Time on Task Effects during Interactive Visual Search

Hayward J. Godwin[1], Simon P. Liversedge[2], Natalie Mestry[3], Haden Dewis[1] and Nick Donnelly[4]

[1]University of Southampton

[2]University of Central Lancashire

[3]Bournemouth University

[4]Liverpool Hope University

Author Note

Correspondence regarding this article should be addressed to Hayward J. Godwin, University of Southampton, School of Psychology, Highfield, Southampton, Hampshire, SO17 1BJ. Tel: +44(0)2380 595078; Email: hayward.godwin@soton.ac.uk.
Data and experimental code for the experiments are available to access online via this web address: https://osf.io/bdxyf/?view_only=4ca5348ed7f44ad1887f31a972feb1af

CRediT Roles

Godwin: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft preparation. Liversedge: Conceptualization, funding acquisition, methodology, writing – review & editing. Mestry: Conceptualization, funding acquisition, methodology, writing – review & editing. Dewis: Data curation, formal analysis, investigation, resources, software, validation, writing – review & editing. Donnelly: Conceptualization, funding acquisition, methodology, writing – review & editing.

Abstract

There is a major shift taking place in airports across the globe, changing from 2D dual-view X-ray screening to 3D CT screening. 3D CT screening is believed to improve target detection since it enables screeners to interact with images of passenger baggage (i.e., rotating and zooming into the displays). The change in screening technology is moving what was once a purely visual search task to an interactive search task. Here, we conducted two experiments with a large sample size during February of 2023 (695 participants) to examine (1) changes in search performance between a simulated dual-view and simulated interactive search task and (2) the effects of time on task upon performance. Consistent with past research, we found that interactive search, when compared with dual-view search, produced higher response accuracy rates coupled with increased reaction times. However, while we found effects of time on task (RTs reduced, and participants became more likely to respond 'absent' as the experiments progressed) there was no evidence that these effects differed across simulated dual-view and simulated interactive searches. The results are discussed in relation to benefits of interactive search for supporting target detection by airport screeners.

Keywords: interactive search, time on task, visual search

Public Significance Statement

We examined the behavior and performance of participants searching for targets in dual-view and interactive displays that resemble those used in airport baggage screening. The results showed participants to be slower but more accurate when searching using interactive displays. The results provided no evidence that effects of time on task differed across simulated dual-view and simulated interactive search displays. The findings have implications for enhancing the detection of threats in baggage screening in the real world.

Time on Task Effects during Interactive Visual Search

During a visual search task, an observer attempts to determine whether a target object is present in the environment around them. Visual searches range from the everyday, such as searching for our favorite cereal in the supermarket, to the societally important, such as when forensic experts are searching for evidence in relation to a crime. Most visual search experiments have been conducted with a focus on theory rather than practical application. However, there are several substantial strands of research that have directly examined real-world visual search. These include radiological searches (e.g., see Nodine & Kundel, 1987), airport baggage screening (for a review, see Donnelly et al., 2019), military searches (e.g., Godwin et al., 2015; Riggs et al., 2018), and forensic searches (e.g., Riggs et al., 2017). It is this second strand of practical applications of visual search – airport baggage screening – that is the focus and foundation of our work here.

Background: Use-inspired Approach in the Current Experiments

The current set of experiments take a 'use-inspired' research approach (Wolfe, 2016), drawing upon airport baggage screening as a real-world task, to guide the research questions being addressed. There is a substantial technological shift currently ongoing in airport screening procedures across the world. In the past, airport screening has involved the use of 2D X-ray images to enable screening officers to examine and search baggage. Though still valuable, X-ray technology is now being phased out and replaced with Computed Tomography (CT) imaging.

The move from X-ray to CT screening technologies has resulted in changes to screening procedures. X-ray screening generally involves the presentation of two different ‘views’ to an airport screener: a *top-down* view (i.e., from above) and a *side* view, and as such is referred to as ‘dual-view X-ray screening’. The new CT screening systems take a different approach. The new CT screening systems present a virtual 3D image of cabin baggage that can be rotated, manipulated, and zoomed into by a screener. The primary difference, then, between X-ray screening and CT screening is the ability to *interact* with the images of cabin baggage. CT screening extends the two views offered in dual-view X-ray screening to an infinite number of different viewpoints under the control of the screener.

The assumption is that providing the opportunity to interact with the CT images will enhance screener performance, primarily in terms of their response accuracy. Improved accuracy is presumed to result from interactivity which allows screeners to rotate the baggage and view it from multiple different perspectives, potentially including viewpoints from which target objects are easier to detect. In addition, screeners may zoom into specific objects for closer inspection. Such interactivity should also permit screeners to view objects from a perspective in which visual clutter and overlap between objects is reduced. Clutter and overlap is known to impair detection and performance in search (Bravo & Farid, 2004a, 2004b, 2006; Godwin et al., 2017, 2020) and to increase the probability of participants simply ‘not knowing’ how to respond (Godwin & Hout, 2023). Overall, then, CT screening should enable searchers to resolve situations wherein the initial viewpoint presented to them is difficult to resolve or reach a decision for. As such, it should therefore enhance response accuracy rates.

The differences between X-ray screening and CT screening are not solely limited to the possibility of interacting with the CT images. Indeed, some differences have the potential to worsen performance in CT screening relative to X-ray screening. First, some authors have noted that image quality can be lower in CT than X-ray screening (e.g., Hättenschwiler et al., 2019). Second, when CT screening is in operation, laptops and liquids are permitted to be retained inside the bags that are being screened. This is in contrast to X-ray screening, where such items must be removed from bags prior to screening. Thus, images viewed in CT screening may be more complex and more cluttered than in X-ray screening. However, this effect of increased clutter might be offset by the fact that screeners are required to search for fewer target items since, liquids and laptops are no longer target items in CT screening. As noted earlier, past research has shown that a reduction in the number of targets being searched for can benefit both response accuracy and response times (see Donnelly et al., 2019).

Interactive Search in CT Screening

Given the significant number of differences, as well as both the magnitude and nature of those differences between X-ray and CT screening, there are many potentially important theoretical questions that arise. *Interactive search*, a term coined by Sauter et al. (2020), is a relatively new area of research, and given its recent emergence, as yet, no single primary method of investigating this type of search behavior has emerged. Studies investigating interactive search have involved participants searching for specific lights on the floor of a room (Smith et al., 2008, 2010); naïve and highly trained participants searching for items relating to a robbery in a residential house (Riggs et al.,

2018); participants searching for coins hidden within a small patch of grass (Riggs et al., 2017); participants clicking and dragging objects that are hiding others in a search task (Solman et al., 2014); participants searching through trays of Lego bricks for targets (Hout et al., 2022; Sauter et al., 2020); and finally, participants searching through 'slices' of images in radiographic screening (Drew et al., 2013). Whilst interesting and informative, these previous studies bear little resemblance to the interactive search that takes place during airport CT screening.

Indeed, relatively few previous studies have directly compared X-ray and CT screening performance. Hättenschwiler et al. (2018, 2019) examined search performance of trained airport screening personnel looking for Improvised Explosive Devices (IEDs) using a 2D X-ray display, a CT display, or a 3D CT display with on-screen alarms to highlight potential targets. They found that sensitivity for detecting IEDs, as measured by the signal detection theory parameter *d'* (Macmillan & Creelman, 2004), was better in CT displays than X-ray displays. However, this was only true if the CT display was accompanied by an automated on-screen alert system (this involved a box being drawn around potential IED targets). In addition, they found that reaction times (RTs) in CT searches were longer than in X-ray searches.

Using a different approach, Parker et al. (2022) compared search performance in a simulated X-ray and simulated CT environment. In the CT condition for this study, the participants were presented with videos of a simulated piece of baggage rotating. The participants could then rewind or fast forward the video to enable a more fine-grained examination. This study was conducted online and recruited members of the public to take part. In contrast to the previous studies, the results indicated that *d'* was higher in

the simulated CT condition even when no alarm system was implemented. The authors did not report response time analyses.

To summarize, these previous studies comparing simulated X-ray and simulated CT environment have delivered only limited evidence for the assumption that interactive CT leads to better target detection than X-ray screening.

The Present Experiments: Time on Task in Interactive Search

Given the limited amount of research examining interactive search to date, there exists a large range of theoretical and applied issues that can be addressed in this domain. Here, we sought to investigate two related issues as a starting point. First, we sought to determine whether, and how, search performance changed between a simulated X-ray and simulated CT screening task. We did this using a large group of participants and using a range of threat items as targets. This issue is an important one to investigate given that it enables us to establish fundamental differences in performance and behavior in the move from non-interactive to interactive search. Moreover, it enables us to assess whether fundamental aspects of visual search that we know hold in simulated dual-view X-ray search maintain during interactive search.

The second issue that we investigated concerned time on task effects in search. Time on task effects relate to any changes in behavior or performance that emerge as engagement in a task progresses over time. There are many potential sources of time on task effects that can influence performance, and they may operate in competition with one another (Lanthier et al., 2013). Time on task effects can include vigilance decrements (Mackworth, 1948), boredom, fatigue (e.g., Horowitz et al., 2003) mind-

wandering (e.g., see Krimsky et al., 2017), and others, all of which might emerge as time progresses. It seems possible that the differences in motoric interaction between X-ray and CT screening may lead to differences in the impact of time on task. For that reason, we believe that time on task effects are particularly important to focus upon here as they may shed light on key differences between dual-view and interactive displays.

From a theoretical standpoint, the present set of experiments can make a number of important contributions to the existing literature. There is, at present, only one theoretical model of interactive search behavior: the interactive Multiple Decision Model, otherwise referred to as the iMDM (Hout et al., 2022). The iMDM is an extension of the Multiple Decision Model of visual search (Wolfe & Van Wert, 2010). Under the iMDM, searches move from one search area to another. Within the visual search-focused MDM, search proceeds as a series of decisions. For each object in a display, the search system asks the question: ‘is this object a target?’. If the answer is ‘no’, the search then proceeds, whilst is the answer is ‘yes’, the searcher responds ‘present’ and terminates the search. Next, the search system asks the question: ‘has the quitting threshold been reached?’. The quitting threshold for a search is governed by a timer. Once that timer has been reached, the search is terminated. The iMDM expands upon this framework by incorporating a series of searches in different areas of the environment. Each area of an environment is searched until the quitting threshold is reached. Once a searcher ceases their examination of that area, the search system then asks a new question, unique to interactive search: “has the *overall quitting threshold* been reached?”. This relates to a second timer that determines when a

searcher decides that they have searched enough areas in order to be confident that no target is present. If this overall threshold has not been reached, a new area is accessed and a new search of that area begins. If this overall threshold has been reached, the search is terminated.

The iMDM serves as a useful starting point, but it is limited in its explanatory scope. In particular, we note that it has no understanding or mechanism for explaining changes in behavior over extended periods of time. Indeed, it makes no consideration of the effects of time-on-task and can only measure transitions between one search to another in terms of ending one trial and moving onto the next. In that regard, it takes trial index as a measure of time. This is true of other models and accounts of search as well (Chun & Wolfe, 1996; Hulleman & Olivers, 2017; Wolfe, 2021), which consider changes in performance over time in terms of arising between one trial and another, with no understanding of the actual time taken to complete those trials, or the time spent searching overall. Re-casting search in terms of the actual time spent searching, and then examining the time on task therefore offers the potential to provide new insights into search behavior and performance that fall out of scope for current models.

