Tracking the Identity of Moving Words: Stimulus Complexity and Familiarity Affects Tracking Accuracy

Jukka Hyönä[1], Lauri Oksama[2], and Esa Rantanen[3]

[1]University of Turku, Finland

[2] Finnish Defence Research Agency, Human Performance Division, Finland

[3]Rochester Institute of Technology, Rochester, NY, USA

Running title: Tracking moving words

Keywords: multiple identity tracking, word length, dynamic attention, situation awareness, air traffic control.

Accepted for publication July 8, 20019 in *Applied Cognitive Psychology*

Abstract

In two experiments, participants tracked the identity and location of moving words. The task bears resemblance to one performed by air traffic controllers who track multiple moving aircraft, where they are identified with relatively complex alphanumeric call signs. In Experiment 1, stimulus familiarity was manipulated by comparing the tracking of familiar words and pseudowords. In Experiment 2, also stimulus complexity was varied by having the participants track short and long words. Stimulus complexity affected tracking: short words were easier to track than long words. Moreover, familiarity of identity affected tracking of short words (short familiar words were easier to track than short pseudo-words) but not of long words. The results are interpreted within the framework of the MOMIT model. Mathematical simulations suggest that observers may not have enough time for full identification of complex identities in a dynamic situation. Practical implications of the results for air traffic control are discussed.

In everyday visual environments we are often required to track people and objects that are constantly moving. For example, in social interactions people constantly update the location of their interlocutors. Navigating in road traffic requires coding of the locations of other cars, road signs, and pedestrians. In team sports, the players must track the movements of both their teammates to make passes as well as the members of the opposing team to anticipate their blocking, whereas the spectators may pay attention to the number on the back of players' shirts to access their identity and track their movements in the field.

Accurate and timely tracking of objects and their identities is even more critical in several current and near-future aviation applications. For example, air traffic controllers must track multiple aircraft as they move on plan view displays, where the moving aircraft are identified with relatively complex alphanumeric call signs (e.g., FIN343). The future air traffic control (ATC) systems will only exaggerate these task demands by placing larger numbers of aircraft under the responsibility of a single controller, who will primarily be monitoring traffic while separation of the aircraft is delegated to automated systems in the cockpits. These same ATC modernization schemes are also changing the pilots' tasks. Pilots, instead of focusing on their own navigation tasks and following instructions from controllers, must now also track movements of other aircraft on the cockpit display of traffic information (CDTI) to maintain self-separation from them. All in all, correct binding of what information to where information (i.e., dynamic identity-location binding) is of fundamental importance in these real-life tasks. If identity is bound to a wrong object, it may have catastrophic consequences (e.g., the air traffic controller giving a clearance to a wrong aircraft).

In the present study, we investigated human performance in tracking multiple moving identities. We were interested in how well people can track moving verbal stimuli (written words), roughly approximating the visual environment of air traffic controllers tracking call signals or sports spectators tracking players' identities as revealed by the name and number on players' shirts. Moreover, we were interested in examining the extent to which semantic information stored in the long-term memory (LTM) can be readily used in tracking of multiple moving identities. We investigated this question by comparing the tracking performance for existing words and pseudowords that do not have a meaning represented in LTM. We were also interested in whether stimulus complexity influences dynamic tracking. In order to do so, we had observers track short versus long words and pseudowords. These questions are relevant theoretically, as described below, but they also have important human performance implications in many naturalistic task environments.

Prior Research on Multiple Identity Tracking

Seminal research on multiple object tracking (MOT) has demonstrated that identity information could not be kept active during tracking. Scholl, Pylyshyn and Franconeri (1999) found that observers could not access color or shape information for tracked targets once they had been visually masked; they concluded that only spatiotemporal properties are encoded during tracking. Pylyshyn (2004) came to the same conclusion by demonstrating that identity information provided during the target designation phase (digits appeared inside target rings but disappeared when objects began to move) was poorly accessed during the probing phase (see, however, Cohen, Pinto, Howe, & Horowitz, 2011).

