



Automation in visual inspection tasks: X-ray luggage screening supported by a system of direct, indirect or adaptable cueing with low and high system reliability

Alain Chavaillaz, Adrian Schwaninger, Stefan Michel & Juergen Sauer

To cite this article: Alain Chavaillaz, Adrian Schwaninger, Stefan Michel & Juergen Sauer (2018) Automation in visual inspection tasks: X-ray luggage screening supported by a system of direct, indirect or adaptable cueing with low and high system reliability, Ergonomics, 61:10, 1395-1408, DOI: [10.1080/00140139.2018.1481231](https://doi.org/10.1080/00140139.2018.1481231)

To link to this article: <https://doi.org/10.1080/00140139.2018.1481231>



© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 26 Dec 2018.



[Submit your article to this journal](#)



Article views: 1100



[View related articles](#)



[View Crossmark data](#)



Citing articles: 8 [View citing articles](#)

Automation in visual inspection tasks: X-ray luggage screening supported by a system of direct, indirect or adaptable cueing with low and high system reliability

Alain Chavaillaz^a, Adrian Schwaninger^b, Stefan Michel^b and Juergen Sauer^a

^aDepartment of Psychology, University of Fribourg, Fribourg, Switzerland; ^bInstitute Humans in Complex Systems, University of Applied Sciences and Arts Northwestern Switzerland, Olten, Switzerland

ABSTRACT

The present study evaluated three automation modes for improving performance in an X-ray luggage screening task. One hundred and forty participants were asked to detect the presence of prohibited items in X-ray images of cabin luggage. Twenty participants conducted this task without automatic support (control group), whereas the others worked with either indirect cues (system indicated the target presence without specifying its location), or direct cues (system pointed out the exact target location) or adaptable automation (participants could freely choose between no cue, direct and indirect cues). Furthermore, automatic support reliability was manipulated (low versus high). The results showed a clear advantage for direct cues regarding detection performance and response time. No benefits were observed for adaptable automation. Finally, high automation reliability led to better performance and higher operator trust. The findings overall confirmed that automatic support systems for luggage screening should be designed such that they provide direct, highly reliable cues.

Practitioner summary: The present study confirmed previous findings showing better detection performance in X-ray images of luggage when supported by automation providing direct, highly reliable cues. Furthermore, participants used adaptable automation only to select their preferred level of automation. This behaviour did not provide the benefits expected under adaptable automation.

Abbreviations: LOA: Level of automation; NC: No cue; IC: Indirect cue; DC: Direct cue; AC: Adaptable cueing; HiRel: high reliability; LoRel: low reliability; CTPA: Checklist of Trust between People and Automation; ARE: automation reliability estimates; TCM: Two-component model

ARTICLE HISTORY

Received 31 October 2017
Accepted 20 May 2018



KEYWORDS

Adaptable automation;
airport security; visual
inspection; system
reliability; performance

1. Introduction

Visual inspection represents an important task in different work domains, such as medical diagnosis, industrial quality control, and security screening. It involves a complex visual search and decision task in which, for example, tumours, faulty devices, or weapons have to be detected. The specific characteristics of the work domain influence the way the task is carried out (Drury 1992; Drury 2001; See 2012). In airport security X-ray screening, this task is characterised by long periods of sustained vigilance resulting in high levels of mental workload (Warm, Parasuraman, and Matthews 2008). Stressors such as time pressure, noise and high task load contribute to the suboptimal working conditions found in this task (McCarley et al. 2004; Michel

et al. 2014; Baeriswyl, Krause, and Schwaninger 2016). Luggage inspection by airport security officers (screeners) represent a challenge for their perceptual and cognitive capacities (Harris 2002). For instance, recognition of prohibited items becomes more difficult when items are rotated, placed in visually complex bags or other objects superimpose the prohibited item (Schwaninger, Hardmeier, and Hofer 2005; Schwaninger et al. 2008). Furthermore, low target prevalence increases the risk of prohibited items being missed by screeners (Wolfe et al. 2007, 2013). Recent developments in airport security scanner technology (X-ray machines) has allowed now the automatic detection of explosives (Wells and Bradley 2012; Sterchi and Schwaninger 2015) and guns (Roomi and

CONTACT Alain Chavaillaz  alain.chavaillaz@unifr.ch  Department of Psychology, University of Fribourg, Rue de Faucigny 2, 1700 Fribourg, Switzerland

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Rajashankari 2012; Mery et al. 2013). Automation in airport security systems is expected to become an increasingly important issue in human factors.

In a visual inspection task, automatic support can guide attention and enhance operator perception (Lee and Sanquist 1996) through textual or pictorial cues, which facilitates target identification. Previous work investigated the impact of two types of cues: indirect and direct ones (Goh, Wiegmann, and Madhavan 2005). Indirect cues indicate the general presence or absence of a target, whereas direct cues additionally show the exact target location on the display. There is research evidence that target detection performance can be improved by providing indirect cues (McCarley et al. 2003; Rice and McCarley 2011) but even more so by direct cues compared to systems without automatic support (Goh, Wiegmann, and Madhavan 2005; Wiegmann et al. 2006).

The degree of performance improvement largely depends on the level of automation reliability. It was demonstrated in different studies that low system reliability brings about low performance in a range of work domains such as process control (Bailey and Scerbo 2007; Ma and Kaber 2007; Rovira, McGarry, and Parasuraman 2007; Wickens and Dixon 2007; Chavailleaz, Wastell, and Sauer 2016a) and visual inspection (Goh, Wiegmann, and Madhavan 2005; Rice and McCarley 2011). Such performance degradations are usually accompanied by low ratings of trust towards automation (Wiegmann, Rich, and Zhang 2001; Dzindolet et al. 2003; Bailey and Scerbo 2007; Ma and Kaber 2007) and low compliance with automation recommendations (Rice and McCarley 2011).

Being sensitive to automation reliability does not necessarily mean that operators are able to estimate it accurately. Several studies reported a systematic underestimation of reliability (Wiegmann, Rich, and Zhang 2001; Wiegmann 2002; Sanchez, Fisk, and Rogers 2004). Three factors were found to have an influence on the magnitude of the estimate: actual reliability, time-on-task, and rate of compliance with automation recommendations. First, perceived reliability ratings are closer to actual system reliability under high-reliability automation than low-reliability automation (Sanchez, Fisk, and Rogers 2004). Second, increasing experience with the same system improves the accuracy of the estimate, though a sudden reliability change can strongly reduce the accuracy score (Wiegmann, Rich, and Zhang 2001). Finally, participants always agreeing with automation recommendations showed the most accurate reliability ratings (Wiegmann 2002).

A similar pattern was found for trust. Operators have difficulties in matching their trust to the current reliability of the system (i.e. trust calibration), with a perfect calibration being hardly observed (Wiegmann, Rich, and Zhang 2001; Wiegmann 2002). A discrepancy between trust and system reliability can result in an inappropriate use of automation (Parasuraman, Molloy, and Singh 1993). Trust ratings that are lower than the corresponding system reliability may result in low compliance with automation recommendations even if they are correct. Conversely, trust levels that are higher than the corresponding system reliability may result in operators showing compliance with automation recommendations even if they are incorrect. The examples demonstrate that joint human-machine performance does not only depend on objective automation properties such as reliability but also on operator perceptions.

Based on studies mentioned above one would assume that automatic support in visual inspection tasks can provide considerable benefits to operators. However, an interesting question would be whether these benefits could be further increased by making use of modern concepts such as adaptable automation (Scerbo 2006). Adaptable automation allows levels of automation (LOA) to be changed by the operator (e.g. switching between indirect and direct cues) to achieve a good match between operator needs and the level of automatic support provided (Sauer, Kao, and Wastell 2012). These operator needs may vary over time (e.g. onset of fatigue) and between operators (e.g. differences in competence levels). Empirical research in multitasking environments showed that operators benefitted more from flexible designs such as adaptable automation than static automation where no switching between support levels was possible (Parasuraman et al. 1993; Kaber and Riley 1999; Inagaki 2003; Sauer, Kao, and Wastell 2012). These results are encouraging and raise the question whether adaptable automation would also be beneficial in single-task environments such as airport security X-ray screening. For instance, operators in adaptable automation may decide to switch from direct cues to no cues to avoid being distracted by cues pointing at irrelevant objects. This may result in faster response time than with direct cues, particularly when automation reliability is low.