One significant body of research that has used time-on-task to measure performance changes in detail is the study of vigilance effects. In studies of vigilance, observers typically spend long periods of time monitoring displays for rare and subtle-to-detect events (for a critical review and discussion, see Hancock, 2013). Here, researchers have found that increasing time on task results in observers being less likely to detect those events (Mackworth, 1948). However, although these findings do have some relation to those being studied here, it has been noted and shown that

vigilance effects are somewhat distinct from those under examination here. Wolfe (2007) reviewed the literature on vigilance and its relation to studies of low prevalence visual search, and argued that, since standard vigilance experiments typically do not involve visual search, and instead involve monitoring simple stimuli for rare events, that the findings from the study of vigilance do not necessarily translate across to studies of visual search. Moreover, Wolfe (2007) found no evidence that vigilance was influenced by extensive periods of search for rare targets.

Taken together, from a theoretical perspective, the current work therefore offers to further our understanding, and provide a substantial evidence base for, time-on-task effects during interactive search and visual search, and to provide a comparison of how these two forms of search are differentially affected by time spent searching. Moreover, by engaging participants in both an interactive search and a visual search, these experiments aid in providing evidence regarding whether one form of search exhibits higher performance over another.

Studying time on task effects is also important from a practical perspective. At present, there is a limit of searching for 20 minutes in airport screening in Europe and the UK. Current UK policy, which was adopted from the EU, stipulates that: “Persons screening cabin baggage by x-ray or EDS equipment shall normally not spend more than 20 minutes continuously reviewing images” (Section 4.1.2.11 from EU regulation 2015/1998). Several recent studies have examined the consequences of moving beyond this 20-minute time limit. One study recruited airport screeners and asked them to search for threats in a lab-based study for four hours, and reported that, as time progressed, participants became more rapid in their responses, and also more biased to

respond ‘absent’ (Ghylin et al., 2007). Another (Meuter & Lacherez, 2016) examined live screening data from an airport and found a similar pattern of results.

Others have reported less clear evidence of consistent shifts in search performance that progress over time. Buser et al. (2019) used a simulated screening task and asked airport screeners to search for one hour, whilst target prevalence levels were set to high prevalence or low prevalence. Target prevalence refers to the proportion of trials on which a target is presented. Past work has found that, when target prevalence is low, searchers adapt their behavior to quit rapidly and respond ‘absent’ (Wolfe et al., 2007). In doing so, they then tend to miss targets once those rare targets do finally appear. Prevalence rates have been cited as an important factor in airport screening given that real targets appear only rarely (for a review, see Donnelly et al., 2019). And indeed, Buser et al. (2019) found that early in their study, participants adapted to the prevalence levels, with a bias towards responding ‘present’ emerging early on for the high prevalence condition, and a bias towards responding ‘absent’ emerging early on for the low prevalence condition. Similar results with a similar design were also reported in a more recent study (Buser et al., 2020). Elsewhere, Buser et al. (2023) examined Threat Image Projection (TIP) performance during live airport screening. TIP images are targets and prohibited items that are inserted at a low prevalence rate by screening software to serve as a check of screener performance. In this study, the authors examined screening sessions of a duration of up to 60 minutes. Their results echoed much of what has been discussed above, finding that screeners exhibited a reduction in RTs, hit rate, and rejection rate (i.e. proportion of trials wherein searchers correctly respond ‘absent’) as the sessions progressed. In sum, these past

studies generally converge on the notion that, during a search for low prevalence targets, searchers become more likely to respond 'absent' and to respond more rapidly as time progresses. Critically, this issue appears to be quite specific to low prevalence searches only and does not emerge for higher prevalence searches. Given this interest in determining optimum safe limits of screening sessions, and to provide further data on the issue of time on task effects in search, here we examined search performance for 20-minute and 30-minute sessions in both simulated X-ray and simulated CT screening in order to develop the first evidence base for potential time limits of interactive visual searches.

For the current project, we conducted two separate experiments. In both experiments, participants were engaged in either a simulated version of CT screening or a simulated version of X-ray screening (hereafter, for brevity, we refer to these as our *Interactive* and *Dual-view* search conditions), and in both experiments, participants were asked to search for either 20 minutes or 30 minutes. In both experiments, targets were presented a low rate of prevalence (20% of trials). We chose a 20-minute time limit to mirror live screening procedures and extended this to a 30-minute time limit in line with previous studies noted above. Extending to 30 minutes provides an opportunity to rigorously test whether extending currently time limits is feasible, and whether this extension impairs performance to any significant degree. Together, these conditions enabled us to understand differences in behavior and performance between dual-view and interactive search, as well as to determine whether performance at the 30-minute time point was poorer than at the 20-minute time point.

The two experiments each focused on one of the key differences between Dual-view and Interactive search. The goal of Experiment 1 was to focus purely upon the effects of interactivity whilst removing the complications imposed by the different sets of items permitted in X-ray screening relative to CT screening. To achieve this, in all conditions, participants searched for simulated guns, knives and IEDs, and we ensured that no prohibited items (i.e., liquids and laptops) were present within any of the bags. Participants were informed of this whilst being trained with regards to the task.

In Experiment 2, on the other hand, we did include items beyond guns, knives and IEDs that would require screeners to make a 'reject' decision with respect to X-ray screening. We took this approach because this represents the situation that has been implemented for X-ray and CT screening in current airport practice. On ecological grounds, therefore, it was important that we evaluated changes in performance between the different display modalities in conditions that mirror as closely as possible the situation that exists in everyday airport screening. Thus, here, participants searched for guns, knives and IEDs in both the Dual-view and the Interactive search conditions. However, unlike Experiment 1, for both conditions, laptops and liquids could also appear in the displays. In some cases, laptops were inside a piece of simulated baggage; in other cases, the laptops were outside of a piece of simulated baggage. Participants in the Dual-view search condition were asked to 'reject' simulated baggage that contained liquids and laptops, whilst those in the Interactive search condition were asked to treat these objects as being safe. Participants in both conditions were asked to 'clear' simulated baggage that contained a standalone laptop that was not inside a bag since such items would require a 'clear' decision from airport screeners.

Predictions

To recap, here our focus of interest was a comparison between Dual-view Search and Interactive Search. We also sought to understand how time on task effects interact with these different forms of search.

Display Type Effects. For both experiments, in line with previous research that has also recruited naïve participants and involved them searching for a range of different threats, we expected RTs for Interactive Search to be longer than for Dual-view Search, and for response accuracy to be higher for Interactive Search than Dual-view Search. We expected response accuracy to be higher for Interactive Search because of the advantages offered by manipulating the displays compared with Dual-view Search. We expected RTs to be longer for Interactive Search than Dual-view Search because participants would need to spend additional time interacting with the displays to acquire additional information that would then serve to enhance the accuracy of their responses.

Time on Task Effects. Moving to time on task effects, we examined these using two measures. First, as noted above, we examined performance in both types of search when participants were given either 20-minute time limit or a 30-minute time limit. This approach has been used in previous studies of time on task effects described above, but the limitation is that it is a rather coarse measure. To highlight this, consider a scenario wherein participants are searching for twenty minutes, and at the 15-minute time point, performance collapses significantly. Taking a mean of these twenty minutes would likely be insensitive to such a shift because it would ignore moment-to-moment changes in behavior. Because of this issue, we also took a more fine-grained measure of time on task effects and additionally considered performance in terms of *the time*

elapsed from the start of the main search trials for each individual trial, rather than taking an average across all trials (see Buser et al., 2023).

Based on previous findings that have examined time on task effects for low prevalence targets (Buser et al., 2019, 2020), we predicted that participants would become biased towards responding 'absent' over time. An increasing bias to respond 'absent' should result in a decrease in response accuracy for target-present trials over time, and an increase in response accuracy for target-absent trials over time. The bias towards 'absent' responses should also occur alongside a concomitant decrease in reaction times. We, therefore, predicted that target-present trial accuracy would be lower, target-absent trial accuracy would be higher, and RTs would be reduced on average for the 30-minute time limit condition than the 20-minute time limit condition, and that there would be a gradual shift in performance along these lines when this is framed in terms of the time elapsed from the start of the trials. As noted above, given the motoric effort associated with engaging in the Interactive Search condition compared with the Dual-view Search condition, we expected the effects of time on task to be greater for the Interactive Search condition (i.e., we expected an interaction to occur). Because of this, we expected the reduction in target-present trial accuracy and RTs as well as the increase in target-absent trial accuracy between the 20 minute time limit and 30 minute time limit to be greater for Interactive Search than Dual-view Search.

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. The experiments reported here were not preregistered. Data and experimental code for the experiments are available to access online via this web address (Godwin et al., 2024): https://osf.io/bdxyf/?view_only=4ca5348ed7f44ad1887f31a972feb1af.

Materials have not been made available given the security-sensitive nature of the work. The data were analyzed using the statistical software package R version 4.3.0, the *lme4* R package (Bates et al., 2015) version 3.1-162, and the *ez* R package version 4.4-0 (Lawrence, 2016).

Ethical Approval

The experiments were given a Favorable Opinion by the Ministry of Defence Research Ethics Committee on 23rd January 2023 (protocol reference: 2181/MODREC/22) and given ethical approval by the University of Southampton's Ethics Committee on 3rd February 2023 (application reference: 79469.A1).

Data Collection Approach

The experiments took place online during February of 2023, with participants recruited from Prolific, a popular crowdsourcing platform. Previous studies have demonstrated that the data quality from participants recruited via Prolific is very high . Participants were directed from Prolific to an in-house data collection server at the

University of Southampton running a piece of software called Just Another Tool for Online Studies (or JATOS: Lange et al., 2015) that was used to run the experiments for each participant. The experiments themselves were coded using jsPsych (de Leeuw, 2015), which has been demonstrated to have a very high level of precision for measuring RTs (Pinet et al., 2017).

Participants

A total of 1231 participants took part across both studies; 618 participants in Experiment 1 and 652 participants in Experiment 2. Participants were paid £7.50 for taking part in the study. After reading an information sheet and consent form, participants completed an online form to indicate their consent to take part in the experiments. Participants had a mean age of 37.31 years (*SD* = 11.38 years). 516 participants reported that they were female (41.92%), 713 reported that they were male (57.92%), and 2 'preferred not to say' (0.16%). A total of 1027 of the participants (83%) identified as white, 89 (7%) identified as Asian, 56 (4%) identified as Black, 43 as Mixed (3%), 14 identified as 'other' (1%), and for 2 participants, information on this was not available. In terms of employment status, 613 (50%) of participants reported being in full-time employment, 211 (17%) reported working part-time, 144 (11%) reported being not in paid work, 123 (10%) reported being 'unemployed (and job seeking)'. For 69 participants (5%), employment information was not available. Fifty-five participants reported their employments status as 'Other' (4%), and 16 participants reported that they were due to being a new job imminently (1%).