More recent studies, however, have provided evidence for successful tracking of identity information. In the study of Horowitz, Klieger, Fencsik, Yang, Alvarez and Wolfe (2007), the to-be-tracked objects were line-drawings of animals that differed from each other in shape and color. The study established a “content deficit”: identity tracking was poorer than simply tracking the location of targets. Yet, identity tracking was successfully accomplished for about two targets. Moreover, they established a unique object advantage: tracking of unique objects was better than that of identical objects. Evidence for content deficit was also found by Botterill, Allen and McGeorge (2011) using butterflies of different shape and color as targets as well as by Ren, Chen, Liu and Fu (2009) for tracking luminosity-matched faces. Moreover, analogously to Horowitz et al. (2007), Botterill et al. (2011), and Ren et al. (2009) estimated the identity tracking capacity to be about two items. Similarly to Horowitz et al., improved tracking performance for distinct objects over homogenous objects was also observed by Makovski and Jiang (2009) using colored digits as targets (an opposite trend was obtained for faces by Ren et al., 2009). Finally, Li, Oksama and

Hyönä (2019) found that observers can track three line drawings of common objects or three color disks with an accuracy of about 85% when the objects were visually masked during the movement phase. In other words, identity tracking can be done with the help of the visuospatial short-term memory, at least for a short time.

More importantly to the present study, Liu, Chen, Liu and Fu (2012) varied the complexity of target identities: In Experiment 1 the targets were 1-, 3- and 4-digit numbers. Target identity was irrelevant to the task; in other words, the participants had to only track the location of the target numbers. Tracking of unique identities was compared to that of identical identities. Liu et al. found that the tracking capacity for simple (1-digit targets) unique stimuli was better than that for simple identical stimuli. A reversed pattern was observed for complex (4-digit targets) stimuli: the tracking capacity was poorer for unique than identical targets. This pattern of results was replicated in Experiment 2 with numbers and also in Experiment 3 that used simple and complex Chinese characters as stimuli. As identity tracking was irrelevant to the task, it is concluded that identity information is involuntarily processed even when it hampered the tracking performance of complex stimuli.

It is noteworthy that in the studies reviewed above, identity information was irrelevant to the task, That is, observers were only asked to track the location of the moving targets. Yet, identity information was found to be readily processed, even when it in some situations hampered the tracking performance. Oksama and Hyönä (2004, 2008) designed a tracking task where target identity is task-relevant, unlike in the more traditional MOT tasks. In this multiple identity tracking (MIT) task, participants are asked to keep track of both the location and the identity of the targets. Using the MIT task, Oksama and Hyönä (2008) found that it is easier to track familiar objects (line drawings of familiar objects or familiar faces) than pseudo-objects (line drawings of pseudo-objects or pseudofaces).

Pinto, Howe, Cohen and Horowitz (2010) examined whether the tracking advantage for familiar objects over pseudo-objects would be confounded with better nameability of familiar objects and faces and or due to low-level perceptual differences between the object categories. In order to control for these potential confounds, they operationalized familiarity as stimulus repetition: during one block the same stimuli always served as the targets and distractors, whereas in another block they were randomly chosen for each trial from a stimulus set. Pinto et al. observed a repetition benefit in tracking: repeated items were tracked better than unrepeated items – a finding generally in line with what Oksama and Hyönä (2008) found.

Theoretical Underpinnings

Some prominent theories of dynamic visual attention seem to provide a negative answer to the human capacity to keep track of identity information of constantly moving objects. According to the object file theory of Kahneman and Treisman (1984; Kahneman, Treisman, & Gibbs, 1992), identity information plays no role in perception of object continuity as objects move in space over time. All that matters is whether or not the object is moving in a spatiotemporally continuous way. If the object’s trajectory is not continuous or logical, perception of object continuity may suffer. But a moving object can change on the fly, for example, from red to blue, from a circle to a triangle, or from a bird to a superman without any apparent influence on the perception of continuity. Similarly, identity information does not play any significant role in the visual index theory of Pylyshyn (1989, 2001), the seminal theory of multiple object tracking. According to this theory, visual tracking is carried out with the help of 4-5 indexes or pointers provided by the early vision system, which convey only location information of target objects. In particular, Pylyshyn’s indexes do not encode object identity or feature information.

If the above theorizing is correct and only spatiotemporal properties play a role in dynamic visual attention, then it would not make any difference for performance efficiency to track an existing word versus a pseudoword. As reviewed above, the recent research has repeatedly demonstrated that identity information is readily processed during MOT. More importantly to the present study, the stimulus familiarity effect obtained by Oksama and Hyönä (2004, 2008) suggests that identity information has an active role in MIT: stimulus familiarity significantly affects tracking efficiency. In order to explain this and other results, we (Oksama & Hyönä, 2008) put forward a model, coined MOMIT (Model of Multiple

Identity Tracking), which, among other things, provides an account for how object identity affects tracking efficiency. According to MOMIT, in order to maintain an up-to-date representation of the whereabouts of target objects, focal attention is continuously switched between the moving elements. The purpose of attention shifts is to update and refresh identity-location bindings. The influence of stimulus properties is mediated by the time taken to identify a target and bind location information to the perceived identity. MOMIT predicts that all the factors that influence target identification and binding time also influence the capacity to maintain multiple identity-location bindings for moving objects. The faster the bindings are serially refreshed, the better the tracking performance is. For example, as LTM representations are available for familiar faces and pictures, they are identified faster than pseudo-objects or pseudo-faces; consequently, familiar objects can be tracked more effectively refreshed and maintained than pseudo-objects during tracking.