1.1. Present study

The goal of the present study was to investigate how different automation modes (indirect cues, direct cues

or adaptable cueing) affect performance under low and high system reliability.

Student participants were asked to detect prohibited items (either a knife or a gun) in a series of X-ray images of hand luggage during two testing phases. In the first phase (pre-test), the screening task was completed to measure participant's ability to detect prohibited items. This was to control for possible differences in ability and aptitude in the experimental groups. In the second phase (main test), participants worked with a support system offering one of the following levels of automatic support: direct cue (indicating the target location in the X-ray image), indirect cue (indicating that a target was present without specifying its location in the X-ray image), or adaptable cueing (operators could freely choose between indirect cue, direct cue and no cue). Furthermore, automation reliability was manipulated at two levels: low reliability (about 60%) or high reliability (about 80%). A control group performed the main test without automatic support. Several dependent variables were obtained, including detection performance (d'), response times, trust, compliance, and reliance on automation.

Several predictions were made based on previous automation studies. Some of these predictions were based on research from other work domains to examine the transferability of the findings to airport security X-ray screening. Better performance was expected when direct cues rather than indirect cues were provided (Goh, Wiegmann, and Madhavan 2005). Based on research in multitasking environments, we also predicted that performance under adaptable automation would be at least as good as than under conditions of static automation (Parasuraman et al. 1993; Sauer, Kao, and Wastell 2012). Furthermore, performance (detection of prohibited items and response time) and trust was expected to be higher for high-reliability automation than low-reliability automation or no cue (Wiegmann, Rich, and Zhang 2001; Goh, Wiegmann, and Madhavan 2005; Bailey and Scerbo 2007; Rovira, McGarry, and Parasuraman 2007; Wickens and Dixon 2007).

2. Methods

2.1. Participants

One hundred and forty student participants from the University of Fribourg (99 females, 41 males), aged from 18 to 40 years [mean (M) = 22.26, standard deviation (SD) = 3.26], took part in this study. They received course credits in return for their participation.

2.2. Ethical considerations

The Ethics Committee of the Department of Psychology at the University of Fribourg (Switzerland) gave their approval for this study.

2.3. Apparatus and stimuli

The X-ray luggage screening simulation was controlled by a Matlab script using the Psychtoolbox (Brainard 1997; Pelli 1997; Kleiner et al. 2007). Stimuli were presented on a 17" LCD flat screen at a resolution of 1280×1024 pixels and a refresh rate of 60 Hz driven by a Dell PC on Microsoft Windows XP operating system. Participants were seated in a dimly lit room at an approximate distance of 0.60 m from the screen and were free to move their head. The height and width of displayed X-ray images of luggage covered about $12.18 \times 13.66^\circ$ of visual angle.

Images from two different versions of the X-Ray Object Recognition Test (X-Ray ORT; Hardmeier, Hofer, and Schwaninger 2005; Schwaninger, Hardmeier, and Hofer 2005) served as stimuli. Each version of the test consists of 256 X-ray images of hand luggage. Each piece of luggage was displayed twice, once with a threat item (target) and once without. Guns and knives were used as threat items because novices are generally more familiar with the shapes of these items (compared to explosives, electronic shock devices, etc.) from every-day life, or at least from every-day multimedia entertainment. Since novices do not know the meaning of colours in X-ray images, the images were presented in grayscale (for more information on the X-Ray ORT, see Schwaninger, Hardmeier, and Hofer 2005). Images were counterbalanced for target point of view (canonical or not), bag complexity (small versus large number of items in the bag) and objects overlap (little versus strong overlap).

2.4. Luggage inspection simulation

A purpose-built simulation environment, called luggage inspection simulation (LIS), was used to simulate the visual inspection task of screeners. Participants had to visually search each X-ray image and decide as accurately and as quickly as possible whether it contained a target item or not (i.e. yes-no task in signal detection theory, Macmillan and Creelman 2005). The left and right mouse buttons were used to carry out the task. The target-presence/target-absence button mapping was counterbalanced across participants and remained constant across the experiment. The target item was either a gun or a knife. In each trial, at first a

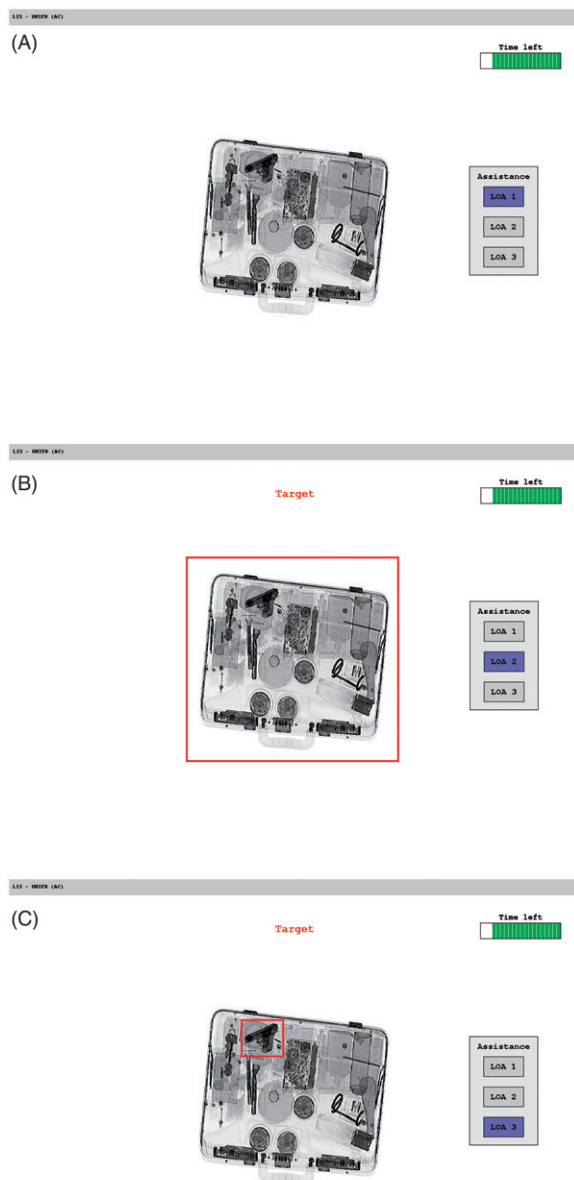


Figure 1. Interface of the luggage inspection simulation (LIS) in adaptable cueing mode set at level of automation 1 (LOA-1, panel A). The control panel for LOA selection is displayed to the right of the X-ray image and the bar countdown timer in the top right corner of the interface. At LOA-2, when a target is detected, the piece of luggage is surrounded by a frame and the word “Target” is displayed above the piece of luggage (panel B). At LOA-3, only the object detected rather than the entire piece of luggage is surrounded by a frame (panel C).

fixation cross was presented for 500 ms in the middle of the white screen, followed by an X-ray image of a piece of hand luggage. In general, the image (Figure 1) remained on the screen until the participant

provided an answer (no target localisation was required). A blank screen was displayed for 500 ms between trials. If the participant did not answer within 20 s, the trial stopped, being scored as a target absent response. During each trial, the remaining time was indicated to participants by a bar countdown timer as depicted Figure 1 (i.e. the number of vertical bars indicated the number of seconds left to answer). The experiment was divided into blocks of 64 trials (50% target-present).