The Prolific platform enables researchers to use a series of filters to prevent advertising a study to certain sets of participants. For both of our experiments, we used the filters listed in Table 1. These filters were included to ensure a high level of data quality. As with many on-line and laboratory-based studies of human visual cognition, we relied on self-report of participants to ensure that the requirements of recruitment were met. Beyond these filters, we also used jsPsych's *browsercheck* plugin to only permit participants to take part in the study if they were using a laptop or desktop computer that had a display resolution of 1024 x 768 or greater. Participants were also prevented from taking part in Experiment 2 if they had already taken part in Experiment 1.

Table 1.

Filters used on Prolific for Experiments 1 and 2

Filter Description
Only include participants who report themselves as fluent English speakers from the UK
Only include participants with an approval rating on Prolific of 95% or above. This means that >= 95% of the studies that they have taken part in were 'approved' by researchers, i.e., the datasets did not have any flaws or issues, or failures of attention tests.
Only include participants who report having normal or corrected-to-normal vision.
Only include participants who report having normal colour vision.

We did not conduct a power analysis prior to data collection given the fact that our experiments used a very different task, and very different stimuli, to any published search studies before this point. Following recommendations regarding transparent reporting of sample size selections (Lakens, 2022), we recruited as large a sample size as possible given the financial resources and time that we had available for this project. We did, however, conduct a sensitivity power analysis using R once the experiments were complete. We set our alpha to 0.05 and used a series of simulations to determine the smallest effect size that we could detect with our analyses. Unless otherwise stated in the Results section, our sample size was sufficient to detect all significant effects with 0.8 power.

Design and Procedure

Both experiments used a 2 (Search Type: Dual-view search, Interactive search) x 2 (Time: 20 minutes, 30 minutes) between-participants design. Participants in each experiment were assigned to one of four conditions: Dual-view search with a 20-minute time limit, Dual-view search with a 30-minute time limit, Interactive search with a 20-minute time limit or Interactive search with a 30-minute time limit. The between-subjects design ensured that there was no cross-contamination between conditions.

After giving their consent to take part, participants were asked to minimize any potential distractions in their environment. They then engaged in training with regards to how to search the displays. This included providing them with individual examples of the different target types, as well as individual examples of the different distractor types in the form they would be displayed in the experiment. Participants engaged in the Dual-

view search condition were shown a dual view of each item and participants engaged in the Interactive search condition were given experience with how to interact with the displays. Once the training was complete, participants undertook 20 practice trials. A target was present on 50% of the practice trials. After each of the practice trials, participants were given feedback as to whether they had given a correct or an incorrect response. In total, the training and practice trials lasted around 15 minutes for each participant.

Once the training and practice trials were complete, the main trials began. No feedback was given to participants in the main trials after each response. A target was presented on 20% of trials in both experiments, to mirror the low target prevalence rates in airport screening (Wolfe et al., 2007). Trial order was randomized for each participant.

A timer started at the onset of the first of the main trials. At the end of each trial, the elapsed time on the timer was compared to the given time limit for that participant (20- or 30-minutes). If that time limit had been reached or exceeded, the experiment ended, and participants were presented with a debrief regarding the study. A consequence of this aspect of our experimental paradigm was that each participant took part in a different number of trials, depending upon their own rate of pace through the study coupled with their given time limit. It should be noted that this approach of ending the study once a time limit has been reached is not standard in visual search experiments, where participants will typically be engaged in an experiment for a fixed number of trials. However, we decided to end participation once the time limit had been reached because live airport screening sessions also terminate once a time limit has

been reached. In other words, we sought to adopt this time limit approach to maximize ecological validity for one of the main factors under examination (i.e., time on task).

A summary of the different trial types, and how often they were presented to participants, is provided in Table 2. In Experiment 1, the targets in both the Dual-view search and Interactive search conditions consisted of guns, knives, and IEDs. These were presented on an equal proportion of trials. In Experiment 2, participants in the Interactive search condition were asked to search for guns, knives and IEDs only, and these were presented on an equal proportion of trials. The Interactive search condition here also involved the presentation of liquids inside bags, laptops inside bags, and laptops outside of bags. In all cases, participants in the interactive search condition were instructed not to treat these liquids and laptops as targets. Participants engaged in the Dual-view search condition of Experiment 2 were asked to search for guns, knives, IEDs, liquids and laptops that were contained within bags. These targets were also presented on an equal proportion of trials. Participants in this condition were asked to treat laptops that were outside of bags as not being targets.

Table 2

Percentage of Trials of each Trial Type in the Different Display Conditions for Experiments 1 and 2

Experiment	Condition	Guns	Knives	IEDs	Liquids	Laptops (inside)	Laptops (outside)	Absent*

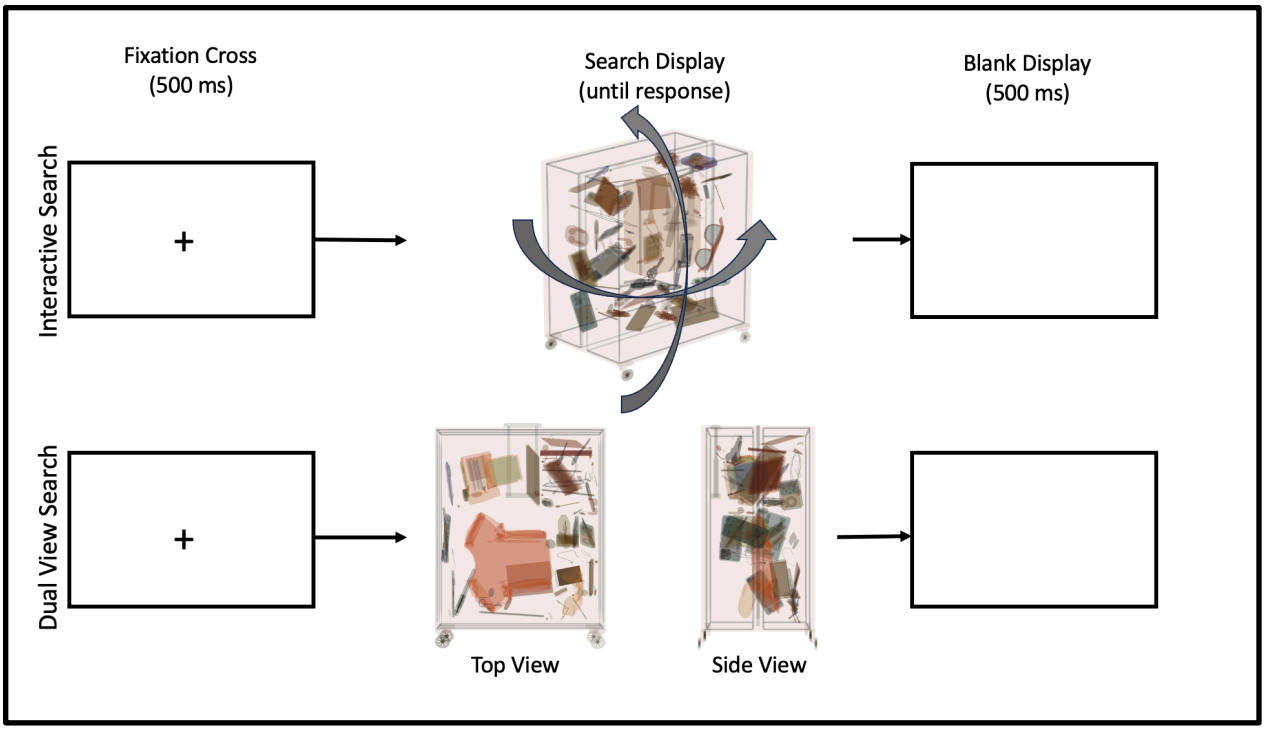
1	Dual-view	6.6%	6.6%	6.6%	-	-	-	80%
	Interactive	6.6%	6.6%	6.6%	-	-	-	80%
2	Dual-view	4%	4%	4%	4%	4%	4%	76%
	Interactive	6.6%	6.6%	6.6%	4%	4%	4%	68%

**'Absent' here indicates target-absent trials that do not contain a liquid or laptop object and require a "clear" decision. Note: Red shaded regions were treated as targets; green shaded regions were treated as non-targets.*

Each trial began with a fixation cross at the center of the display for 500 ms, after which the search display was presented. The search display remained in view until participants responded. After a response had been given, the trial ended and was followed by a blank screen for 500 ms. The trial procedure for each trial is presented in Figure 1.

Figure 1

Trial Procedure for the Interactive Search Condition and the Dual-view Search Condition.



At the end of the study, we used a technique that has been recommended for maximizing data quality in online experiments (Meade & Craig, 2012). This involves simply asking participants whether their data should be included as part of our analyses. The datasets from participants answering 'yes' to this question were retained for the final analyses; data sets of participants answering 'no' to this question were removed from our final analyses.

Apparatus

Participants took part in the study using their own computers and laptops. They were asked to respond that a bag should be 'rejected' (i.e., contained a target) using the R key on their keyboards and that a bag should be 'cleared' (i.e., deemed safe) using the W key on their keyboards.

Participants who were engaged in the Interactive search condition used their computer mice and trackpads to rotate and zoom the interactive displays. In this condition, we checked the rotation and zoom level of the camera on each screen refresh. If the rotation of the camera had changed by >= 0.2° or the zoom level had changed by >= 0.1 (for the zoom threshold, the 0.1 is in an arbitrary unit recorded in terms of the distance between the camera and the center point of the display), then the updated rotation and/or zoom level were recorded.

Stimuli

Stimuli Presentation. The stimuli were presented using a customized version of jsPsych's *canvasKeyboardResponse* plugin. This plugin presents a trial containing a single HTML canvas element and receives keyboard responses. We adapted the standard 2D HTML canvas element from this plugin to present 3D WebGL rendered images. The 3D images were presented and controlled using the *three.js* JavaScript library . The simulated baggage was created in the same manner for the X-ray and CT conditions: the conditions only differed in terms of the viewpoint(s) presented to the participants. We describe these in detail below.

The images presented to participants in both the Dual-view and Interactive search conditions were rendered from a viewpoint of a virtual camera. In the Interactive search condition, interactivity was achieved using three.js' *Orbit Controls* functionality. This enabled the participants to move the virtual camera from which the simulated baggage was being viewed, including rotating the view using their computer's mouse or trackpad, as well as zooming into and out of the image. These Orbit Controls had their

viewpoint reset to their original viewing position after each trial. We set the Orbit Controls' minimum distance to 0.75 and their maximum distance to 6 (again, these values are in arbitrary units). This prevented participants from zooming so far into the simulated baggage that they moved the camera to a position practically behind the baggage (which caused all objects to disappear). It also prevented participants zooming so far out that the baggage became a barely discernible point in the distance. In addition to limiting the zoom distances, we also disabled participants' ability to pan the Orbit Controls as during piloting, we found that panning became confusing for participants, such that when a participant panned the object to one side they then struggled to bring it back to the screen, meaning that the simulated baggage disappeared from view and was difficult to find it again. Disabling the ability to pan the images prevented this from happening.

In the Dual-view search condition, the viewing area was split into two regions: one containing the top-down view and the other containing the side view. Each region presented the view from one of two cameras. These cameras were placed in different locations to create the top-down and side views respectively. For the Dual-view search condition, Orbit Controls were disabled, thereby preventing participants from rotating the simulated baggage or zooming into and out of it.

