual and decision processes. This is the Think Aloud technique and resembles one verbalizing their thoughts as if talking to themselves. For example, a verbal protocol fragment might be "That orange area looks dark; I don't like it; I'll strip out organics and see if that wire leads from the cell phone to it". The facilitator did not probe for explanations. The technique is strictly having the participant verbalize thought processes. The facilitator called out each image analysis tool to the data recorders to reduce participant workload.

The second stage began as the TIP feedback message appeared and varied dependent on: a) whether a threat was detected, b) the nature of the feedback message (hit or miss), and c) content of the verbal protocols during the initial 45-seconds. Several cognitive engineering and knowledge elicitation techniques were employed by the facilitator during this stage and was characterized by direct interaction between facilitator and participant using specific probes, Teachback technique, or demonstration by the TSO.

Teachback and Critical Incident Technique were the primary methods and were specific probes targeted at TSO actions, cognitive processes or decisions made while screening the bag. The focus was to isolate the cues, anomalies, and threat characteristics that were perceived while viewing an image, the image analysis tools used and the purpose for using them. The verbal protocols, responses to probes, and interaction during Teachback dialogs were recorded independently by two trained observers. The data was maintained separately for each bag screened. Because of the randomness of TIP projections it was not uncommon to see the same threat on more than one occasion as a result of the limited size of the test set. Even though the same TIP was presented more than once it was presented in a different bag and in a different orientation,– thus constituting a unique presentation.

Results

Job and work experience

The mean job longevity for EP Screener personnel was 80.02 months (n = 66, SD = 29.80 months) and 71.30 months (n = 53, SD = 32.55 months) for mid-tier and low tier combined. These were not significant differences ($F_{1, 117} = 2.314$, $p < .13$). There were no significant differences in mean job longevity between Smiths Detection (mean months = 77.86, SD = 31.02, n = 35) and Rapiscan (mean months = 82.45, SD = 28.67, n = 31) EP Screeners ($F_{1, 64} = .39$, $p < .53$).

Information was collected on the participant's previous employment prior to working as a Transportation Security Officer (TSO). This was an exploratory component to determine if there were trends regarding prior law enforcement, security or X-ray experience (e.g. food services or industrial inspection). Based on the findings of this limited sample the workforce is comprised of individuals who came from a broad range of previous job careers.

Threat detection performance

Threat detection metrics, although captured both manually and as part of the TIP automated database during the study, are not reported or analyzed. It became apparent that hits, misses, and false alarms in many cases may not have been accurate markers of threat detection performance. The source of error variance was two-fold: a) the knowledge elicitation and cognitive engineering techniques employed, and b) the validity of the metrics themselves.

Verbal protocols

Verbal protocols generate a substantial amount of information that require using segmenting, coding and analysis procedures (Ericsson & Simon; 1996). This study is no exception and produced a comparatively larger database than similar efforts using comparable techniques. This resulted from relatively high number of participants, task complexity, inclusion of two different systems, and number of events (bags screened). Each bag constituted an experimental trial and generated a unique set of verbal protocols. Consequently it is impractical to provide an exhaustive compilation of the verbal protocols. The protocol analyses were executed separately for each system. The data were sorted by threat category (e.g. IED, knife, etc.) and by specific threat image. This process facilitated organizing, segmenting and interpreting the vast number of verbal protocols. Each image analysis function used was coded sequentially as occurred such that they were lined with decision and perceptual processes. Encoding participant group enabled a comparison between groups across each specific threat type (see Table 1).

Whilst not practical to include an extensive summary of protocols from 6,000 events, illustrative verbal protocol from the three performance groups are offered to demonstrate that different performance groups use very different cues for threat detection and X-ray image analysis. Table 1 provides these comparisons along key areas of interest:

Table 1. Overall differences and sample verbal protocols across key dimensions

	Performance Tier Groups		
Dimension	**EP Screener**	**Mid Tier**	**Low Tier**
Cues	Thin black straight line indicative of a knife	9-volt battery is automatic bag check	Many organics are a problem
Major Difference	EPs respond to subtle, distinctive cues that are linked directly to threats whereas the other performance groups respond to generic cues		
Anomalies	Changes in the texture and density of organics occurs with underlying or overlapping explosive masses	Too much organics	Too much clutter
Major Difference	EPs respond using rule-based, specific visual elements while other performance tiers respond generically to global assessments such as colour		
Knowledge	Greenish-blue Polaroid batteries are rare and would not be carried without associated equipment	9-volt batteries are often found in TIPs	Too many batteries
Major Difference	EPs determine the validity and reason for an item in context whereas others apply general rules. The process is more analytical and hypothesis based.		
Analogues	Curling irons don't have thin non-voltage carrying wires or components in handle	Camera is too green	Camera has no lens
Major Difference	EPs recognize when common items vary and have been tampered with or contain additional, un-expected materials; then isolate the cause for the discrepancy		
Appraisal	Assessing relationship among bag contents (e.g. Men don't normally carry curling irons)	Shoes come in pairs	No or little evidence at determining relationships among articles in a bag

Major Difference	EPs assess the "theme" of a bag; detecting subtle deviations or anomalies from the norm		
Tools	Use of black/white tool to highlight small wires or "High/Low" incrementally to define the edges of organics	Black/white tool lightens image. Tool use generally to treat image holistically and remove colour.	Rigid, consistent use of the same pattern of tool use regardless of image dynamics
Major Difference	EPs select tools to achieve specific purposes related to the image or information desired; others use tools procedurally		
Strategies	Often systematic and methodical method; although individualized	Procedural applications	Often focus on specific objects/areas and not systematic approach
Major Difference	EPs have well defined strategies that are diverse and involve testing assumptions		
Approaches	Scanning as image scrolls on; noting anomalies and areas of concern – examine entire bag then focus on "hot spots"	Generic approaches as presence of batteries and organics or organics and electronics are considered a threat.	If cannot immediately clear a bag then suspect it
Major Difference	EPs use well defined and disciplined approaches.		
Techniques	Focus on items that don't "fit' with contents. Use of image tools to extract additional information to confirm or dismiss suspicions.	Typically a consistent use of the same tools in the same sequence	Typically a consistent use of the same tools in the same sequence
Major Difference	EPs have substantial command over the image analysis features, using them to analyze and test assumptions about the contents of bags.		

The synchronization of TIP databases with the verbal protocols provided a unique opportunity to examine differences between EP personnel and other participants to specific threat images (TIPs) because it was possible to isolate the verbal protocols in response to each TIP projection (Note: A TIP projection could occur in any test bag, consequently each TIP appeared in different bag environments). In some cases a threat was salient and un-obscured whilst in other instances it could be masked within dense clutter.

Table 2 illustrates the differences of perceptual cues or physical characteristics of an IED across different groups. Representative verbal protocols are provided for each group. The sample sizes are number of participants within each group that were presented with the image.

The most prominent feature of this IED was a curling iron containing a small metallic detonator in the handle. EPs focused on anomalies with the curling iron, emphasizing physical characteristics that drew comparisons with typical curling irons. The texture, hue, and presence of unexpected elements (including additional wiring) were physical cues that led to successful threat detection despite the entire IED was not visible. The query regarding presence of two curling irons in the bag indicates another anomaly as did the association of the curling iron to an unknown dense organic. These protocols are indicative of examining the suspect item in relationship to other bag contents.

Table 2. Differences among groups (threat characteristics) for an IED

EP (n = 20)	Mid Tier (n = 9)	Lowest Tier (n = 10)
▪ Extra wire coming out of curling iron ▪ Why two curling irons? ▪ Two different shapes in handle – doesn't belong ▪ Additional component in handle; det clear with B/W ▪ Texture of barrel different than normal ▪ Two metal things in handle; why two wires? ▪ Heavy organic next to curling iron ▪ Two many cords around it ▪ Darker handle than normally see, should not be blue inside ▪ Something concealed in handle	▪ Mass right here could be something, curling iron looks ok ▪ Just the shoes with massive shanks ▪ Automatic TIP; looks like pocket knife ▪ See curling iron; not a normal curling iron ▪ Don't know what to make of this bag; curling iron – nothing coming off of it ▪ Dark, indicator of being Disney gun – saw trigger real fast ▪ We have couple of pairs of shoes, curling iron, wires ▪ Straight TIP – the thing had a detonator in back	▪ There is something in shoe (*was not in shoe*) ▪ Looks like we have a mass here, looking for wires ▪ Looks ok to me, clear ▪ Too many wires going to nothing, could be a knife ▪ All electronics one side, all shoes on the other ▪ Some kind of hair thing; don't think there is a TIP ▪ Long metal bar going through here, plus whatever so dark in the bag