Present Study

As it appears from the literature review presented above, the previous research has mostly used as stimuli objects and faces (for exceptions, see Liu et al., 2012; Pylyshyn, 2004) that are analogous representations of identities in that they represent in a rather direct way their reference objects. The present study stands in contrast to the previous research in that the written words used as stimuli are abstract representations of real-life objects. That is, the word meaning can be accessed via decoding of the abstract letter symbols the written word consists of. The question we posed in the present study is whether such abstract identities can be readily accessed and utilized in dynamic tracking. As mentioned above, the call signals monitored by air traffic controllers represent a real-life example of a visual environment where moving abstract symbols are tracked. Moreover, the employed experimental paradigm was inspired by an in-depth task analysis of air traffic controllers and fighter pilots observing radar screens (Haavisto & Oksama, 2007).

In two experiments, we investigated the tracking of multiple moving words and pseudowords that were either short (Experiment 1 and 2) or long (Experiment 2). As noted above, written words significantly differ from stimuli typically tracked in earlier research, as they encode semantic identity information in an abstract form (via abstract letter codes), as opposed to moving faces or objects for which identity is inherent in the visual configuration of the stimuli (i.e., in an analog format). Since Paivio's (1971) seminal work, much evidence has been accumulated to demonstrate that stimulus materials differ in their recall and in recognition efficiency (e.g., pictures are superior to words; Madigan, 1983; Paivio, 1991; Stenberg, 2006). However, this issue has not been addressed before in the context of dynamic visual attention. If we are to replicate previous results of MIT (Oksama & Hyönä, 2008), we should find that existing words are tracked better than novel words. According to MOMIT (Oksama & Hyönä, 2008), as identification time is predicted to be intimately related to tracking efficiency, differences in the ease of identification (words are faster to decode than pseudowords) are assumed to translate into tracking accuracy.

Also stimulus complexity affects the ease with which identity information is encoded. Thus, if identity information is indeed utilized in multiple object tracking, stimulus complexity should also influence tracking efficiency. It is known that long words take more time to be identified than short words (see e.g., Calvo & Meseguer, 2002; Hyönä & Olson, 1995; Just & Carpenter, 1980; Kliegl, Grabner, Rolfs, & Engbert, 2004). Thus, if the ease of access of identity information has an impact on MIT, tracking moving words that are long should be more demanding than tracking short words.

In the two experiments, the MIT task developed by Oksama and Hyönä (2004, 2008) was employed. In this task, observers are asked to keep track of designated moving targets, that is, where each target is located at any one time. All the moving elements have distinct identities, analogous to many real-life tracking tasks. At the end of a trial, the movement stops, all objects are masked and one of the designated targets is probed (by flashing a frame around the target). The participant has to recognize the probed target by selecting the right one among all objects that moved (the objects are displayed next to each other on a separate screen).

Experiment 1

In Experiment 1, we examined the effect of stimulus familiarity on multiple identity tracking. We used short words that can be identified with a single eye fixation positioned on the stimulus. Stimulus familiarity was manipulated by using familiar words and pronounceable pseudowords as targets. By manipulating the stimulus familiarity we explored the role of LTM in tracking. By definition, a LTM representation is available for familiar words but not for pseudowords. Existing mental representations were predicted to result in faster identification and, as a consequence, in more efficient identity-location binding and identity tracking (Oksama & Hyönä, 2008). In order to assess the tracking capacity for verbal stimuli, the target set-size was manipulated (3-6 targets to be tracked). Moreover, in order to vary tracking difficulty, object speed was also manipulated using three speed conditions (slow, medium, fast)[1].

Method

Participants. Thirty participants (psychology students of the University of Turku) were recruited for the experiment. Their mean age was about 26. All participants had normal or corrected-to-normal vision. They gave a written informed consent for their participation.