Object Library. We created a set of custom 3D objects for the study using Blender (Hess, 2013). These 3D objects were created based on an existing library of 2D X-ray images that we have used in past research (Godwin et al., 2010). All objects were created as closely as possible to their actual size to ensure that their proportions were

maintained when included in the simulated baggage. The bags were created within the maximum limits allowed for carry-on baggage in the UK.

The categories of objects that we created were guided from the results of an unpublished survey conducted by our funders that asked participants what objects they typically pack into their carry-on baggage on flights. To ensure variety of our stimuli, we created 10 different bag structures to be used in the trials and filled them with objects, simulating an item of carry-on baggage. For target objects, we created 10 guns, 10 knives, 10 IEDs, 10 liquids (for Experiment 2 only) and 10 laptops (for Experiment 2 only). We created six different items of distractor objects from the following categories: hardback books, paperback books, cameras, mobile phones, charging cables, clothing, coins, electronic tablets, food, headphones, music players, notepads, paper, passports, pens, pencils, sunglasses, wallets, and wash kits. To make the simulated luggage appear ‘messy’ and more realistic, we also created three different items of distractor objects from the following categories: crumbs, ear buds, fluff, hair, paperclips, pen lids, pills, ripped paper, and wrappers.

As part of the baggage creation process we categorized distractor objects as either being ‘large’ or ‘small’. This was based on an examination of the overall distribution of the different distractor object volumes. On each trial, an object could only be presented once, and there was a limit of three instances of each object category. This was, except for the ‘messy’ objects, for which there could be 10 copies of each specific object with no more than 50 copies of objects from that object’s category (e.g., there could be no more than 10 copies of one specific paperclip, with no more than 50 paperclips in total per bag).

Creating Simulated Baggage. A set of example stimuli is presented in Figure 2. On each trial, a new piece of simulated baggage was created. This process began with randomly selecting a bag to use and fill with objects. Each bag had a defined ‘safe zone’ wherein objects could be placed. This was implemented to avoid objects overlapping with bag components such as wheels, zips, or the beveled corners of bags.

Figure 2

Grid Display of Random Selection of Simulated Baggage from Experiment 1 and 2

Note: Figure displays a 5 x 5 grid of a random selection of simulated baggage from a single run of the experiment. Items were selected randomly from the object library and fitted into each bag using the bag packing algorithm.

After a bag had been selected, the first object was then added to the simulated baggage. On target-present trials, the first object added was a target: this approach was taken to ensure that the target could fit within the bag and thereby mean that the trial was genuinely a target-present trial. On target-absent trials, the first object added was a distractor from the 'large' object category. Once selected, the object was rotated at random from 0° to 360° in the *x*, *y*, and *z* planes. There were then 100 'attempts' at finding a location where the object could fit within the bag. Each of these attempts involved selecting a random position on the *x*, *y*, and *z* axes within the areas of the bag's safe zone and then checking that the object still fit within the confines of the bag at that location. This check placed a virtual cube around the object and then examined whether that cube did not leave the bounding area of the bag.[1] If it did fit at that location, the object was placed there. After 100 failed attempts at finding a viable first object location, a new object was selected. This process continued until a location was found that allowed the first object to fit within the bag.

[1] Virtual cubes are often used for comparing the placement, as well as any overlap or collisions, in 3D spaces because a direct comparison of, say, individual pixels and their overlap between two 3D objects is extremely complex and would likely require a very lengthy period of time to complete. For this reason, it is commonplace when using 3D spaces to determine whether a collision or overlap has occurred by comparing the location of two virtual cubes.

Once the first object had been added, the remaining objects were added in a series of up to 700 more ‘attempts’. The first 350 attempts involved focusing on finding viable locations for ‘large’ objects and the second 350 attempts involved focusing on finding viable locations for ‘small’ objects. A similar approach has been used in previous work (Parker et al., 2022), and was necessary for the simple reason that, placement of the smaller objects substantially reduces potential locations for lager objects and therefore precludes their inclusion. Each attempt involved trying to find a viable location in the bag for a given object that would not: (i) result in the object extending beyond the defined safe zone for the bag and (ii) result in the object overlapping any other objects. The positioning of objects as part of these 700 attempts was guided by a novel algorithm designed to very rapidly find viable locations in 3D space. Overall, an average of 73.85 objects were presented in each bag (*SD* = 11.99 objects).

Results

Analytic Approach

We present the results of our analyses for both Experiments 1 and 2 together. We do this for brevity and to enable a clearer view of any consistent (or inconsistent) patterns of effects across the two experiments. Given the differences in the task (in terms of targets being searched for and objects included in the displays) between Experiments 1 and 2, we explicitly chose not to undertake formal statistical between experiment comparisons.

We examined response accuracy rates and RTs using Generalized Linear Mixed Models (GLMMs). We used this type of model for both accuracy and RTs as they have

been argued to be beneficial for reaction time data as well as accuracy data (Lo & Andrews, 2015). These models have a substantially higher level of statistical power than ANOVAs, and are better able to capture fine-grained effects, when present in a dataset. This is, in part, because these models take as their input every trial in the study, rather than averages across many trials (as in the case of other statistical procedures such as ANOVAs). They are also particularly beneficial to use when cell counts are imbalanced, as is the case here (for example, because target-present trials only account for 20% of the trials, and because participants did not complete a set number of trials since the study ended when a time limit was reached).

Our model fitting process for the GLMMs was as follows. We began with a full model containing a full set of interactions, along with a full set of interactions as random slopes for each participant (for the one within-participant factor only, with participant ID being included as a random factor in all cases). We then iterated through different model variants in a systematic order. The factor or interaction accounting for the least variance was removed (trimmed) and the overall fit of this new model was then compared to the previous 'untrimmed' model using a likelihood ratio test. If the newer model was a significantly better fit, then it became the 'winning' model. This process repeated until a 'final' model was reached. The final model was reached when no further factors or interactions could be removed without impairing the overall fit of the model.

For both response accuracy and RTs, we used the same set of fixed factors in our GLMMs. These were: Display Type (Dual-view search, Interactive search), a categorical factor focusing on the type of display being searched; Time Limit (20 mins, 30 mins), a categorical factor representing the time limit for that participant's searching;

Presence (Target Present, Target Absent), a categorical factor that represented whether a target was present or absent on each trial; and finally Time Elapsed, which was a continuous factor, represented in terms of the number of minutes since the main trials of the study began, and was used to capture time on task effects. This factor was centered following standard procedures prior to running the models.

In terms of the dependent variables, we took the following approach. For our response accuracy analyses, we coded accuracy as a binary variable, with '1' representing a correct response and '0' representing an incorrect response. Given the binary variable used here, our GLMMs that examined response accuracy used a binomial link function. For our RT analyses, we log-transformed the values prior to analysis because the values were spread across a wide range (we discuss this point further below). We used a gamma link function for the GLMMs for RTs (Lo & Andrews, 2015).

We did not examine in detail the interactive behavior that participants engaged in during for the interactive search trials. However, we note that, on every trial, participants either moved the display or used the zoom function at least once.

Data Cleaning

An extensive set of data cleaning procedures was undertaken before conducting the analyses. We began with data from 618 participants in Experiment 1 and 652 participants in Experiment 2.

Inclusion of First Participations Only. Due to an error on our part in the filtering of participants on Prolific, some participants were permitted to engage in the

experiments more than once. Because of this, we filtered out any second or subsequent engagements of the experiments from the same participants. For Experiment 1, data from 519 (84%) participants was retained; for Experiment 2, data from 530 (81%) participants was retained.

Inclusion of Completed Datasets Only. Of the participants included at this stage, 464 (89%) completed Experiment 1 and 445 (84%) completed Experiment 2.

Inclusion Based on Honesty Declaration. From the set of completed participants, we then removed those who reported that their data should not be used due to concerns over data quality. For Experiment 1, 458 (99 %) reported that their data should be used, and for Experiment 2, 443 (99%) reported that their data should be used.

Inclusion Based on Rendering Time. Whilst developing the experiments and software, we tested them using a range of different computers with different specifications, including lower-specification machines. When doing so, we found no evidence of the computers slowing down or failing to render the images in the trials. However, for the experiments, a check was made to ensure that participants using computers that were slow to load and produce the search displays were excluded. In order to achieve this, the time between the start of the process to select, position and render the objects in each display, and the end of that process was examined. For a clear majority of trials and participants, this process was very rapid (*Mdn* = 191.2 ms, with 92% of trials being rendered in < 500 ms). However, for some participants, rendering was very slow (maximum of 53.79 seconds). With that in mind, slow-rendering trials were identified as those that rendered in more than 2.5 standard

deviations from the mean time to render. Any participants whose data showed signs of one or more slow-rendering trials was removed from the analyses. This resulted in us retaining 446 (97%) participants in Experiment 1 and 426 (96%) participants in Experiment 2.

Inclusion Based on Behavior. The final steps for our data cleaning process involved identifying and removing outliers in terms of participant behavior. First, we identified participants with very fast or very slow RTs. Although it is difficult to be certain, participants who are not engaging with the task can often exhibit very rapid RTs, wherein they respond as rapidly as they can in order to quickly end the study and receive payment. The counterpart situation exists for very long RTs which can be indicative of participants being distracted from the task, perhaps even leaving their computer for a period of time before returning and providing a response. Indeed, there were 343 trials wherein the RT was 0 ms at this stage of the cleaning process: in this instance, participants were likely pressing a response key repeatedly in order to end the trial and move on. The longest trial for the participants included at this stage was 80.87 minutes in duration, where the participant likely left the computer for an extended period of time.

Given these extreme values for RTs, we defined a ‘fast’ RT for a trial as being one that was < 250 ms. This, we set based on outliers used in previous studies, but also because it is implausible that a participant could have meaningfully engaged in a search of the displays and made a decision in such a short time. We defined a ‘slow’ RT for a trial as being one that was > 60,000 ms (i.e., 1 minute). We set this value as being just over double the time limit screeners in some airports are given to process each display

(e.g., several airports currently use 25 seconds as a time limit in CT 3D screening). Given the fact that the participants in the current study did not receive the same level of training as screeners, and that the displays did not offer any of the image enhancements that airport screeners can use, this seemed a reasonable limit by which participants should have been able to complete a search. Based on these limits, we removed participants who exhibited more than 5 fast or more than 5 slow RTs. After these removals, we retained 421 (94%) participants in Experiment 1 and 410 (96%) participants in Experiment 2. Finally, for the participants retained at this stage, we removed trials that were less than 250 ms or greater than 60 seconds in duration. When examining the data at this stage, we also detected a trial in the study that was completed far after the time limits (~43 minutes after the study began) and we removed this trial accordingly. Overall, these filters resulted in us retaining 60,566 trials from Experiment 1 (99% of the dataset) and 61,807 trials from Experiment 2 (99% of the dataset).

Next, all participants who scored less than 50% accuracy (i.e., below chance) on either target-present or target-absent trials were removed. This was done to ensure that all participants included in the final dataset were able to complete the tasks to some degree of competency. After applying this filter, 340 participants were retained in Experiment 1 (81%) and 355 participants in Experiment 2 (85%).