To support participants in their detection task, the LIS provided an automatic support system with three LOA. The design of the system was based on the work of Goh, Wiegmann, and Madhavan (2005) and the LOA corresponded to the lowest three LOA from the model of Sheridan and Verplank (1978). At LOA-1, the support system provided no cue that signifies the presence (or absence) of a target (Figure 1(A)). At LOA-2, the support system provided an indirect cue: when the automation detected the presence of a target, the piece of luggage was surrounded by a red frame and the word ‘Target’ was displayed in red above the frame (Figure 1(B)). The absence of a target was signalled by the absence of the frame and the words ‘Target absent’ written in green. At LOA-3, the support system provided a direct cue: it indicated the exact location of the target by a red frame (the warning ‘Target’ was written above the X-ray image; Figure 1(C)). In the adaptable cueing mode, a control panel appeared on the right side of the interface, which allowed participants to select different LOAs (Figure 1(A)). No miscues were used (i.e. an object other than the target is cued in a target-present trial) but participants were not informed about it. Objects cued in false-alarm trials were selected for their visual similarity with a potential target. Please note that the X-ray images enjoyed a high-level of ecological validity (i.e. they are based on real-life luggage inspection) even though the weapon displayed in the figure is easy to detect.

2.5. Design

The study used an incomplete 4×2 design with seven groups (Table 1). The first between-subjects factor was automation mode. During the testing session, participants had to perform the visual inspection task either without automatic support (no cue, NC) or with one of the three automation modes. In the indirect cue mode (IC), the support system was fixed at LOA-2, whereas in the direct cue mode (DC), it was fixed at LOA-3. In adaptable cueing mode (AC), participants

Table 1. Experimental design.

Automation mode	Reliability
No cue	–
Indirect cue	High Low
Direct cue	High Low
Adaptable cueing	High Low

$n = 20$.

could freely select one of the three LOAs and change it at any time. The second between-subjects factor was system reliability. In the high-reliability condition (HiRel), the system had a detection performance of $d' = 1.774$ (hit rate = 81.25%, false-alarm rate = 18.75%, see 'Dependent variables' section for the formula). In the low-reliability condition (LoRel), the detection performance of the system was $d' = 0.637$ (hit rate = 62.5%, false-alarm rate = 37.5%). The response bias of the system (criterion c , see 'Dependent variables' section for the formula) was set to 0 in both reliability conditions. Participants were randomly assigned to one of the seven experimental conditions.

2.6. Dependent variables

Participant performance was assessed by three measures. *Detection performance* referred to participant ability to indicate the presence or absence of a target in X-ray images and was measured by $d' = z(H) - z(FA)$, where H refers to hit rate and FA to false-alarm rate of participants. *Response bias* corresponded to participant response behaviour, that is, the tendency to respond 'yes' or 'no'. It was computed by the following formula: $c = -0.5 \times [z(H) + z(FA)]$ using hits and false alarms from participants. For more information on these measures, see Green and Swets (1966) and Macmillan and Creelman (2005). Response time in milliseconds (ms) was measured as the time that elapsed between the onset of the X-ray image and the participant pressing the mouse button. Response time was measured and analysed separately for 'target-present' and 'target-absent' trials.

Two measures were used to evaluate the participants' propensity to heed the advice of the automation (based on Meyer 2001). *Compliance* referred to participants' inclination to confirm that a target was present when the automation had indicated one. *Reliance* corresponded to participant's propensity to confirm that no target was present when the automation had indicated the absence of a target in the X-ray image. Both measures are expressed as percentages

(corresponding to the methodological approach adopted for instance by Rice and McCarley 2011).

For participants working under adaptable cueing, two additional measures of automation use were employed. *Preferred LOA* referred to the level of automation selected most of the time by a participant (i.e. the LOA with the highest frequency across trials)¹. *Frequency of LOA changes* was determined by the mean number of switches between automation levels during a trial.

Trust was assessed by the Checklist of Trust between People and Automation questionnaire (CTPA; Jian, Bisantz, and Drury 2000). Participants had to rate 12 items on a seven-point Likert scale (ranging from 'not at all' to 'totally agree'). An item example was 'I am confident in the system'.

Perceived reliability of automation was measured by one item ('How reliable was the support system during task completion (0–100%)?'). Since system reliability varied across groups, a direct comparison for this variable would be biased. Therefore, we computed the index *Accuracy of Reliability Estimate* (ARE) that indicated the difference between perceived and objective reliability of automation. A score of zero indicates a perfect calibration, whereas a negative score corresponds to an underestimation of automation reliability and a positive one to an overestimation.

Self-confidence was measured by a single item adapted from Lee and Moray's study (1992): 'How confident were you in your ability to detect dangerous objects?'. This item was rated on a 10-point Likert scale (ranging from 'not at all' to 'completely').

Subjective workload was measured by the NASA-TLX (Hart and Staveland 1988). Six items reflecting specific facets of workload (i.e. mental demands, physical demands, temporal demand, performance, frustration, and effort) were rated on a 20-point Likert scale (ranging from 'not at all' to 'extremely').

2.7. Procedure

The experiment was divided into two phases (pre-test and main test) using the LIS described above. The purpose of the pre-test was to measure participant's ability to detect threat objects in hand luggage without automatic support. This was to control for possible differences in ability and aptitude in the experimental groups. The main test investigated the impact of automation mode and system reliability on dependent variables.

During the pre-test, participant's ability to detect target objects in hand luggage was tested with

128 X-ray images from an older version of the X-Ray ORT (Hardmeier, Hofer, and Schwaninger 2005; Schwaninger, Hardmeier, and Hofer 2005) and not used again in the main test of the experiment. At the start of a session, the instructions about the task and the response modalities were displayed on the screen. The pre-test started with a practice block of eight trials followed by two experimental blocks of 64 trials each (50% target-present trials). In the practice block, participants were informed whether their response was correct or not. No feedback was given in the experimental blocks. To become familiar with the target items (i.e. guns and knives), each of the two sets of target items were presented for 10 s at the beginning of the practice block and again before the first experimental block. A break of 5 min was scheduled between experimental blocks. In each trial, a fixation cross was presented for 500 ms in the centre of the white screen, followed by an X-ray image displayed for 4 s and a blank white screen shown for 16 s. As soon as a response was given, the trial stopped and the next one started. If participants did not provide a response within 20 s, the trial stopped and the next one started. Giving no response was scored as a 'target absent' response.

At the beginning of the main test, participants were informed about the automation mode by means of on-screen instructions. No details were given about automation reliability other than that automation might sometimes fail. Participants were made familiar with the automation mode during a practice block of 32 trials prior to completing four experimental blocks of 64 trials (50% target-present). A 5-min forced break was scheduled between experimental blocks. The same trial sequence (i.e. fixation cross, X-ray image, blank white screen) as in the pre-test was used, except that the X-ray images stayed on screen for 20 s (rather than 4 s) or until the participant made a response. Again, if no response was given, it was considered as a target-absent response. Furthermore, as in the pre-test, each target set (i.e. guns and knives) was displayed for 10 s at the beginning of the practice block and before the first experimental block. During the practice block, participants were informed whether they provided a correct response or not. Participants in the AC condition started each phase of the experiment at LOA-1. They could only change LOAs when an X-ray image was on screen. The default LOA for each new trial corresponded to the last LOA selected in the previous one. After the completion of the last experimental block, participants were asked to complete a series of questionnaires (trust towards

automation, perceived automation reliability, self-confidence in their ability to achieve the task, and subjective workload). Participants took about 45 min to complete the experiment.

2.8. Data analysis

Data from the pre-test were analysed to examine whether participants' ability to detect a target object (i.e. d') was equivalent between main-test groups. Levene's test showed equal variances across main-test groups, $F(6,133) = 0.832$, $p = 0.547$. We employed an alpha level of 0.20 in the analysis of variance, following a procedure of null hypothesis testing adopted by Onnasch (2015). The one-way ANOVA showed no significant difference between the seven groups, $F(6,133) = 1.054$, $p = 0.394$, $\eta^2_{\text{partial}} = 0.045$. The results indicated that participants in each experimental condition had similar screening ability. Therefore, it was not necessary to use screening ability as a covariate in the subsequent analyses.

Due to the incomplete 4×2 design, two separate analyses of variance were carried out (following a procedure used by Rice and McCarley 2011). In a first step, a one-way analysis of variance (ANOVA) (including all seven experimental groups) was carried out to evaluate the difference between no cue and the six automation conditions. For this reason, only multiple comparisons involving the control group (i.e. no cue) are reported in the results section. In a second step, the influence of automation mode (IC, DC, and AC) and system reliability (LoRel, HiRel) was assessed by a two-way ANOVA including all six experimental conditions involving automation. If required, Keppel-modified Bonferroni corrections were used to adjust the level of significance for multiple comparisons, following an approach used by Rice and McCarley (2011). All main effect and interactions are reported in Table 2.