Apparatus. The stimuli were presented on a 19-inch Samsung SyncMaster monitor with a refresh rate of 100 Hz, a resolution of 1280 by 1024 pixels controlled by NVIDIA GeForce FX5200 card, AMD Sempron 2400+, 1.66 GHz, 1.0 GB of RAM computer, and the E-prime software (Schneider, Eschman, & Zuccolotto, 2002a, 2002b). The software that generated the motion sequences was written in Visual Basic.

Task. The participants' task was to track the identity of targets (3-6 targets out of 6 moving objects) when they moved about on the screen. After the movement stopped, all the objects were masked and one of the target objects was probed by flashing a frame around it. After that the screen was cleared and a response screen appeared where all the six stimuli present during the tracking phase were arranged into an array of two rows and three columns. A response frame (100 x 100 pixels) surrounded the stimuli.

Participants were asked to select (point and click within a framed picture) the probed target as quickly as possible. They were asked to guess if they did not know the answer. A computer mouse was used to collect the responses. The mouse pointer was initially positioned in the middle of the stimulus array.

Procedure. Participants were tested individually. They were seated approximately 57 cm from the display; a chinrest was used to reduce head movements and to control the viewing distance. Participants were given written instructions prior to the experiment, which outlined the general procedure and explained the trial sequence.

At the beginning of each trial, all six objects were displayed for 1 s. After that, a black frame flashed on and off for ten times (flashing duration was 150 ms; total flashing time was 3000 ms) around the designated targets (3-6, depending on the trial). The objects then began to move in a random and continuous fashion around the screen. The frames were not visible during the movement phase. The participants tracked them for 5, 7, or 9 s, after which the movement stopped, and the objects were simultaneously masked with six Xs that were 60 pixels in width and 10 pixels in height (i.e., 1.5 deg x 0.3 deg). This was followed by the flashing of a black frame (75 x 75 pixels, 2 pixels in width, 1.9 x 1.9 deg) for five times (flashing duration was 150 ms; total flashing time was 1500 ms) around one of the targets (i.e., the probed item). After a response was given to the probe, the response screen was cleared and an inter-trial screen was presented. The next trial was initiated by the participant pressing the space bar, or after the maximum inter-trial interval (3000 ms) expired. Participants were provided with 10 practice trials; feedback was given after each response during the practice session. The entire session took about 100 min.

Stimuli. Two sets of six stimuli were used, six familiar words and six pseudowords. The familiar words and pseudowords were presented in separate blocks. The order of blocks was counterbalanced across participants so that half the participants performed the pseudoword trials first, while the other half performed the word trials first. Each participant completed two blocks of 72 trials for each experimental set (i.e., the familiar words and pseudowords), altogether 288 trials (12 repetitions of the 24 conditions). The order of trials was randomized separately for each participant within each block.

As the familiar words, we selected the following five-character, highly common Finnish words: ’ruusu’, ’takki’, ’kello’, ’tuoli’, ‘hauki’, and ‘kissa’ (the English equivalents are rose, coat, clock, chair, pike, and cat, respectively; note that pike is a very common fish in Finland). The words were selected to be concrete nouns that represent different semantic categories. Pseudowords were created from familiar words by changing the order of the syllables and sometimes changing some of the characters or the order of the characters within a syllable (in order to avoid creating other existing words). The pseudowords used were ‘kuiha’, ’ruolu’, ’hilpu’, ’olkel’, ’tiolu’, and ’sikso’ (all are pronounceable in Finnish). The stimuli were presented in Arial 16 font. The stimuli (11-15 pixels in height and 41 pixels in width) appeared in black on a white background, subtending a visual angle of 0.3—0.4 x 1.1—1.3 deg. The computer screen subtended a visual angle of 27 deg horizontally and 35 deg vertically.

In order to ensure that the pseudwords were indeed more difficult to recognize than the real words, we conducted a lexical decision experiment, where a group of 19 participants was asked to decide as fast as possible whether or not the presented letter strings were existing words. The lexical decision times demonstrated a reliable word type effect, *t*(18)=4.46, *p*<.001; the pseudowords took 55 ms longer on average to recognize as such than the real words (M=617 ms, SD=98 versus M=562 ms, SD=96).

To keep the total number of objects and the probability of guessing constant, the animation always consisted of six moving objects. The object combinations within a target set were constructed so that each object was selected equally often as a target (each object appeared eight times as a target in the target-set three, twelve times in the target-set four, etc.).