In comparison, only occasionally did mid-tier personnel recognize anomalies with the curling iron. Generally the curling iron is verbalized but dismissed as a potential threat. There was only one verbal protocol indicative of tampering. Low tier participants failed to identify anomalies and did not detect the threat. In nearly all cases attention to the curling iron was tangential.

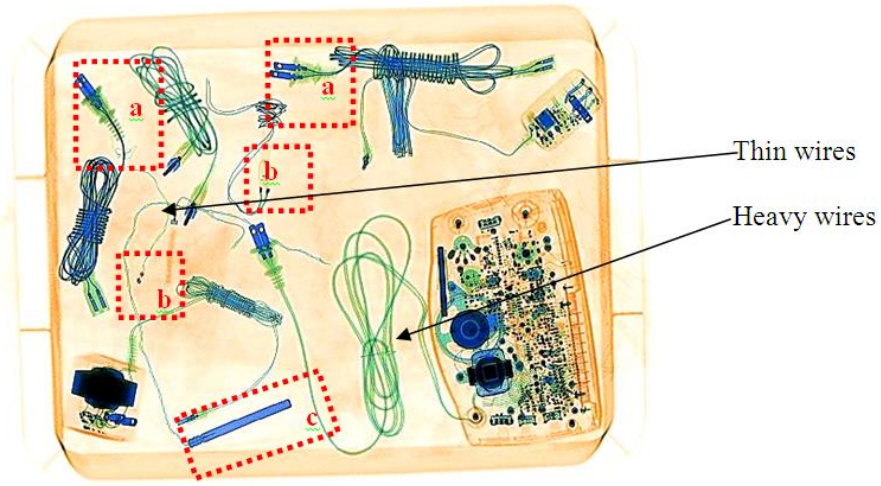

Figure 1. Illustration of visual cues used by EP Screeners

Differences between EP personnel and other groups were representative across the full range of threat objects. There was considerable consistency of the findings within groups. For example, specific knowledge of the physical cues and anomaly of

Polaroid batteries was almost universal among EPs whilst uncommon among other participants. Figure 1 provides an illustration of the unique knowledge and visual cues employed by EPs in detecting detonators. Visual cues such as: a) ordinary use wires usually terminate in plugs or jacks, b) fine wires that end in dots indicate a possible detonator, and c) presence of a, straight, slender cylinder are coupled with knowledge (e.g. thick wires are not capable of carrying the small amount of current needed for detonators) to successfully identify detonators. Figure 1 was prepared for the prototype training program. Whilst the differences are subtle and seemingly small, collectively they represent knowledge that improves threat detection.

A prototype IED training module for new hires was developed from a modest part of the protocol analysis results. The training used elements that discriminated between EPs and other participants. Training images used uncomplicated bags and threats that highlighted features used by EPs. Whilst the content was based on the protocol analysis, the research team built and captured images that increased the saliency of the cues and approaches suitable for an inexperienced group.

Summary

Although individual differences exist amongst EP Screeners in the use of image analysis tools, strategies, and routine approaches to screening, there is a genuine consistency distinguishing them from other screening officers. EP Screeners often began screening as the leading edge of the bag scrolled into view. They examine bags for relationships existing among the contents, identify anomalies, and focus on items that appear out of place. For example, recognition that electronics have only minimal, low density organics. They apply knowledge that 9-volt batteries are uncommon in typical electronics carried aboard aircraft.. They appreciate hair management products (hair dryers, curling irons) don't have dark components in the handles or include fine wires incapable of handling AC electrical current.