3. Results

3.1. Performance

3.1.1. Detection performance

The one-way ANOVA using detection performance (d') of human-automation team as dependent variable revealed an effect of automation mode (Table 2 and Figure 2). Multiple comparisons for each of the six automation modes compared to NC showed that detection performance in DC mode under high reliability was higher than in NC mode (HiRel),

Table 2. *F*-value, significance level and effect size for the main effect of the one-way (seven-level) ANOVA and for the main and interaction effects for automation conditions and system reliability

Variable	One-way ANOVA			Automation			Reliability			Automation × reliability		
	<i>F</i> _a	<i>p</i>	η^2_{partial}	<i>F</i> _b	<i>p</i>	η^2_{partial}	<i>F</i> _c	<i>p</i>	η^2_{partial}	<i>F</i> _b	<i>p</i>	η^2_{partial}
Performance												
Detection ability	3.358	0.004	0.132	3.733	0.027	0.061	3.901	0.051	0.033	4.292	0.016	0.070
Response bias	1.969	0.74	0.082	1.271	0.284	0.022	3.078	0.082	0.026	0.029	0.972	0.001
Response time												
Target present	11.000	<0.001	0.332	27.272	<0.001	0.324	5.835	0.017	0.049	0.369	0.734	0.005
Target absent	4.968	<0.001	0.183	6.456	0.002	0.102	5.233	0.024	0.044	4.572	0.012	0.074
Use of automation												
Compliance	–	–	–	7.911	0.001	0.124	80.083	0.001	0.417	0.085	0.918	0.002
Reliance	–	–	–	0.644	0.647	0.011	23.105	<0.001	0.171	0.437	0.647	0.008
Subjective measures												
Trust	–	–	–	1.578	0.221	0.027	14.176	<0.001	0.112	1.645	0.198	0.029
ARE	–	–	–	4.566	0.012	0.076	21.875	<0.001	0.165	1.306	0.275	0.023
Self-confidence	0.812	0.562	0.035	0.904	0.408	0.016	0.046	0.830	<0.001	1.666	0.194	0.028
Perceived workload	0.706	0.645	0.031	1.875	0.158	0.032	0.032	0.840	<0.001	0.174	0.840	0.003

Significant effects are in boldface.

ARE: automation reliability estimates.

^a*dl*=(6,133).

^b*dl*=(2,114).

^c*dl*=(1,114).

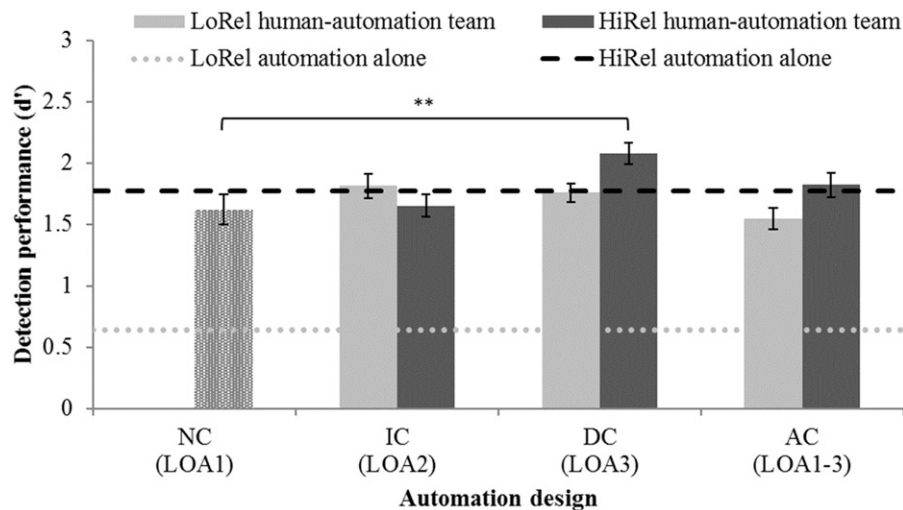


Figure 2. Mean detection performance (d') of participants as a function of automation mode (NC: no cue, IC: indirect cue; DC: direct cue, AC: adaptable cueing) and system reliability (LoRel: low reliability; HiRel: high reliability). $**p < 0.01$. The dashed line indicates detection performance (d') of the automation in high-reliability conditions, whereas the dotted line indicates detection performance of the automation in low-reliability conditions. The error bars denote standard errors.

$t(38) = 3.392$, $p = 0.003$, $r(38) = 0.482$. All other comparisons were not significant, t 's < 1.500 , $p > 0.10$.

A two-way ANOVA was carried out, using detection performance (d') of human-automation team as dependent variable with system reliability (LoRel, HiRel) and automation mode (IC, DC, and AC) as between-participant factors. It showed a main effect of automation mode (Table 2). *Post hoc* analysis revealed that participants in DC mode showed significantly better detection performance than participants in AC, $t(78) = 2.587$, $p = 0.027$, $r(78) = 0.281$, and marginally better detection performance than participants in IC mode, $t(78) = 2.056$, $p = 0.063$, $r(78) = 0.227$ (Table 3).

There was no difference between IC and AC, $t(78) = 0.531$, $p = 0.500$, $r(78) = 0.060$. Furthermore, participants were marginally better when system reliability was high than when it was low (Tables 2 and 3). Finally, the interaction between automation mode and system reliability was significant. Simple effects showed that there was no significant difference between the three automation modes under low system reliability, $F(2,114) = 2.402$, $p = 0.095$, $\eta^2_{\text{partial}} = 0.040$. However, a different pattern emerged under high system reliability, $F(2,114) = 5.623$, $p = 0.005$, $\eta^2_{\text{partial}} = 0.090$. Detection performance for participants in DC mode was significantly better than in the IC

Table 3. Mean scores (and standard deviations) for participants' performance, subjective measures and use of automation of the experimental groups.

Score	NC	Low reliability			High reliability		
		IC	DC	AC	IC	DC	AC
Performance							
Detection (d')	1.62 (0.55)	1.81 (0.45)	1.76 (0.34)	1.55 (0.40)	1.65 (0.41)	2.08 (0.40)	1.82 (0.44)
Response bias (c)	0.18 (0.36)	0.09 (0.29)	0.03 (0.34)	-0.01 (0.28)	0.01 (0.25)	0.03 (0.34)	-0.11 (0.38)
Response time (s)							
Target present	2.22 (0.46)	2.75 (0.44)	1.69 (0.53)	2.17 (0.95)	2.37 (0.45)	1.43 (0.39)	2.01 (0.69)
Target absent	4.28 (1.04)	5.89 (1.73)	3.56 (1.75)	4.33 (2.20)	3.96 (0.93)	3.69 (1.49)	4.10 (1.35)
Use of automation							
Compliance (%)	-	59.77 (11.19)	65.47 (8.09)	66.43 (7.83)	73.59 (9.25)	79.84 (7.00)	81.89 (8.97)
Reliance (%)	-	64.18 (6.57)	66.80 (10.32)	61.76 (13.18)	73.34 (11.76)	74.53 (8.84)	74.06 (13.90)
Subjective measures							
Trust	-	2.77 (0.83)	3.47 (0.96)	3.27 (0.74)	3.79 (1.04)	3.76 (0.75)	3.85 (1.07)
ARE (%)	-	-11.25 (14.22)	-0.39 (9.62)	-9.35 (14.32)	-17.40 (14.22)	-15.00 (14.50)	-24.43 (5.14)
Self-confidence	5.60 (2.09)	6.30 (1.98)	5.15 (1.79)	5.70 (1.38)	5.65 (1.63)	5.85 (1.69)	5.45 (1.67)
Subjective workload	9.65 (2.48)	9.98 (2.09)	8.75 (2.08)	9.63 (1.60)	9.83 (2.76)	9.02 (2.69)	9.29 (2.73)

NC: no cue; IC: indirect cue; DC: direct cue; AC: adaptable cueing; ARE: accuracy of reliability estimate.

mode, $t(38) = 3.332$, $p = 0.002$, $r(38) = 0.476$, and marginally better than in AC mode, $t(38) = 2.007$, $p = 0.071$, $r(38) = 0.310$ (Table 3). No significant difference was observed between IC and AC modes, $t(38) = 1.323$, $p = 0.282$, $r(38) = 0.210$.