Movement sequences. Object speed was varied using three conditions, slow, medium, and fast. The experimental trials consisted of 167, 233, or 300 static frames presented one after another for 30 ms each (i.e., for 5, 7 or 9 s). In the slow speed condition, items moved a minimum of 1 and a maximum of 5.7 pixels per frame. As each frame had duration of 30 ms, the resulting item velocities were in the range of

0.9 to 4.8 deg/s (the average speed was 2.6 deg/s). In the medium speed condition, items moved a minimum of 3 and a maximum of 12.7 pixels per frame, so the resulting object velocity ranged from 2.6 to 10.9 deg/s (the average speed was 6.3 deg/s). In the fast speed condition, items moved a minimum of 7 and a maximum of 18.4 pixels per frame, thus the resulting velocities were in the range of 6.0 - 15.7 deg/s (the average speed was 10.7 deg/s). Note that in applied settings (e.g., team sports, road traffic, ATC, CDTI and other cockpit displays) speeds of the objects to be tracked in terms of VA/s vary greatly. The sources of this variance are the actual speeds of the objects, but also display scale and the speed and altitude of the airborne display platform ("ownship"). The present study did not investigate effects of local speed fluctuations on tracking performance (for such effects, see Meyerhoff, Papenmeier, Jahn, & Huff, 2016).

Initial object positions were generated at random. Movement direction for each object was chosen randomly from among the eight cardinal and ordinal compass directions. Each object was assigned a movement duration that was randomly selected from 7 to 37 in 30 ms increments (210–1110 ms), and speed, randomly selected from 1 to 5.7 pixels per frame in the slow condition; from 3 to 12.7 pixels per frame in the medium condition; or from 7 to 18.4 pixels per frame in the fast condition. The movement duration determined the time for how long the object maintained a certain direction and speed. When the movement duration expired, new random speed, direction and duration values were assigned to the object.

Random object motion created many possible collisions between objects and between objects and the edges of the display. Several actions were taken to avoid these collisions. First, an invisible cushion (which went through the vertices of the 75 x 75 pixel picture square) surrounded the objects. Thus, objects could not intersect each other. Second, before an object was moved to a new position, a possible collision to another object's cushion area and to the edges of the display was checked. If a collision was going to happen, a reverse direction was chosen to these objects (e.g., if one object was to the south of another object, the northern object moved to the north and the southern to the south). Also new random duration and speed values were assigned to the objects that were about to collide. Third, edge collisions were prevented in a similar manner: a new direction (randomly selected from the three possible reverse directions), speed, and duration were assigned to the objects. This procedure yielded a sequence of frames in which each element moved in a random, independent and continuous way for some period of time (210 - 1110 ms, or until a collision was about to happen), and then changed direction and speed abruptly and began to move in a new direction.

Thirty-six animation sequences (trajectory files) were generated and stored offline and were divided into two sets of eighteen files for the two experimental blocks. The same trajectory files were used for both word types (i.e., familiar words and pseudowords) and each set-size and object speed condition. One trajectory file was used four times in the block, one time in each target set, duration (the durations of 5, 7, 9 s were used to make tracking time less predictable; the data for the three durations were collapsed for the analyses) and speed condition. Thus, the trajectories within each target set and duration were different. The chosen target objects were different in different target sets. This procedure ensured that any differences in the tracking performance related to set-size and object type were not due to differences in the trajectory patterns (see Scholl & Pylyshyn, 1999, for a similar procedure).

Design. Target set-size (3-6) was manipulated to estimate how many moving word and pseudoword objects participants are able to track. Moreover, in order to further manipulate task difficulty, object speed was also varied. It has been shown that increase in speed lowers tracking accuracy (e.g., Alvarez & Franconeri, 2007; Holcombe & Chen, 2012; Oksama & Hyönä, 2008). Thus, there were three manipulated factors in the experiment: number of targets (3-6), familiarity of textual information (familiar words versus pseudowords), and object speed (2.6 deg/s, 6.3 deg/s, 10.7 deg/s; coined ‘slow’, ‘medium’, and ‘fast’). All the factors were within-participants variables. The dependent variable was the error rate in the tracking performance.

Results[2] and Discussion

Error rates were submitted to a linear mixed effects analysis using SPSS 25.0 (SPSS, Inc; Chigaco, Illinois). As fixed effects, set-size, word type and speed as well as their interactions entered in the model. Participants were entered as a random effect. Figure 1 shows the mean error rate in performance accuracy as well as their standard errors, as a function of target set-size and word type in the three different speed conditions. According to the Bayesian information criterion, the best fitting covariance structure was heterogeneous compound symmetry. Other covariance structures were tested, but they did not improve the model fit. Heterogeneous variance structure fits also well with the nature of the tracking task as it includes different difficulty levels with assumed different variances (e.g., small and easy-to-track set-sizes and object speeds produce smaller variances than large and difficult-to-track set-sizes and object speeds).