The final stage of the data cleaning process involved removing trials that ended after the Time Limit for the given Time Limit condition. This resulted in the removal of approximately 1% of the dataset from Experiment 1 and 1% of the dataset from Experiment 2.

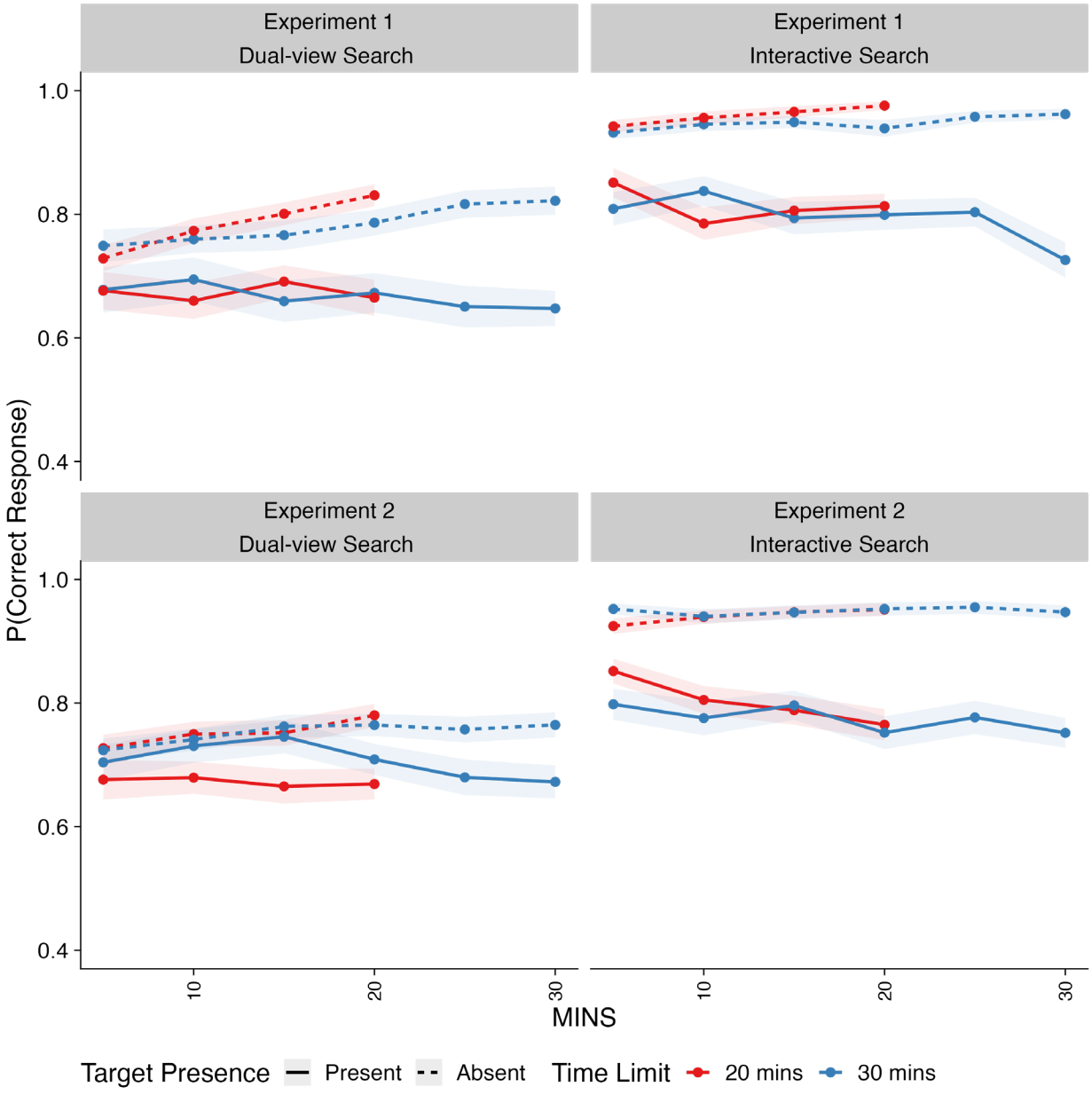
Final Dataset. The final dataset comprised 42,440 trials from 340 participants in Experiment 1 and 48,451 trials from 355 participants in Experiment 2.

Response Accuracy

We examined response accuracy as a function of time on task for the two experiments. We ran separate GLMMs for target-present and target-absent trials. Descriptive statistics for the results of our analyses of response accuracy are presented in graphical form in Figure 3. The best-fitting models for each experiment are presented in Table 3. Across the four GLMMs, there were three main effects (Display Type, Time Elapsed, and Time Limit), and two interactions (Display Type x Time Elapsed and Display Type x Time Limit). We discuss each of these in turn.

Figure 3

Response Accuracy Rates for Dual-view and Interactive Search as a function of Time Elapsed for Present and Absent Trials in Experiments 1 and 2



Note: Shaded areas represent ±SE. Time-on-task has been binned into increments of five minutes for the purposes of visualization only.

Table 3.

Best-fitting GLMMs for Response Accuracy Rates in Experiments 1 and 2, broken down for GLMMs focusing on Experiment 1 Present Trials, Experiment 1 Absent Trials, Experiment 2 Present Trials and Experiment 2 Absent Trials

	Experiment 1 Present Trials			**Experiment 1 Absent Trials**			**Experiment 2 Present Trials**			**Experiment 2 Absent Trials**		
Predictors	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*
(Intercept)	1.06 (0.04)	0.99-1.14	**<0.001**	3.02 (0.10)	2.82-3.22	**<0.001**	1.05 (0.04)	0.98-1.12	**<0.001**	2.69 (0.08)	2.53-2.86	**<0.001**
Display Type (Interactive - Dual-view)	0.71 (0.08)	0.56-0.86	**<0.001**	2.67 (0.19)	2.29-3.05	**<0.001**	0.52 (0.07)	0.39-0.66	**<0.001**	2.63 (0.17)	2.31-2.96	**<0.001**
Time Elapsed	-0.16 (0.04)	-0.23--0.09	**<0.001**	0.30 (0.04)	0.22-0.38	**<0.001**	-0.08 (0.03)	-0.13--0.03	**0.002**	0.19 (0.03)	0.13-0.26	**<0.001**
Display Type x Time Elapsed				0.16 (0.07)	0.02-0.30	**0.024**				0.13 (0.06)	0.01-0.25	**0.033**
Time Limit (30 mins - 20 mins)							-0.00 (0.07)	-0.14-0.13	0.971			
Display Type x Time Limit							-0.36 (0.14)	-0.63--0.08	**0.010**			

Observations	8475	33965	9660	38921

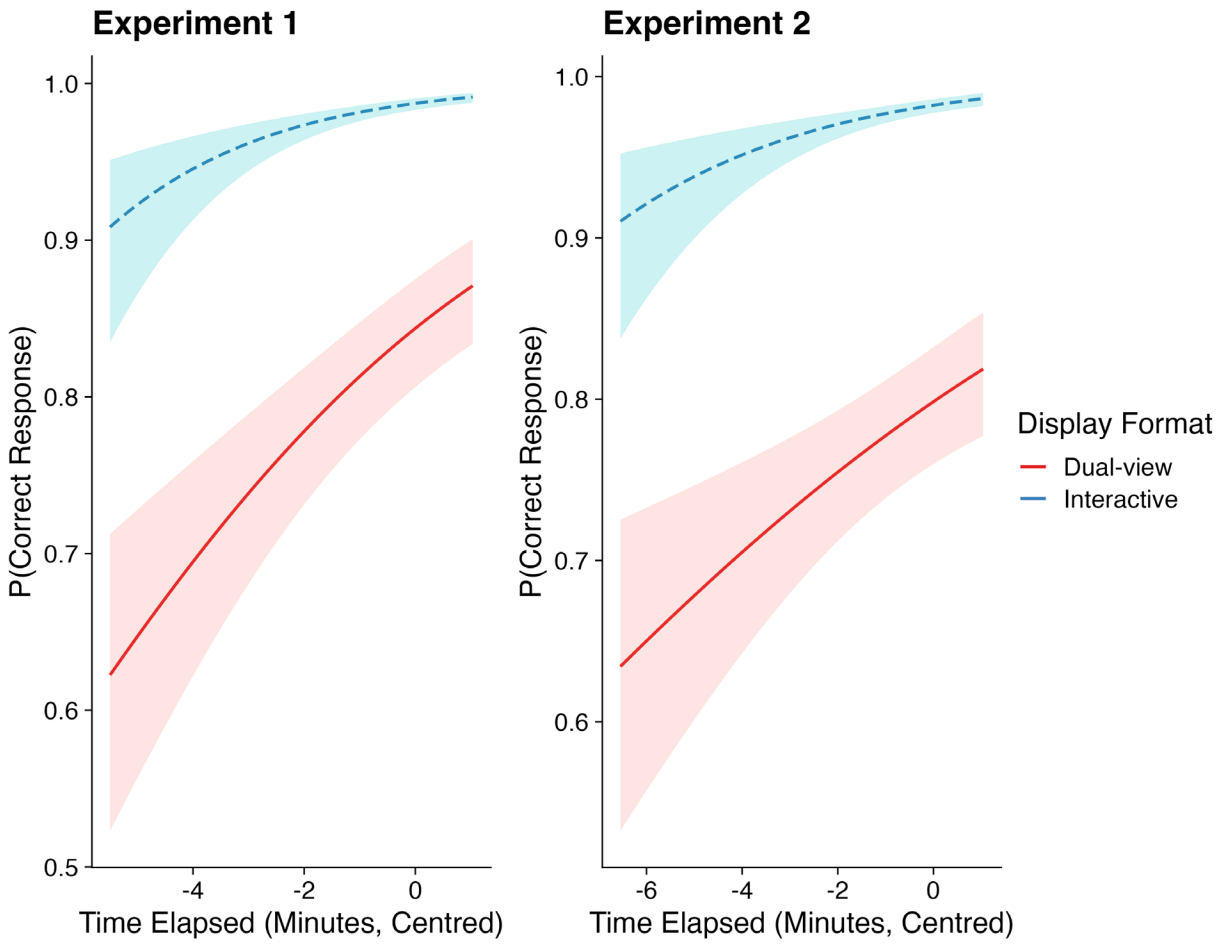
Effects of Display Type. For both target-present and target-absent trials in Experiment 1 and Experiment 2, we found an effect of Display Type. Response accuracy was higher for Interactive Search than Dual-view Search for target-present and target-absent trials. This was in line with our expectations that Interactive search would confer advantages to the recognition and detection of objects in the displays.

Effects of Time Elapsed. As time went on, participants were more accurate in their responses for target-absent trials and were less accurate in their responses for target-present trials. This finding was observed for both experiments and Display Types. Overall, this points to a general pattern wherein participants were less likely to respond 'present' as the experiment progressed, in line with previous studies of airport screeners, as well as our predictions.