3.1.2. Response bias

Overall, participants were rather unbiased in their answers (overall $c = 0.02$, $SD = 0.33$). The main effect of automation mode was not significant (Tables 2 and 3), showing no difference between automation modes and the no cue condition.

The two-way ANOVA with automation mode and system reliability showed no main effect of automation mode. Moreover, there was no significant effect of system reliability (Table 3). Finally, there was no interaction between automation mode and system reliability (Table 2 for F -values).

3.1.3. Target present response times

The one-way ANOVA revealed a main effect of automation mode on response time when the target was present (Figure 3 and Table 3). *Post hoc* analyses showed that response time to the presence of a target was slower in NC than in both DC modes, $t_{NC-HiRel} DC(38) = 4.261$, $p < 0.001$, $r(38) = 0.569$ and, $t_{NC-LoRel} DC(38) = 2.829$, $p = 0.019$, $r(38) = 0.417$. Furthermore, faster response times were observed for participants in NC than for participants working under the IC mode with low reliability, $t(38) = -2.848$, $p = 0.018$, $r = 0.419$. All other comparisons involving NC were not significant, all t 's < 1.140 , $p > 0.10$.

The two-way ANOVA revealed a significant effect of automation mode for response time when a target was present (Table 2). Detailed analyses revealed that participants detected the target in the DC mode faster

than in the AC and IC modes, $t(78) = 3.484$, $p < 0.001$, $r(78) = 0.367$ and $t(78) = 7.382$, $p < 0.001$, $r(78) = 0.641$ (Table 3 and Figure 3). Furthermore, participants were faster to detect the target presence under high than low system reliability (Tables 2 and 3). No interaction was observed.

3.1.4. Target absent response times

The one-way ANOVA for the response time when the target was absent was also significant, $F(6,133) = 4.968$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.183$ (Figure 3 and Table 3). The pairwise comparisons including the NC mode showed that participants in NC were significantly faster to indicate the absence of a target than participant working under IC mode with low reliability, $t(38) = 3.280$, $p = 0.005$, $r(38) = 0.470$. All other comparisons did not reach the significance level, all t 's < 1.500 , $p > 0.10$.

The two-way ANOVA showed a main effect of automation on response time when no target was in the luggage (Table 2). Multiple comparisons revealed that only participants in IC condition were significantly slower to respond than participants in DC mode, $t(78) = 3.589$, $p = 0.001$, $r(78) = 0.376$, and marginally slower than in AC mode, $t(78) = 1.956$, $p = 0.079$, $r(78) = 0.216$ (Table 3). No significant difference was found between participants of DC and AC modes, $t(78) = 1.633$, $p = 0.158$, $r(78) = 0.182$. Moreover, participants were faster to answer when the system reliability was high than low (Tables 2 and 3). Finally, the interaction between automation design and system reliability was significant (Table 2). Further analyses revealed that high system reliability was only beneficial in IC mode, where participants were faster to answer under high than low system reliability, $t(38) = 3.755$, $p < 0.001$, $r(38) = 0.520$ (Table 3

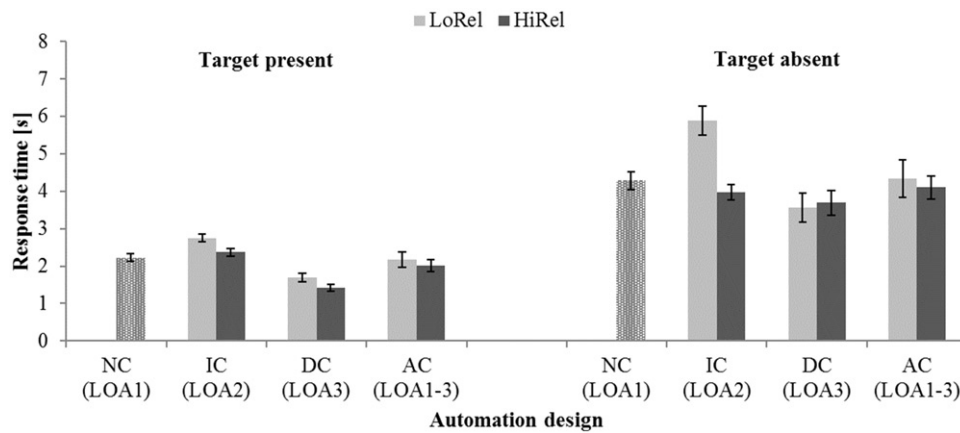


Figure 3. Mean response time as a function of automation mode (NC: no cue; IC: indirect cue; DC: direct cue; AC: adaptable cueing) and system reliability (LoRel: low reliability; HiRel: high reliability) for both target present and target absent trials. The error bars denote standard errors. LOA: level of automation.

and Figure 3). However, no such difference was observed in the two other modes, all t 's < 1.

3.2. Use of automation

3.2.1. Compliance

Compliance was affected by automation mode (Table 2). Multiple comparisons showed that participants in IC mode were less compliant than participants in DC and AC modes, $t(78) = 3.738$, $p < 0.001$, $r(78) = 0.390$ and $t(78) = 3.027$, $p = 0.005$, $r(78) = 0.324$, respectively (Table 3). Participants in AC and DC modes did not differ significantly, $t(78) = 0.751$, $p = 0.500$, $r(78) = 0.085$. As expected, participants were more compliant when system reliability was high than when it was low (Tables 2 and 3). However, no interaction was observed, $F(2,114) = 0.085$, $p = 0.918$, $\eta^2_{\text{partial}} = 0.002$.

3.2.2. Reliance

Automation mode did not influence participants' reliance on automation (Table 2). As expected, high system reliability induced more reliance than low system reliability, $F(1,114) = 23.105$, $p < 0.001$, $\eta^2_{\text{partial}} = 0.171$ (Table 3). No interaction was observed, $F(2,114) = 0.437$, $p = 0.647$, $\eta^2_{\text{partial}} = 0.008$.

3.2.3. Preferred LOA

In AC modes, most participants mainly opted for LOA-3 (Figure 4). There was no significant difference between system reliability levels, $\chi^2(2) = 2.939$, $p = 0.230$.

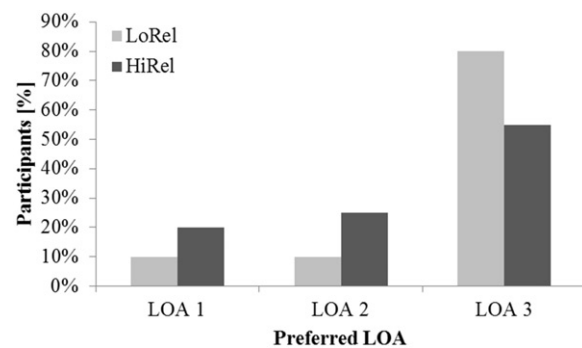


Figure 4. Preferred LOA of participants working under adaptable cueing condition (%) as a function of system reliability (LoRel: low reliability; HiRel: high reliability). Preferred LOA corresponds to the automation level used most of the time during the task.

3.2.4. Frequency of LOA changes

LOA stability was very high. Participants overall switched LOAs about 0.09 times (SD = 0.22) during the testing session. No difference was observed between system reliability conditions, $t(38) = 0.806$, $p = .874$, $r(38) = 0.130$.

3.3. Subjective measures

3.3.1. Trust

Since there was no automation involved in the NC mode, trust was not measured in this condition. The two-way ANOVA revealed no main effect of automation mode on trust. As expected, higher trust ratings were observed under high- than low-reliability condition (Table 3). Finally, there was no interaction between automation mode and system reliability. All F -values are displayed in Table 2.