EP Screeners are attuned to subtle perceptual features and anomalies within bags (e.g. thin, straight black lines are not characteristic of most carry-on items; typical organics are not dense and have varied textures; highly saturated orange hues without definitive shape is uncommon; multiple pairs of shoes are not miss-matched in size or vary significantly in colour or density). Perhaps one of the most illustrative examples of recognizing perceptual nuances is when an EP Screener, while screening a particularly cluttered test bag with considerable organics, detected a change in the texture and shading within an area of overlapping organic material. By incrementally changing the contrast of the image using the High/Low image analysis function the shape of a glove began to emerge. The officer changed the image using the Black/White (monochrome) functionality, removing colour, and the outline of the "glove" became pronounced against the background of other organics. After confirming the anomaly existed the officer questioned "why is only one glove packed?", "why would the glove appear stiff as if it something was inside it?" and "what is that small metal piece inside?" These perceptual and cognitive processes were representative of EP Screener expertise throughout the study and were identified successfully using the applied knowledge elicitation techniques.

Acknowledgment

This work was funded by the U.S. Department of Homeland Security.

References

Barrientos, M., Gibb, G.D., & Eldredge, D. (2011). *Test and Evaluation Report: Exceptional Performing Screeners Study - Analysis of the National Threat Image Projection Database Characterizing Transportation Security Officer and Threat Image Performance*, Washington, D.C., USA: U.S. Department of Homeland Security.

Dubois, D.A., Shalin, V. L., & Borman, W.C. (1995). *A Cognitively-Oriented Approach to Task Analysis and Test Development*. Arlington, VA, USA: Office of Naval Research, (Contract No. N00014-93-C-0199).

Dubois, D. (2002). Leveraging hidden expertise: Why, when and how to use cognitive task analysis. In D. Dubois (Ed.), *Creating, implementing, and managing effective training and development*, Jossey-Bass Publishing of Wiley Company, San Francisco, CA., USA.

Ericsson, K. & Simon, H. (1996). Protocol analysis: *Verbal reports as data*. MIT Press, Cambridge, MA., USA.

Gibb, G, Fischer, B. & Cabeca, B. (2007) Methodological approach to advancing airport screener X-ray threat detection skills. In D. de Waard, B. Hockey, P. Nickel, & K. Brookhuis, *Human Factors Issues in Complex System Performance* (pp. 443-454). Shaker Publishing , Maastricht, The Netherlands.

Hoffman, R. Shadbolt, N, Burton, A. & Klein, G. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes, 62*, 129 – 158.

Hoffman, R. Crandall, B. & Shadbolt, N. (1998). Use of the Critical Decision Method to elicit expert knowledge: A case study in the methodology of Cognitive Task Analysis. *Human Factors, 40*, 254 – 276.

Johnson, L & Johnson, N. (1987). Knowledge elicitation involving Teachback interviewing, (91-108). In A. Kidd, (Ed.), *Knowledge Elicitation for Expert Systems: A Practical Handbook*, Plenum Press, New York, USA.

Klein, G.A., Calderwood, R. & Clinton-Cirocco, R. (1986). Rapid Decision Making on the Fire Ground. In Proceedings of the 30th Annual Meeting of the Human Factors Society, (pp. 576-580), Dayton, OH., USA.

Klein, G., Calderwood, R. & MacGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics, 19*, 462-472.

Lipshitz, R, Klein, G. Orasanu, J & Salas, E. (2001). Focus Article: Taking Stock of Naturalistic Decision Making. *Journal of Behavioral Decision Making, 14*, 331-352.

Salas, E. & Klein, G. (2001). Expertise and naturalistic decision making: An overview. In E. Salas & Klein, G. (Eds.), *Linking Expertise and Naturalistic Decision Making* (pp. 3-10). Lawrence Erlbaum Associates, Mahwah, NJ., USA.

Someren, W., Barnard, Y, & Sandberg, J. (1994). *The Think-Aloud Method: A Practical Guide to Modeling Cognitive Processes*, Academic Press, London, UK.

Woods, D. (1993). Process-tracing methods for the study of cognition outside of the experimental laboratory. In G. Klein, Orasanu, J, Calderwood, R. & Zsambok, C. (Eds.), *Decision Making in Action: Models and Methods*, Ablex Publishing Co., NJ, USA.