Display Type x Time Elapsed Interaction. There was an interaction between Display Type and Time Elapsed for the target-absent trials in both Experiments 1 and Experiment 2 (see Figure 4). Our sensitivity power analysis revealed that, in both cases, our acquired sample size was not sufficient to detect effects as small as those we observed. We do not consider this to be problematic because compressed effects arose due to a ceiling effect in response accuracy for the Interactive Search condition, as can be seen from Figure 4. Here, we had predicted that the tendency for participants to be more likely to respond 'absent' as the experiment progressed would be greater for the Interactive Search condition than the Dual-view Search condition, but this prediction was not borne out in the results due to a ceiling limiting the possibility of observing further improvement.

Figure 4

Probability of Producing a Correct Response for Dual-view and Interactive Search during Target-Absent Trials as a function of Time Elapsed in Experiments 1 and 2



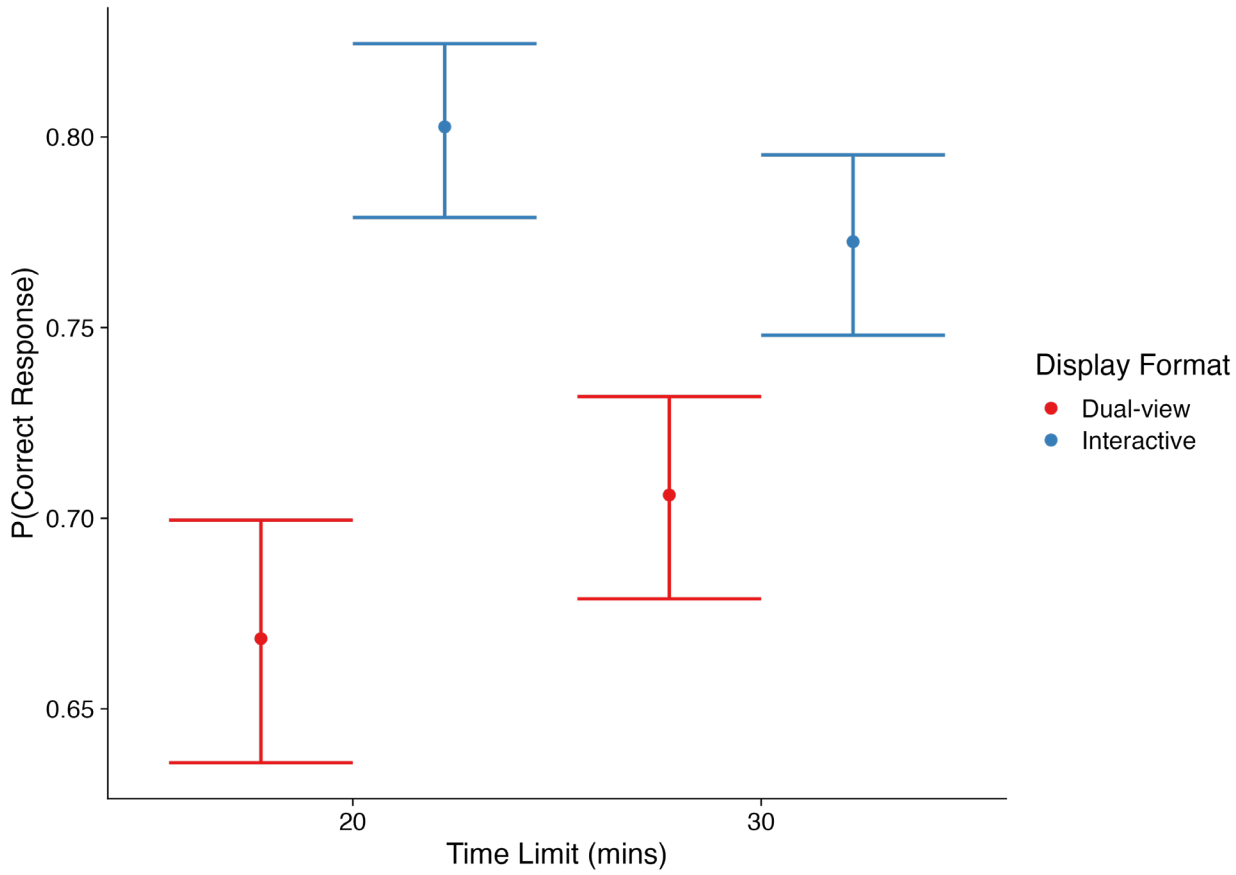
Note: Shaded areas represent ± SD

Display Type x Time Limit Interaction. There was an interaction between Display Type and Time Limit for target-present trials in Experiment 2 but not Experiment 1. While the power analysis suggested that our sample size limited our ability to reliably detect such a small effect as was observed for this interaction, this interaction was not an effect that we predicted. For these reasons, we did not examine this interaction

further, though we have plotted out these effects for visualization purposes (see Figure 5).

Figure 5

Probability of Producing a Correct Response for Dual-view and Interactive Search as a function of Time Limit for Target-present Trials in Experiment 2



Note: Bars represent ± SD

Summary: Response Accuracy. The analyses of response accuracy rates for Experiments 1 and 2 produced highly consistent patterns of effects. Participants exhibited higher response accuracy during the Interactive Search than the Dual-view Search condition. As the experiment progressed, participants became more likely to respond 'absent', resulting in decreasing accuracy rates for target-present trials and

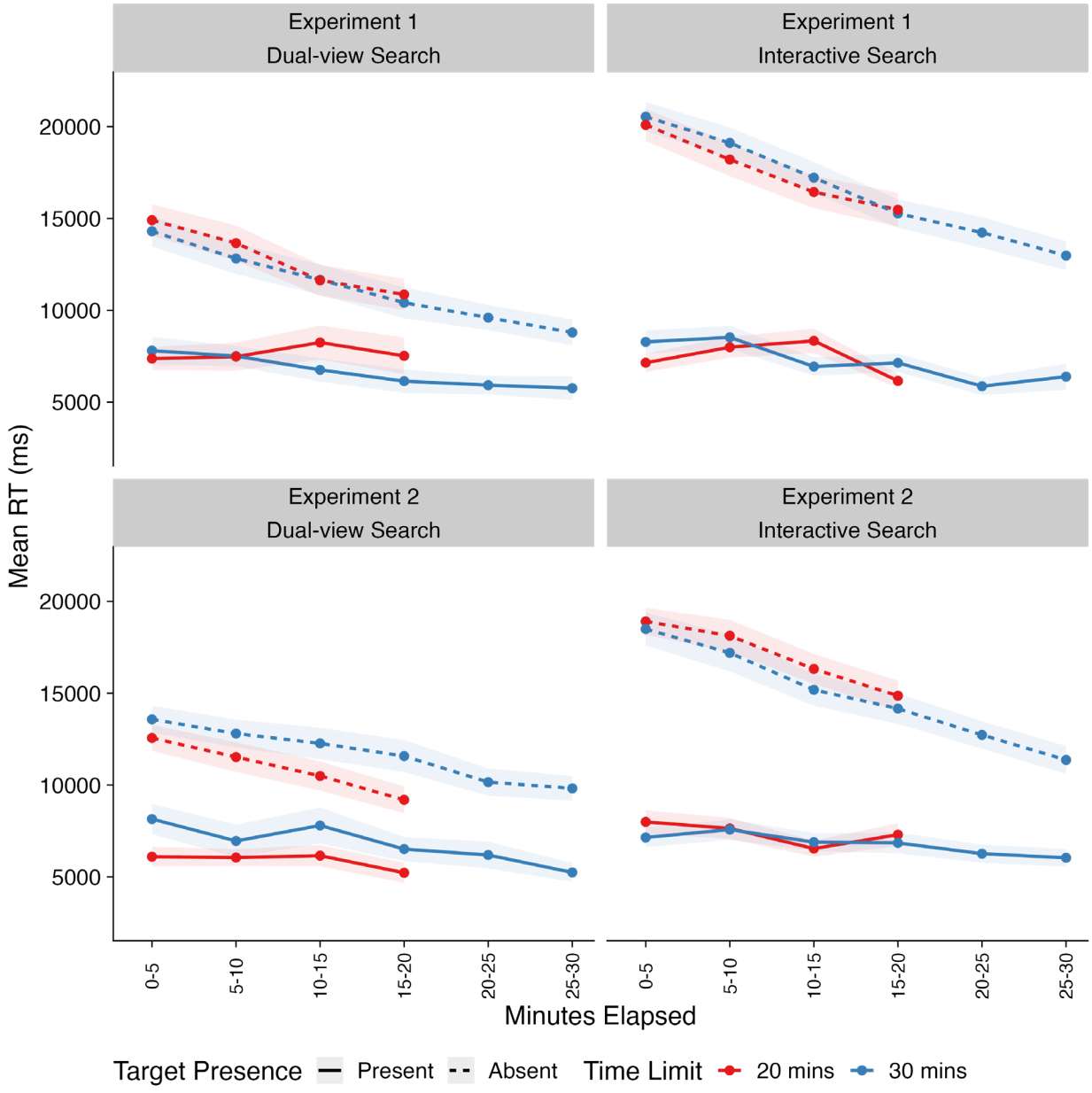
increasing accuracy rates for target-absent trials. Since response accuracy rates were so much higher in Interactive Search than Dual-view search, this increase of accuracy rates for target-absent trials resulted in an interaction between Display Format and Time Elapsed. We had predicted an interaction, but not the interaction that we observed. The one that we observed was the result of ceiling effects for target-absent trials during Interactive Search.

Response Times

Our second set of analyses focused on participants' reaction times. Here, as is standard in visual search tasks, we focused on correct-response trials only. Descriptive statistics relating to the RTs are presented in Figure 6, and the final GLMMs for RTs in Experiment 1 and 2 are presented in Table 4. Across the four GLMMs, there were effects of Time Elapsed and Display Type only, with one interaction arising between Time Elapsed and Time Limit. We will now discuss each of these in turn.

Figure 6

Mean Response Times for Dual-view and Interactive Search as a function of Time Elapsed for Present and Absent Trials in Experiments —and 2



Note: Shaded areas represent ±SE. Time-on-task has been binned into increments of five minutes for the purposes of visualization only.