A significant main effect was found for the number of targets, F(3, 96.5) = 148.38 $p < .001$, performance deteriorated as the number of targets increased. More importantly, the main effect of word type was significant, F(1, 333.35) = 8.81, $p = .003$; tracking of familiar words produced less errors than that of pseudowords. The interaction between the number of targets and word type was significant, F(3, 306.60) = 3.15, $p = .025$. As is evident from Figure 1, performance deteriorated more strongly for pseudowords than for familiar words as a function of target set-size. The interaction partly reflects the floor effect observed for set-size 3. The main effect of object speed proved significant, F(2, 183.01) = 95.81, $p <.001$; performance deteriorated as the target velocity increased. Finally, the Set-Size x Object Speed interaction was statistically significant, F(6,213.85) = 17.72, $p < .001$. As is evident from Figure 1, the faster the targets moved and the more targets there were to be tracked, the steeper the tracking performance deteriorated. The Object Type x Object Speed and the 3-way interactions were non-significant, F(2, 337.85) = 2.00, $p = .14$, and F< 1, respectively.

Insert Figure 1 about here

We also tested the quantitative fit of the mathematical formulation of MOMIT to the data (see Appendix). It was conducted in a similar fashion as described in Oksama and Hyönä (2008). These simulations yielded a very nice overall fit with an estimate of about 250 ms for the binding refresh time with 19 ms difference between familiar and pseudowords (see Appendix, Table A, run 1—run 4). This refresh time estimate is highly similar to the duration of a single eye fixation made during text reading (for a review, see Rayner, 1998).

The results of Experiment 1 demonstrate that tracking of short words is influenced by word familiarity and number of tracked words and their relative speed. The word familiarity effect is consistent with the idea of active role of identity information (active LTM involvement in binding) in multiple identity tracking. It is also consistent with the MOMIT model (Oksama & Hyönä, 2008), which assumes it to reflect identification and binding time when (re-)establishing identity-location bindings for the to-be-tracked targets. Identification and binding take longer for pseudowords than familiar words particularly with larger set-sizes.

The set-size effect reflects capacity limitations in MIT. Observers were able to track three words with nearly flawless performance, while the performance started to deteriorate with four targets, particularly with faster object speeds (see Figure 1). An analogous Set Size x Object Speed interaction has been found for line drawings (Oksama & Hyönä, 2008). A possible explanation for this interaction is that with faster speeds and more targets the likelihood increases for the targets to change direction. In our experimental paradigm object collisions were prevented: when an object came close to another object or near the screen border, the object changed direction. Previous studies on multiple object tracking have demonstrated that trajectory changes hamper tracking performance (Ericson & Beck, 2013; Meyerhoff, Papenmeier, Jahn, & Huff, 2013).

Experiment 2

The results of Experiment 1 demonstrated that when observers track moving targets that are short familiar words or pseudowords they make use of identity information. This became apparent in that familiar words were easier to track than pseudowords. This familiarity effect was bigger for larger set-sizes. In Experiment 2, we studied whether observers are able to access identity information in MIT when complex textual identity information is used. Rather complex textual call signs are typical, for example, in ATC (e.g., SAS2459). Stimulus complexity was manipulated by varying word length. As the identity of long words consists of more visual elements (letters) than that of short words, their identification takes more time (see e.g., Calvo & Meseguer, 2002; Hyönä & Olson, 1995; Just & Carpenter, 1980; Kliegl et al., 2004). One reason for this is the fact that visual acuity drops off quite dramatically outside the foveal area; consequently, longer words cannot be identified with a single fixation, but two or more fixations are often needed for their identification (see e.g., Hyönä, 1995, but see also McDonald, 2006). Assuming that (1) identity-location bindings are refreshed one at a time with the help of focal attention (Oksama & Hyönä, 2008) and that (2) bindings are lost if not frequently refreshed, in MIT observers are likely to have only limited time to update the bindings for the to-be tracked objects. These time constraints may in turn lead to a situation where the identity of complex visual stimuli cannot be fully accessed. This may be particularly the case when the object identities are unfamiliar to the observer, as is the case with long pseudowords, and with larger target set-sizes.