3.3.2. Accuracy of reliability estimate

Since participants in the NC condition did not experience automation, no reliability estimate was computed in this condition. Overall, participants tended to underestimate system reliability by about 13%. The two-way ANOVA showed a main effect of automation mode. Participants in AC mode underestimated system reliability more strongly than participants in DC mode, $t(78) = 2.92$, $p = 0.006$, $r(78) = 0.314$ (Table 3). Participants in IC mode did not differ from the two other groups, $t's < 2.14$, $p > 0.05$. Furthermore, participants underestimated system reliability more strongly when it was low than when it was high (Table 3). There was no significant interaction. All F -values are reported in Table 2.

3.3.3. Self-confidence

Overall, ratings were made in about the middle of the scale ($M = 5.67$, $SD = 1.75$). The one-way ANOVA revealed that the main effect of automation mode was not significant (Tables 2 and 3). Non-significant results were also observed for the two-way ANOVA (Table 2).

3.3.4. Subjective workload

The overall workload level was at mid-scale ($M = 9.45$, $SD = 2.37$). The one-way ANOVA showed no significant effect of automation condition was observed (Tables 2 and 3). Non-significant results were also observed for the two-way ANOVA (Table 2).

4. Discussion

The goal of the present study was to investigate the effects of different automation modes (indirect cues, direct cues or adaptable cueing) under different reliability levels (low versus high) in an X-ray luggage screening task. With direct cues under high system reliability, participants showed higher detection performance compared to all other automation modes. Moreover, participants detected targets faster when direct cues were available compared to all other automation modes. Adaptable automation did not provide additional benefits compared to static automation. As expected, under high automation reliability, participants achieved higher detection performance and faster response times. Participants were also more inclined to follow automation recommendations (higher compliance and reliance), and expressed higher levels of trust. Finally, participants systematically underestimated the actual reliability of automation.

These findings confirm previous research that showed a larger benefit for direct cues than for indirect ones (Goh, Wiegmann, and Madhavan 2005) but these benefits only seem to take effect under high system reliability. This interaction between automation mode and system reliability is an important result since it confirmed the limited utility of low-reliability automation. Generally, our findings can be interpreted in the framework of Drury's two-component inspection model (TCM), which proposes that an inspection task contains both search and decision components (Drury 1975). Studies on the applicability of TCM for X-ray image inspection have provided converging evidence for a search and decision component (Koller, Drury, and Schwaninger, 2009; Wales, Anderson, Jones, Schwaninger, and Horne, 2009). Within this framework, the benefit of direct cues for detection performance and target present response times observed in this study can be interpreted as direct cues guiding attention to target items (and hence reduce search time). This may have facilitated participants' compliance with automation recommendations when it indicated the presence of a target. In contrast, when automation indicated the absence of a target, there was no such restriction of the search area, which may explain similar levels of reliance in DC and IC modes.

The lack of precise information provided by the cue in IC mode is aggravated by low system reliability, especially in target-absent trials. In the current study, participants in the IC mode working with a low-reliability system needed more time to report the absence of a target than in all other conditions. This is an interesting finding, which may be related to the general difficulty of IC to provide adequate support. In target-absent trials, this is particularly difficult because of the long target search times associated with IC. In the framework of the TCM, the combination of low reliability and rather imprecise indications of target locations offered by IC may have raised the stopping threshold for the search component, resulting in longer search times, but it did not influence the decision process. As trust ratings were lower in low-reliability conditions, this may have also influenced response times. For example, Yeh and Wickens (2001) have argued that trust levels determine when operators finish searching the display for a target. Overall, this observation provides further evidence for higher efficiency of DC mode over IC mode in visual inspection.

An important research question refers to the potential of adaptable automation in luggage screening, with this study being the first to address this question. Surprisingly, a clear benefit of adaptable automation

compared to static automation modes was not found. At first sight, this seems to contradict findings from multitasking environments, which demonstrated benefits of adaptable automation for performance (Parasuraman et al. 1993; Sauer, Kao, and Wastell 2012). However, considering that with high system reliability, direct cues (DC mode) are much more beneficial than indirect cues (IC mode), participants would have benefited most from adaptable automation (AC) only if they had primarily chosen the DC mode. In the present study, this was not the case. About a third of participants in the AC condition did not predominantly use the DC mode (i.e. DC was used more than 50% of the time by 67.5% of participants while no cue and IC were predominantly used by 15 and 17.5% of the participants, respectively). The reduced usage frequency of the most powerful LOA in adaptable cueing may be related to general difficulties of estimating levels of system reliability accurately. Furthermore, in some instances operators may have opted for manual control (i.e. no cue) due to a general need for increased latitude in system operation, which is also advocated by models in work psychology (Karasek and Theorell 1990). Generally, our results indicate that it might be wrong to assume a general benefit of adaptable automation over static automation observed in multitasking environments without considering the benefits of certain static automation modes in a specific single-task environment (like DC for luggage screening in our study).

When adaptable cueing was available, it was interesting to note that participants did not make frequent changes between LOAs and mainly worked in their preferred LOA. This suggests that the main advantage of adaptable automation is that it allows each participant to select the LOA he or she is most comfortable with. In contrast, the second potential advantage of adaptable automation was not much made use of. This is to adapt LOA according to changing operational needs during the course of a working session (e.g. to cope with increasing fatigue). These observations are consistent with the findings of previous research from domains such as process control, in which a strong preference of one LOA was also found (Chavaillaz, Wastell, and Sauer 2016b; Sauer, Chavaillaz, and Wastell 2017). This suggests that operators' preference to keep working with the same LOA may be found in more than one work domain.

The present study also allows us to examine the question of the consequences of using three principal options of work design, which are fundamentally different from each other: *human alone* (here: no cue),

machine alone (here represented by performance lines in Figure 2) and combined *human-machine team* (here: three automation modes). They can be compared to each other with regard to their detection performance. The data in Figure 2 indicate that human alone and human-machine team outperformed machine alone on the detection task in the low-reliability condition. Furthermore, Figure 2 shows that under high system reliability, performance of the human-automation team is better than machine alone and human alone in the DC mode. This confirms the advantages of direct cues for verifying the validity of automation suggestions but also shows that it is not a simple matter to achieve benefits of combined human-automation team performance compared to alternative work designs (i.e. machine alone or human alone). This may also raise questions about further alternatives in work design in the luggage screening environment which could be examined in future studies. For example, this may involve a loosely coupled human-automation team, in which the automation decision is only shown to the human after he or she has taken a first decision. However, it would still allow the human to revise this first decision in the light of the automation's decision before the ultimate decision is taken by the human.

Perceived reliability (as a factor determining the way automation is used) showed overall that participants underestimated system reliability by about 13%. Previous research also found that operators tended to underestimate system reliability (Wiegmann, Rich, and Zhang 2001; van Dongen and van Maanen 2013). It is interesting to note that automation mode seems to have influenced the magnitude of the underestimation. Participants assessed automation reliability more accurately in DC (deviation of about 8%) than in the two other automation modes (about 16%). This observation suggests that the quality of feedback on the validity of automation recommendations influences the magnitude of the underestimation of system reliability. In DC mode, participants can assess more easily whether the automation recommendation is correct (i.e. is the cued object really prohibited?) than in IC (where the whole piece of luggage is cued). In line with Wiegmann, Rich, and Zhang (2001), we observed a better estimate of system reliability under low than high reliability. This might be explained by the frequency of automation failures. Under low reliability, participants are often confronted with automation failures, which may have made them less salient. In contrast, under high reliability the fewer occurrences of automation failures made them more conspicuous.

This might have biased the ratings of system reliability. Overall, the findings suggest that perceived reliability is not only affected by actual system reliability but also by the capabilities of an automatic system.