Table 4

Best-fitting GLMMs for Response Times in Experiments 1 and 2, broken down for GLMMs focusing on Experiment 1 Present Trials, Experiment 1 Absent Trials, Experiment 2 Present Trials and Experiment 2 Absent Trials

	Experiment 1 Present Trials			**Experiment 1 Absent Trials**			**Experiment 2 Present Trials**			**Experiment 2 Absent Trials**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	8.56 (0.04)	8.48-8.63	**<0.001**	9.60 (0.04)	9.51-9.68	**<0.001**	8.49 (0.04)	8.42-8.57	**<0.001**	9.45 (0.04)	9.38-9.53	**<0.001**
Time Elapsed	-0.09 (0.02)	-0.13--0.05	**<0.001**	-0.15 (0.01)	-0.18--0.12	**<0.001**	-0.08 (0.02)	-0.12--0.04	**<0.001**	-0.12 (0.01)	-0.15--0.09	**<0.001**
Time Limit (30 mins - 20 mins)				-0.04 (0.09)	-0.21-0.12	0.622						
Display Type (Interactive - Dual-view)				0.32 (0.07)	0.19-0.46	**<0.001**	0.14 (0.07)	0.01-0.28	**0.039**	0.41 (0.07)	0.28-0.55	**<0.001**
Time Elapsed x Time Limit				-0.06 (0.03)	-0.11--0.01	**0.015**						
Observations	6090			29362			6953			32654		

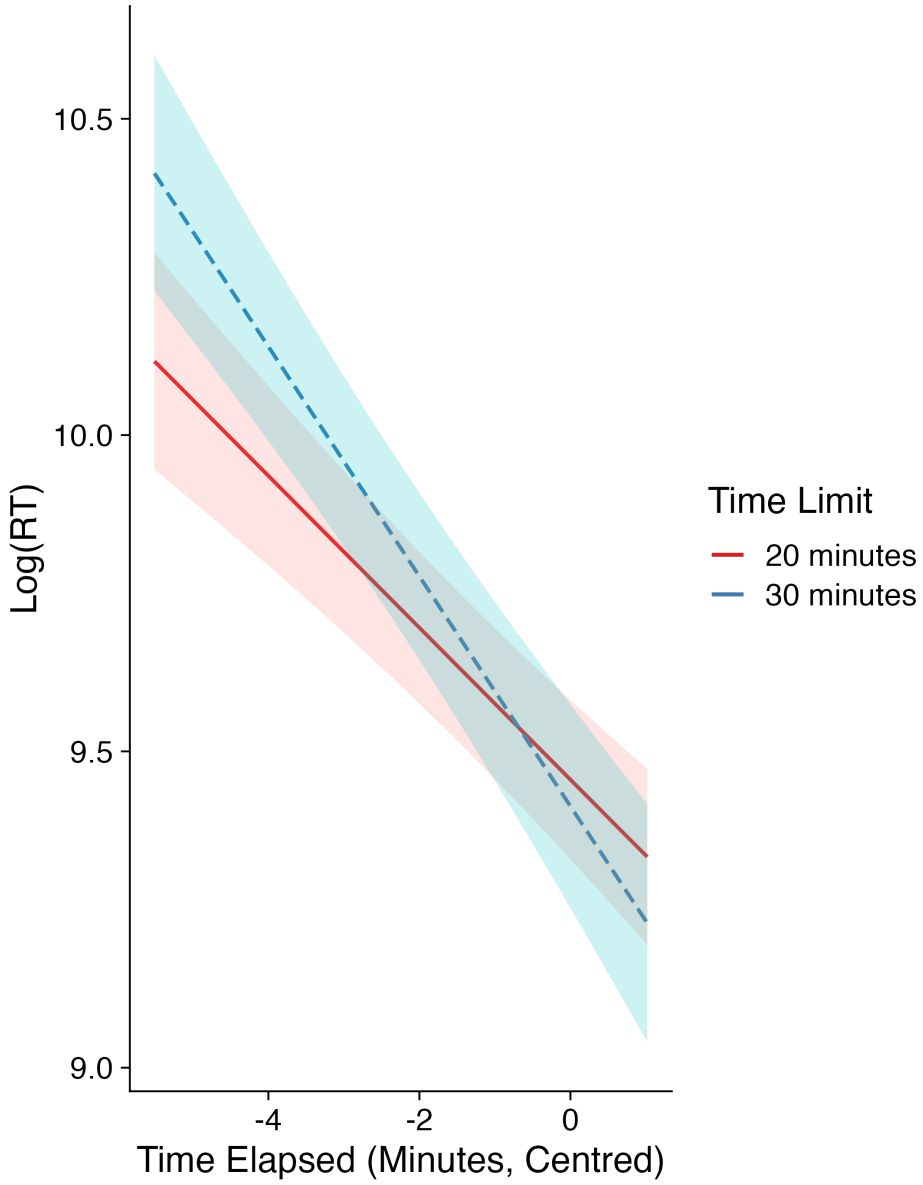
Effects of Display Type. For target-absent trials in Experiment 1, and for both target-present and target-absent trials in Experiment 2, we found a main effect of Display Type. This did not interact with any other factors. In these trials, RTs were longer for Interactive Search than Dual-view search, consistent with previous findings and our predictions.

Effects of Time Elapsed. In all cases, we found that there was an effect of Time Elapsed. Here, as the experiment progressed, RTs became more rapid.

Time Elapsed x Time Limit Interaction. For target-absent trials in Experiment 1, we found evidence of an interaction between Time Elapsed and Time Limit. However, rather than demonstrating that performance worsened over time in one Time Limit condition than another, instead this interaction arose because participants in the 30-minute Time Limit condition exhibited longer RTs at the start of the experiment than those in the 20 minute condition, as can be seen in Figure 7. One possibility is that this could have been the result of expectation effects for participants in the 30-minute Time Limit shifting their behavior based on how long they thought the experiment would take to complete. However, we did not inform participants about the exact time limit for their condition, so this cannot explain our result. Instead, it simply appears that there were some baseline differences between participants at the beginning of the experiment.

Figure 7

Response Times for Twenty- and Thirty-minute Conditions as a function of Time Elapsed (Centered)



Note: Shaded areas represent ± SD

Summary: Response Times. Our findings from the analyses of the RTs from Experiments 1 and 2 were generally consistent with one another. One clear effect that occurred for both experiments, and for target-present and target-absent trials was that RTs reduced as the experiments progressed. This was in line with our predictions. In target-present and target-absent trials for Experiment 2, as well as for target-absent trials in Experiment 1, RTs for Interactive Search were longer than for Dual-view search, again in line with our predictions.

Discussion

Two substantial online experiments (695 participants across both experiments) were conducted that compared performance and behavior when participants were searching interactive displays, compared with dual-view displays. The task used in the experiments was inspired by ongoing changes to airport screening methods. Our primary aim was to compare performance between simulated interactive and dual-view displays, alongside addressing how time on task affects performance when examining those displays. We addressed time on task effects both at a coarse level, by asking participants to search either for 20 minutes or for 30 minutes, as well as a fine-grained level, by analyzing search behavior as a function of the time elapsed from the start of the main search trials. For Experiment 1, participants in all conditions searched for the same target types in order to avoid confounds associated with differences in the number of targets being searched for across experimental conditions (Dual-view vs. Interactive Search). For Experiment 2, search occurred in line with stipulations that currently exist in airports with regards to permitted items that are in place for 3D CT and 2D X-ray

screening. Participants in our Dual-view Search condition in Experiment 2 were asked to search for guns, knives and IEDs, as well as laptops and liquids, but only when these items were inside a piece of baggage. Participants in our Interactive Search condition in Experiment 2 were asked to search for guns, knives and IEDs only. Thus, in Experiment 2, we adopted realistic stimuli that were assessed under ecologically valid search instructions by our participants. We made a set of predictions based on our key areas of interest in the experiments, focusing on predictions relating to Display Type effects, Time on Task Effects and the interaction between Display Type and Time on Task. We will now discuss each of these in turn.

Display Type Effects. We predicted that Interactive Search would produce higher response accuracy rates than Dual-view Search because Interactive Search enables searchers to rotate, inspect and zoom into the search array from a range of different perspectives, thereby aiding in the identification of targets and the rejection of distractors. Additionally, there is evidence from past research of accuracy benefits to interactive search (Hättenschwiler et al., 2018, 2019; Parker et al., 2022). In addition, we predicted that RTs for Interactive Search would be longer due to the level of inspection afforded in Interactive Search displays to acquire additional information. Our results across both experiments, and across both target-present and target-absent trials were consistent with these predictions. The results confirm previous findings suggesting that adding interactivity to the displays substantially improved accuracy rates but at a cost in the form of increased RTs for Interactive compared with Dual-view Search. While improving overall accuracy, Interactive Search is more time consuming than Dual-view Search.

Time on Task Effects. We examined time on task effects by asking participants to engage in our experiments with different Time Limits and charted how their performance shifted over time. We predicted that, as time progressed, participants would become more likely to respond 'absent' and respond more rapidly, in line with previous studies (Buser et al., 2019, 2020, 2023; Ghylin et al., 2007; Meuter & Lacherez, 2016). We quantified time on task effects using two approaches: a fine-grained approach that looked at the time elapsed for each trial during the experiments, and a coarser approach that involved providing different time limits for different conditions in the experiments (for a similar analytic approach, see Buser et al., 2023). Our fine-grained analyses of response accuracy and RTs showed clear time on task effects and supported our predictions, as well as aligning neatly with past research that has demonstrated similar effects. Here, participants were more likely to respond 'absent' as the experiments progressed, and their RTs became more rapid as well. The primary effect of time on task, at least in our experiments here, was to induce a response bias in participants such that they terminated search more rapidly and were biased towards absent responses. Critically, this finding only seems to have been found in other studies that, like ours, involved targets being presented a low rate of prevalence (Buser et al., 2019, 2020, 2023; Ghylin et al., 2007; Meuter & Lacherez, 2016).

Given the fine-grained effects of time on task, what coarser effects were there, at the level of the different overall time limits? For response accuracy, we found no evidence of Time Limit effects, with no evidence of overall differences in response accuracy for the two different Time Limits, for either of our two experiments, or for target-present and target-absent trials. We also found no evidence of significant effects

of Time Limit upon the RTs. Overall, we found no evidence that was a collapse in performance after a specific time point had been reached: rather, as time goes on during a search, there is a gradual progression towards faster RTs and a greater likelihood of target-absent responses – at least for the time limit conditions that we used here.

Interaction between Display Type and Time on Task Effects. The final set of questions to address are in relation to any interaction between the display types and time on task effects. We expected to find interactions between time on task effects and display type as a consequence of the increased motoric effort required by interactive search, with the expectation that any time on task effects would be exaggerated in Interactive Search compared with Dual-view Search. We found no evidence to support these predictions.

Key Messages. There are two key messages that can be drawn from our results. First, changing from a dual-view to an interactive display is beneficial to baggage search in terms of response accuracy, but this comes at a cost in the form of longer RTs. Second, in both experiments we obtained very consistent time on task effects. Over time, searchers respond more rapidly and are more likely to make absent responses. These effects occurred to a similar degree for both dual-view and interactive displays. Thus, we found no evidence that moving from dual-view to interactive displays will increase (or ameliorate) the influence of time on task.

Constraints on Generality

Given the large sample size and the fact that we recruited members of the general public to take part in our experiments, we believe that our findings will generalize well outside of the experiments presented here. We do not have any reasons to suspect that our findings are purely dependent upon the characteristics of the participants, materials, or the context in which the experiments took place. Using simpler stimuli would likely reduce the differences in performance between dual-view and interactive searches as search could likely be resolved with reduced or no interaction. Reducing the number of targets being searched for, or increasing the target prevalence, would likely increase response accuracy. All of this said, however, we recognize that the use of naïve participants in a study that seeks to generalize findings to expert screeners may represent a concern. Thus, having illustrated these important effects under experimental conditions with naïve participants, it will be valuable in future research to repeat the study with experienced screeners (thereby allowing for more reasonable generalization). Furthermore, we note that the costs of failing to detect targets here is not equivalent to the costs of real-world searchers, including airport screeners, failing to detect targets. Future research will therefore also need to carefully examine reward and cost structures for errors during visual search.