In Experiment 2, we used as stimuli familiar short 5–character words (the ones used in Experiment 1) and longer 9-10 –character words as well as length-matched, long and short pronounceable pseudowords. Long words were presented on two lines (separated by a hyphen at the syllable boundary). This was done for two reasons. First, long words and pseudowords were equated in this way with the short stimuli with respect to the horizontal extent. Thus, when focally attended the whole word fell on the foveal region of the eye. Second, the displays used, for example, in ATC typically contain information presented on multiple lines. It should also be noted that in running text of Finnish words are often hyphenated, made possible by the syllable structure of Finnish without significantly hindering word recognition. Target set-size was varied from three to six but only one speed condition was used (i.e., the medium speed condition of Experiment 1).

As discussed above, MOMIT (Oksama & Hyönä, 2008) predicts that both familiarity and complexity of textual information influence the tracking performance. Due to the additional time needed to identify long targets and update their identity-location bindings, MOMIT predicts that tracking performance should be poorer for long than short words and also poorer for pseudowords than familiar words. However, two contrasting predictions can be made concerning the role of identity information in tracking of moving long words. According to the first alternative, accessing the identities of long words (especially those of long pseudowords) takes considerable amount of time, which in turn may significantly disrupt their tracking. Thus, a sizeable familiarity effect is predicted that is greater than the one observed for short words in Experiment 1. This prediction is based on the assumption that identification time can increase (or refresh rate can slow down) without any limits. However, in dynamic environments this assumption may not hold (see above). Thus, according to the second alternative, dynamic environments constrain observers to limit the time allotted to serially refresh and update identity-location bindings. For example, observers may opt for making only one eye fixation on the to-be-refreshed targets. If so, there may be insufficient time to fully access the identity information for longer words and pseudowords during each refresh cycle, which would lead to a negligible familiarity effect. This may be particularly the case with larger set-sizes. Both hypotheses predict a Word Length x Word Type interaction and/or a 3-way interaction between word length, word type and set-size.

For Experiment 2, we recruited a sample of Finnish Air Force applicants. Fighter pilots need excellent executive and attentional skills. Although only a small minority of the applicants were eventually chosen to be trained as fighter pilots, the trainee requirements heavily constrain who will apply for the Air Force. The experiment was carried out as a part of the selection process, although it was not part of the test package used in making the final selection (the participants knew it). All in all, the study participants were well equipped and motivated to successfully carry out the MIT task.

Method

Participants. One hundred participants took part in the experiment. Their mean age was about 20. The experiment was conducted as a part of a large test battery that was administered to Finnish Air Force applicants. Participants were told that the experiment was not part of the test package used in the air force selection process and they were free not to participate in it if they wished. However, all applicants agreed to take part in the experiment and gave their informed consent. They all had normal, uncorrected vision. A pre-selection was made on the basis of their previous school achievement (in mathematics and English).

Apparatus. The stimuli were presented on 19-inch Eizo FlexScan F730 monitors with a refresh rate of 100 Hz, a resolution of 1280 by 1024 pixels controlled by Matrox G400 cards, Pentium 3, 500 MHz, 128 Mt, RAM computers.

Stimuli. The short familiar word and pseudowords were the same 5-character words as in Experiment 1. The long targets were the following common 9-10-character Finnish words ’ambulanssi’, ’kalastaja’, ’kalenteri’, ’kuningatar’, ‘pyöräilijä’, and ‘ravintola’ (the English equivalents are ambulance, fisherman, calendar, queen, cyclist, and restaurant). The long pseudowords (all pronounceable in Finnish) were created in the same way as the short pseudowords in Experiment 1; they were ‘elsiokause’, ‘halsejoune’, ‘kolketika’, ‘niokarusve’, ‘osvatuosi’, and ‘seivomike’. Long words were presented on two lines (separated by a hyphen positioned at a syllable boundary) to equate them for the horizontal extent with the short words and make them comparable to information formats encountered in ATC. It also allowed us to use the same movement trajectories for short and long stimuli. Long and short words were presented in the same font and font size. Finally, it should be noted that as words in Finnish tend to be long, in normal text words often appear hyphenated; thus, our participants are accustomed to read hyphenated words. The long words subtended a visual angle of 0.8—1.0 deg x 1.3—1.5 deg (31—39 pixels in height and 52—60 in width with the hyphen; without the hyphen, their width was 45–55 pixels = 1.2 deg – 1.4 deg). The computer screen subtended a visual angle of 25 deg horizontally and 32 deg vertically. Visual masks (75 x 75 pixels) that replaced the words at the end of movement phase consisted two rows of six Xs that were 60 pixels in width and 32 pixels in height, i.e., 1.5 deg in width x 0.8 in height.