There are several implications of the present study for the design of automatic support systems at airport security checkpoints. First, pointing out the exact target location (DC mode) appears to be the most powerful form of support (see also Goh, Wiegmann, and Madhavan 2005) because it guides attention to the target location and therefore accelerates the search process (i.e. DC supports searching under time pressure). This provides some backing for the current design of automatic target detection devices, which only makes use of direct cues (Wells and Bradley 2012). Second, despite the benefits of adaptable automation observed in multitasking environments (Kidwell et al. 2012; Sauer, Kao, and Wastell 2012; Sauer and Chavailleaz 2017), the present work did not show evidence for similar advantages over powerful static automation in the form of direct cueing. This may be due to the ease with which automation suggestions can be verified. In luggage inspection tasks, automation suggestions can be directly checked with the cued object. In contrast, verifying automation recommendations in more complex tasks (e.g. in process control; Sauer, Kao, and Wastell 2012) usually requires several checks to be made. However, this advantage may decrease, or even disappear, if miscues occur (i.e. a non-target object is cued in a target-present image). In such a case, the cued location remains the first location on the image to attract attention but loses its delimiting function of the search area. This may result in an increase in search time compared to NC and IC modes. More research is needed to determine whether miscues can effectively remove the delimiting function of the cue and, if this was possible, how many miscues are required to produce such an effect. Furthermore, since operators might not always use the most suitable LOA to complete their tasks (a third of the participants in the current study did not do so), further research should be conducted before the use of adaptable automation in target detection tasks and devices can be recommended. Given the considerable differences between work domains due to differing task requirements (e.g. process control and luggage screening), there is a need to carry out domain-specific research so that the resulting domain-specific findings will allow us to determine to what extent they are transferable across different work domains.

Note

1. In the adaptable automation condition, participants could change between LOA1-3 anytime. For each trial, the LOA selected last was used to determine the preferred LOA.

Acknowledgements

Thanks are also due to Debora de Felice, Malina Gruener, Nadine Weber, Alexandre Cudré, and Beat Vollenwyder for their help with data collection.

Funding

This work was supported by the Swiss National Science Foundation (SNSF) under [Grant No 100014_134566].

References

- Baeriswyl, S., A. Krause, and A. Schwaninger. 2016. "Emotional Exhaustion and Job Satisfaction in Airport Security Officers – Work-Family Conflict as Mediator in the Job Demands-Resources Model." *Frontiers in Psychology* 7: 663. doi:10.3389/fpsyg.2016.00663
- Bailey, N. R., and M. W. Scerbo. 2007. "Automation-Induced Complacency for Monitoring Highly Reliable Systems: The Role of Task Complexity, System Experience, and Operator Trust." *Theoretical Issues in Ergonomics Science* 8 (4): 321–348. doi:10.1080/14639220500535301.
- Brainard, D. H. 1997. "The Psychophysics Toolbox." *Spatial Vision* 10: 433–436. doi:10.1163/156856897X00357
- Chavailleaz, A., D. Wastell, and J. Sauer. 2016a. "Effects of Extended Lay-off Periods on Performance and Operator Trust Under Adaptable Automation." *Applied Ergonomics* 53(Pt A): 241–251. doi:10.1016/j.apergo.2015.10.006
- Chavailleaz, A., D. Wastell, and J. Sauer. 2016b. "System Reliability, Performance and Trust in Adaptable Automation." *Applied Ergonomics* 52: 333–342. doi:10.1016/j.apergo.2015.07.012
- Drury, C. G. 1975. "Inspection of Sheet Materials – Model and Data." *Human Factors* 17 (3): 257–265. doi:10.1177/001872087501700305.
- Drury, C. G. 1992. "Inspection Performance." In *Handbook of Industrial Engineering*, edited by Gavriel Salvendy. 2nd ed., 2282–2314. New York: Wiley.
- Drury, C. G. 2001. "Human Factors and Automation in Test and Inspection." In *Handbook of Industrial Engineering: Technology and Operations Management*, edited by Gavriel Salvendy. 3rd ed., 1887–1920. New York: John Wiley.
- Dzindolet, M. T., S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. 2003. "The Role of Trust in Automation Reliance." *International Journal of Human-Computer Studies* 58 (6): 697–718. doi:10.1016/S1071-5819(03)00038-7.
- Goh, J., D. A. Wiegmann, and P. Madhavan. 2005. "Effects of Automation Failure in a Luggage Screening Task: A Comparison Between Direct And Indirect Cueing." In *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society*. Vol. 49, 492–496. Orlando, FL: SAGE Publications. doi:10.1177/154193120504900359

- Green, D. M., and J. A. Swets. 1966. *Signal Detection Theory and Psychophysics*. New-York, NY: Wiley.
- Hardmeier, D., F. Hofer, and A. Schwaninger. 2005. "The X-ray object recognition test (X-ray ORT)—a reliable and valid instrument for measuring visual abilities needed in X-ray screening." *Paper presented at the Proceedings of the 39th IEEE Carnahan Conference on Security Technology*, 189–92. Las Palmas, Spain. doi:10.1109/CCST.2005.1594876.
- Harris, D. H. 2002. "How to Really Improve Airport Security." *Ergonomics in Design: The Quarterly of Human Factors Applications* 10 (1): 17–22. doi:10.1177/106480460201000104.
- Hart, S.G., and L.E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In *Human Mental Workload*. Vol. 52 of *Advances in Psychology*, edited by Peter A. Hancock and Najmedin Meshkati, 139–383. Amsterdam: North-Holland.
- Inagaki, T. 2003. "Adaptive Automation: Sharing and Trading of Control." In *Handbook of Cognitive Task Design. Series: Human Factors and Ergonomics*, edited by Erik Hollnagel, 221–245.. Mahwah: Lawrence Erlbaum Publishers.
- Jian, J.-Y., A. M. Bisantz, and C. G. Drury. 2000. "Foundations for an Empirically Determined Scale of Trust in Automated Systems." *International Journal of Cognitive Ergonomics* 4 (1): 53–71. doi:10.1207/S15327566IJCE0401_04.
- Kaber, D. B., and J. M. Riley. 1999. "Adaptive Automation of a Dynamic Control Task Based on Secondary Task Workload Measurement." *International Journal of Cognitive Ergonomics* 3 (3): 169–187. doi:10.1207/s15327566ijce0303_1
- Karasek, R., and T. Theorell. 1990. *Healthy Work: Stress Productivity and the Reconstruction of Working Life*. 1st ed. New York: Basic Books.
- Kidwell, B., G. L. Calhoun, H. A. Ruff, and R. Parasuraman. 2012. "Adaptable and Adaptive Automation for Supervisory Control of Multiple Autonomous Vehicles." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56 (1): 428–432. doi:10.1177/1071181312561096.
- Kleiner, M., D. Brainard, D. Pelli, A. Ingling, R. Murray, and C. Broussard. 2007. "What's New in Psychtoolbox-3." *Perception* 36 (14): 1.
- Koller, S. M., C. G. Drury, and A. Schwaninger. 2009. "Change of search time and non-search time in X-ray baggage screening due to training." *Ergonomics* 52 (6): 644–656. doi:10.1080/0014013080526.
- Lee, J. D., and N. Moray. 1992. "Trust, Control Strategies and Allocation of Function in Human-Machine Systems." *Ergonomics* 35 (10): 1243–1270. doi:10.1080/00140139208967392.
- Lee, J. D., and F. Sanquist. 1996. "Maritime Automation." In *Automation and Human Performance: Theory and Applications. Series: Human Factors in Transportation*, edited by Raja Parasuraman and Mustapha Mouloua, 365–384. Mahwah, NJ: Erlbaum.
- Ma, R., and D. B. Kaber. 2007. "Effects of In-Vehicle Navigation Assistance and Performance on Driver Trust and Vehicle Control." *International Journal of Industrial Ergonomics* 37 (8): 665–673. doi:10.1016/j.ergon.2007.04.005.
- Macmillan, N. A., and C. D. Creelman. 2005. *Detection Theory: A User's Guide*. 1st ed. Mahwah, NJ: Psychology Press.
- McCarley, J. S., A. F. Kramer, C. D. Wickens, E. D. Vidoni, and W. R. Boot. 2004. "Visual Skills in Airport-Security Screening." *Psychological Science* 15 (5): 302–306. doi:10.1111/j.0956-7976.2004.00673.x.
- McCarley, J. S., D. A. Wiegmann, C. D. Wickens, and A. F. Kramer. 2003. "Effects of Age on Utilization and Perceived Reliability of an Automated Decision-Making Aid for Luggage Screening." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 47 (3): 340–343. doi:10.1177/154193120304700319.
- Mery, D., G. Mondragon, V. Rizzo, and I. Zuccar. 2013. "Detection of Regular Objects in Baggage Using Multiple X-ray Views." *Insight - Non-Destructive Testing and Condition Monitoring* 55 (1): 16–20. doi:10.1784/insi.2012.55.1.16.
- Meyer, J. 2001. "Effects of Warning Validity and Proximity on Responses to Warnings." *Human Factors* 43 (4): 563–572. doi:10.1518/001872001775870395.
- Michel, S., N. Hattenschwiler, M. Kuhn, N. Strebel, and A. Schwaninger. 2014. "A Multi-Method Approach Towards Identifying Situational Factors and their Relevance for X-ray Screening." Paper presented at the International Carnahan Conference on Security Technology (ICCST). October 13-16. Rome, Italy. doi:10.1109/CCST.2014.6987001.
- Onnasch, L. 2015. "Crossing the Boundaries of Automation—Function Allocation and Reliability." *International Journal of Human-Computer Studies* 76: 12–21. doi:10.1016/j.ijhcs.2014.12.004
- Parasuraman, R., R. Molloy, and I. L. Singh. 1993. "Performance Consequences of Automation-Induced 'Complacency'." *The International Journal of Aviation Psychology* 3 (1): 1–23. doi:10.1207/s15327108ijap0301_1.
- Parasuraman, R., M. Mouloua, R. Molloy, and B. Hilburn. 1993. "Adaptive Function Allocation Reduces Performance Costs of Static Automation." In *7th International Symposium on Aviation Psychology*, edited by Richard S. Jensen and David Neumeister, 178–185. Columbus: Ohio State University.
- Pelli, D. G. 1997. "The VideoToolbox Software for Visual Psychophysics: Transforming Numbers into Movies." *Spatial vision* 10 (4): 437–442. doi:10.1163/156856897X00366
- Rice, S., and J. S. McCarley. 2011. "Effects of Response Bias and Judgment Framing on Operator Use of an Automated Aid in a Target Detection Task." *Journal of Experimental Psychology: Applied* 17 (4): 320–331. doi:10.1037/a0024243.
- Roomi, M. M., and R. Rajashankari. 2012. "Detection of Concealed Weapons in X-ray Images Using Fuzzy K-NN." *International Journal of Computer Science, Engineering and Information Technology* 2 (2): 187–196. doi:10.5121/ijcseit.2012.2216.
- Rovira, E., K. McGarry, and R. Parasuraman. 2007. "Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task." *Human Factors* 49 (1): 76–87. doi:10.1518/001872007779598082.
- Sanchez, J., A. D. Fisk, and W. A. Rogers. 2004. "Reliability and Age-Related Effects on Trust and Reliance of a Decision Support Aid." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 48 (3): 586–589. doi:10.1177/154193120404800366.