Practical Implications and Future Directions

Overall, then, we found that Interactive Search offers a substantial benefit to target detection compared with Dual-view Search. The cost, though, is that Interactive Search is significantly slower than Dual-view Search. However, taking this narrow view of the costs and benefits will need to be considered against a much broader range of

factors. For example, the change in permitted items introduced by CT screening may mean that, although response times may be longer than dual-view screening, the fact that liquids are now being left in baggage may be of substantial benefit. Finding liquids in baggage requires airports to manually search and then re-screen that baggage. No longer having to do so in 3D CT screening may mean that the costs of increased RTs compared with dual-view X-ray screening are outweighed by the gains associated with permitted liquids to remain in baggage. (The same is true, of course, for laptops as well).

Next, in terms of time limit effects, whilst we found very limited evidence of any real effects of increasing the time limit from 20 minutes to 30 minutes, there was evidence of a tendency to reduce 'present' responses over time. The reduction in the likelihood of responding 'present' was found across both display types and so is not linked to interactive or dual-view search specifically. The effect is worthy of further study. It is apparent that, beyond a certain time point, the practical consequence of reducing 'present' responses may reach a level that is not permissible from a safety and efficiency perspective. We reach this conclusion despite finding no overall difference between the 20- and 30-minute conditions.

We regard this project as just the beginning of a new avenue of research in this area. There are a number of future directions and next steps that would need to be taken to translate the findings from these experiments into the real world. First, and as noted above in the Constraints on Generality section, it is important that we establish a replication of the findings reported here in a sample of airport screener participants. It would also be helpful to compare these findings alongside live 3D CT screening

datasets. In addition, it would be helpful and informative to update the experimental paradigm to include more of the functions offered by 3D CT screening, such as alarms that highlight potential IEDs in baggage, ‘slice’ functions through the displays, the functionally to remove specific objects from the display once clicked, and image enhancement functions.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bravo, M. J., & Farid, H. (2004a). Recognizing and segmenting objects in clutter. *Vision Research*, *44*(4), 385–396. https://doi.org/10.1016/j.visres.2003.09.031

Bravo, M. J., & Farid, H. (2004b). Search for a Category Target in Clutter. *Perception*, *33*(6), 643–652. https://doi.org/10.1068/p5244

Bravo, M. J., & Farid, H. (2006). Object recognition in dense clutter. *Perception & Psychophysics*, *68*(6), 911–918. https://doi.org/10.3758/BF03193354

Buser, D., Schwaninger, A., Sauer, J., & Sterchi, Y. (2023). Time on task and task load in visual inspection: A four-month field study with X-ray baggage screeners. *Applied Ergonomics*, *111*, 103995. https://doi.org/10.1016/j.apergo.2023.103995

Buser, D., Sterchi, Y., & Schwaninger, A. (2019). Effects of Time on Task, Breaks, and Target Prevalence on Screener Performance in an X-ray Image Inspection Task. *2019 International Carnahan Conference on Security Technology (ICCST)*, 1–6. https://doi.org/10.1109/CCST.2019.8888408

Buser, D., Sterchi, Y., & Schwaninger, A. (2020). Why stop after 20 minutes? Breaks and target prevalence in a 60-minute X-ray baggage screening task. *International Journal of Industrial Ergonomics*, *76*, 102897. https://doi.org/10.1016/j.ergon.2019.102897

Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology*, *30*, 401–408. https://doi.org/10.1006/cogp.1996.0002

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Donnelly, N., Muhl-Richardson, A., Godwin, H. J., & Cave, K. (2019). Using eye movements to understand how security screeners search for threats in X-Ray baggage. *Vision*, *3*(2), 24. https://doi.org/10.3390/vision3020024

Drew, T., Vo, M. L.-H., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013). Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision*, *13*(10), 3. https://doi.org/10.1167/13.10.3

Ghylin, K. M., Drury, C. G., Batta, R., & Lin, L. (2007). Temporal Effects in a Security Inspection Task: Breakdown of Performance Components. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(2), 93–97. https://doi.org/10.1177/154193120705100209

Godwin, H. J., & Hout, M. C. (2023). Just say 'I don't know': Understanding information stagnation during a highly ambiguous visual search task. *PLOS ONE*, *18*(12), e0295669. https://doi.org/10.1371/journal.pone.0295669

Godwin, H. J., Liversedge, S. P., Kirkby, J. A., Boardman, M., Cornes, K., & Donnelly, N. (2015). The influence of experience upon information-sampling and decision-making behaviour during risk assessment in military personnel. *Visual Cognition*, *23*(4), 415–431. https://doi.org/10.1080/13506285.2015.1030488

Godwin, H. J., Liversedge, S. P., Mestry, N., Dewis, H., & Donnelly, N. (2024). *Time on Task Effects during Interactive Visual Search OSF Repository*. https://osf.io/bdxyf/

Godwin, H. J., Menneer, T., Cave, K. R., Helman, S., Way, R. L., & Donnelly, N. (2010). The impact of Relative Prevalence on dual-target search for threat items from airport X-ray screening. *Acta Psychologica*, *134*(1), 79–84. https://doi.org/10.1016/j.actpsy.2009.12.009

Godwin, H. J., Menneer, T., Liversedge, S. P., Cave, K. R., Holliman, N. S., & Donnelly, N. (2017). Adding depth to overlapping displays can improve visual search performance. *Journal of Experimental Psychology. Human Perception and Performance*, *43*(8), 1532–1549. https://doi.org/10.1037/xhp0000353

Godwin, H. J., Menneer, T., Liversedge, S. P., Cave, K. R., Holliman, N. S., & Donnelly, N. (2020). Experience with searching in displays containing depth improves search performance by training participants to search more exhaustively. *Acta Psychologica*, *210*, 103173. https://doi.org/10.1016/j.actpsy.2020.103173

Hancock, P. A. (2013). In search of vigilance: The problem of iatrogenically created psychological phenomena. *American Psychologist*, *68*(2), 97–109. https://doi.org/10.1037/a0030214

Hättenschwiler, N., Mendes, M., & Schwaninger, A. (2019). Detecting Bombs in X-Ray Images of Hold Baggage: 2D Versus 3D Imaging. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *61*(2), 305–321. https://doi.org/10.1177/0018720818799215

Hättenschwiler, N., Merks, S., & Schwaninger, A. (2018). Airport Security X-Ray Screening of Hold Baggage: 2D Versus 3D Imaging and Evaluation of an on-Screen Alarm Resolution Protocol. *2018 International Carnahan Conference on Security Technology (ICCST)*, 1–5. https://doi.org/10.1109/CCST.2018.8585713

Hess, R. (2013). *Blender Foundations: The Essential Guide to Learning Blender 2.5*. Taylor & Francis.

Horowitz, T. S., Cade, B. E., Wolfe, J. M., & Czeisler, C. A. (2003). Searching Night and Day: A Dissociation of Effects of Circadian Phase and Time Awake on Visual Selective Attention and Vigilance. *Psychological Science*, *14*(6), 549–557. https://doi.org/10.1046/j.0956-7976.2003.psci_1464.x

Hout, M. C., White, B., Madrid, J., Godwin, H. J., & Scarince, C. (2022). Examining the effects of passive and active strategy use during interactive search for LEGO® bricks. *Journal of Experimental Psychology. Applied*, *28*(1), 35–51. https://doi.org/10.1037/xap0000295

Hulleman, J., & Olivers, C. N. L. (2017). The impending demise of the item in visual search. *Behavioral and Brain Sciences*, *40*, e132. https://doi.org/10.1017/S0140525X15002794

Krimsky, M., Forster, D. E., Llabre, M. M., & Jha, A. P. (2017). The influence of time on task on mind wandering and visual working memory. *Cognition*, *169*, 84–90. https://doi.org/10.1016/j.cognition.2017.08.006

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLOS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lanthier, S. N., Risko, E. F., Risko, E. F., Smilek, D., & Kingstone, A. (2013). *Measuring the separate effects of practice and fatigue on eye movements during visual search*.

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments* [Computer software].

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01171

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6–21. https://doi.org/10.1080/17470214808416738

Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory: A User's Guide*. Psychology Press.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Meuter, R. F. I., & Lacherez, P. F. (2016). When and Why Threats Go Undetected: Impacts of Event Rate and Shift Length on Threat Detection Accuracy During Airport Baggage Screening. *Human Factors: The Journal of the Human Factors*

and Ergonomics Society, *58*(2), 218–228. https://doi.org/10.1177/0018720815616306

Nodine, C. F., & Kundel, H. L. (1987). Using eye movements to study visual search and to improve tumor detection. *RadioGraphics*, *7*(6), 1241–1250. https://doi.org/10.1148/radiographics.7.6.3423330

Parker, M. G., Muhl-Richardson, A., & Davis, G. (2022). Enhanced threat detection in three dimensions: An image-matched comparison of computed tomography and dual-view X-ray baggage screening. *Applied Ergonomics*, *105*, 103834. https://doi.org/10.1016/j.apergo.2022.103834

Pinet, S., Zielinski, C., Mathôt, S., Dufau, S., Alario, F.-X., & Longcamp, M. (2017). Measuring sequences of keystrokes with jsPsych: Reliability of response times and interkeystroke intervals. *Behavior Research Methods*, *49*(3), 1163–1176. https://doi.org/10.3758/s13428-016-0776-3

Riggs, C. A., Cornes, K., Godwin, H. J., Liversedge, S. P., Guest, R., & Donnelly, N. (2017). The importance of search strategy for finding targets in open terrain. *Cognitive Research: Principles and Implications*, *2*(1), 14. https://doi.org/10.1186/s41235-017-0049-4

Riggs, C. A., Godwin, H. J., Mann, C. M., Smith, S. J., Boardman, M., Liversedge, S. P., & Donnelly, N. (2018). Rummage search by expert dyads, novice dyads and novice individuals for objects hidden in houses. *Visual Cognition*, *26*(5), 334–350. https://doi.org/10.1080/13506285.2018.1445678

Sauter, M., Stefani, M., & Mack, W. (2020). Towards Interactive Search: Investigating Visual Search in a Novel Real-World Paradigm. *Brain Sciences*, *10*(12), Article 12. https://doi.org/10.3390/brainsci10120927

Smith, A. D., Hood, B. M., & Gilchrist, I. D. (2008). Visual search and foraging compared in a large-scale search task. *Cognitive Processing*, *9*(2), 121–126. https://doi.org/10.1007/s10339-007-0200-0

Smith, A. D., Hood, B. M., & Gilchrist, I. D. (2010). Probabilistic cuing in large-scale environmental search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 605–618. https://doi.org/10.1037/a0018280

Solman, G. J. F., Hickey, K., & Smilek, D. (2014). Comparing target detection errors in visual search and manually-assisted search. *Attention, Perception, & Psychophysics*, *76*(4), 945–958. https://doi.org/10.3758/s13414-014-0641-3

Wolfe, J. M. (2016). Use-inspired basic research in medical image perception. *Cognitive Research: Principles and Implications*, *1*(1), 17. https://doi.org/10.1186/s41235-016-0019-2

Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, *28*(4), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology. General*, *136*(4), 623–638. https://doi.org/10.1037/0096-3445.136.4.623

Wolfe, J. M., & Van Wert, M. J. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology*, *20*(2), 121–124. https://doi.org/10.1016/j.cub.2009.11.066