- Sauer, J., and A. Chavaillaz. 2017. "How Operators Make Use of Wide-Choice Adaptable Automation: Observations from a Series of Experimental Studies." *Theoretical Issues in Ergonomics Science*, 1–21. doi:10.1080/1463922X.2017.1297866.
- Sauer, J., A. Chavaillaz, and D. Wastell. 2017. "Experience of Automation Failures in Training: Effects on Trust, Automation Bias, Complacency and Performance." *Ergonomics* 73 (9): 1–14. doi:10.1080/00140139.2015.1094577.
- Sauer, J., C.-S. Kao, and D. Wastell. 2012. "A comparison of adaptive and adaptable automation under different levels of environmental stress." *Ergonomics* 55 (8): 840–853. doi:10.1080/00140139.2012.676673.
- Scerbo, M. W. 2006. "Dynamic Function Allocation." In *International Encyclopedia of Ergonomics and Human Factors*, edited by Waldemar Karwowski. 2nd ed., 1080–82. Boca Raton, FL: Taylor & Francis.
- Schwaninger, A., A. Bolfig, T. Halbherr, S. Helman, A. Belyavin, and L. Hay. 2008. "The impact of image based factors and training on threat detection performance in X-ray screening." Paper presented at the Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT, June 1-4. Fairfax, VA, 317-324. doi:10.13140/RG.2.1.1299.3526.
- Schwaninger, A., D. Hardmeier, and F. Hofer. 2005. "Aviation Security Screeners Visual Abilities and Visual Knowledge Measurement." *Aerospace and Electronic Systems Magazine, IEEE* 20 (6): 29–35.
- See, J. E. 2012. "Visual Inspection: a Review of the Literature." Sandia Report (SAND2012-8590). October 2012. Sandia National Laboratories, New Mexico.
- Sheridan, T. B., and W. L. Verplank. 1978. *Human and Computer Control of Undersea Teleoperators. Technical Report*. Cambridge, MA: MIT Man-Machine Systems Laboratory.
- Sterchi, Y., and A. Schwaninger. 2015. "A First Simulation on opTImizing EDS for Cabin Baggage Screening Regarding Throughput." Paper presented at the International Carnahan Conference on Security Technology (ICCST), September 21–24, Taipei, Taiwan, 55–60.
- van Dongen, K., and P.-P. van Maanen. 2013. "A Framework for Explaining Reliance on Decision Aids." *International Journal of Human-Computer Studies* 71 (4): 410–424. doi:10.1016/j.ijhcs.2012.10.018.
- Warm, J. S., R. Parasuraman, and G. Matthews. 2008. "Vigilance Requires Hard Mental Work and is Stressful." *Human Factors* 50 (3): 433–441. doi:10.1518/001872008X312152.
- Wales, A. W. J., C. Anderson, K. L. Jones, A. Schwaninger, and J. A. Horne. 2009. "Evaluating the two-component inspection model in a simplified luggage search task." *Behavior Research Methods* 41 (3): 937–943. doi:10.3758/BRM.41.3.937.
- Wells, K., and D. A. Bradley. 2012. "A Review of X-ray Explosives Detection Techniques for Checked Baggage." *Applied Radiation and Isotopes* 70 (8): 1729–1746. doi:10.1016/j.apradiso.2012.01.011.
- Wickens, C. D., and S. R. Dixon. 2007. "The Benefits of Imperfect Diagnostic Automation: A synThesis of the Literature." *Theoretical Issues in Ergonomics Science* 8 (3): 201–212. doi:10.1080/14639220500370105.
- Wiegmann, D., J. S. McCarley, A. F. Kramer, and C. D. Wickens. 2006. "Age and Automation Interact to Influence Performance of a Simulated Luggage Screening Task." *Aviation, Space, and Environmental Medicine* 77 (8): 825–831.
- Wiegmann, D. A. 2002. "Agreeing with Automated Diagnostic Aids: A study of Users' Concurrence Strategies." *Human Factors* 44 (1): 44–50. doi:10.1518/0018720024494847.
- Wiegmann, D. A., A. Rich, and H. Zhang. 2001. "Automated Diagnostic Aids: The Effects of Aid Reliability on Users' Trust and Reliance." *Theoretical Issues in Ergonomics Science* 2 (4): 352–367. doi:10.1080/14639220110110306.
- Wolfe, J. M., D. N. Brunelli, J. Rubinstein, and T. S. Horowitz. 2013. "Prevalence Effects in Newly Trained Airport Checkpoint Screeners: Trained Observers Miss Rare Targets, Too." *Journal of Vision* 13 (3): 1–9. doi:10.1167/13.3.33.
- Wolfe, J. M., T. S. Horowitz, M. J. van Wert, N. M. Kenner, S. S. Place, and N. Kibbi. 2007. "Low target prevalence is a stubborn source of errors in visual search tasks." *Journal of Experimental Psychology: General* 136 (4): 623–638. doi:10.1037/0096-3445.136.4.623.
- Yeh, M., and C. D. Wickens. 2001. "Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration." *Human Factors* 43 (3): 355–365. doi:10.1518/001872001775898269.