In order to confirm that the long hyphenated pseudowords take longer to recognize than long hyphenated words, a lexical decision experiment was run with 19 participants. This pretest indeed demonstrated the assumed difference in the ease of recognition, *t*(18)=3.97, *p*<.001. There was a 55 ms advantage in average in recognition latency for real words (M=579 ms, SD=68 vs. M=634 ms, SD=78).

Design. There were three manipulated factors in the experiment: number of targets (2-6), familiarity of words (existing words versus pseudowords), and stimulus length (5 versus 9-10 characters). Set-size and stimulus familiarity were within-participants variables while word length was a between-participants variable. There were 24 trials in each of the 10 conditions of the within-participants variables. The order of the pseudoword and word trials was counterbalanced across participants so that half of the participants performed the word blocks first, while the other half of the participants performed the pseudoword blocks first. As in Experiment 1, the dependent variable was response time to probes of the objects.

Procedure. The procedure and the experimental task were identical to those of Experiment 1. Object speed was comparable to the medium-speed condition of Experiment 1.

Results

As in Experiment 1, a linear mixed effects analysis was performed on the error rates. As fixed effects, set-size, word type and word length as well as their interactions were entered in the model. Participants were entered as a random effect. Figure 1 shows the mean error rate in performance accuracy with standard errors, as a function of target set-size and word type in the two word length conditions. According to the Bayesian information criterion, the best fitting covariance structure was again heterogeneous compound symmetry. Other covariance structures were tested, but they did not improve the model fit.

A significant main effect was found for the number of targets, F(4, 190.07) =240.65, p < .001; performance deteriorated as the number of targets increased. The main effect of word length also proved significant, F(1,104.74) = 7.19, p = .009; long words produced more errors than short words. However, the main effect of word type missed significance (F (1,606.15)=3.69, p=.055. The main effects were qualified by two reliable interactions (the other interactions were clearly non-significant, F<1.26), Set-Size x Word Length, F(4,190.07) = 3.31, p = .012, and Word Type x Word Length, F(1,606.15)=4.03, p = .045. As can be seen from Figure 2, the nature of the former interaction is such that performance deteriorated steeper for long than short words as a function of target set-size. The interaction also partly reflects the floor effect in tracking accuracy obtained for set-size 2.

To examine in more detail the nature of the Word Type x Word Length interaction, we conducted a separate LMM for short and long words. In the analysis of short words, a significant main effect emerged for the number of targets, F(4,97.71) = 128.35, p < .001, and word type, F(1,300.67) = 9.52, p = .002, but their interaction was not significant, F(4,176.63) = 1.98, p = .10. That is, the word type effect observed in Experiment 1 was replicated for short words. However, in the analysis of long words, a significant main effect was found only for the number of targets, F(4,92.44) = 119.16, p < .001, but the main effect of word type and the interaction were clearly non-significant (both F < 1). That is, a word type effect was not established for long words (see Figure 2).

Insert Figure 2 about here

According to MOMIT (Oksama & Hyönä, 2008), MIT is achieved by switching the focal attention between the to-be-tracked targets. Focal attention is needed to frequently refresh the identity-location bindings that would otherwise decay from the episodic buffer. To estimate the refresh time for long words and pseudowords, we ran MOMIT simulations for the long word data (see Appendix, Table A, Run 5 and 6). These simulations yielded an estimate of about 300 ms for the refresh time with only 1 ms difference between familiar and pseudowords. The time estimate is clearly shorter than the time needed to make two fixations on the word during reading (about 450-500 ms), and it is also about 50 ms longer than the estimate obtained for the refresh time of the short words (see Experiment 1). As discussed below, these refresh time estimates may explain why no familiarity effect was observed with long stimuli.

Discussion

Experiment 2 established three main findings. First, we observed dynamic tracking of long words to be significantly poorer than that of short words. Second, the familiarity effect (real words versus pseudowords) established for short words in Experiment 1 was replicated in Experiment 2 using short and long words. Third, both the word length and word type effect were more robust with larger set-sizes. The pattern of data is illustrated in Figure 2.

The poorer tracking performance for long than short words is readily explained by MOMIT (Oksama & Hyönä, 2008) as a longer refresh time for more complex than less complex items (see above and Appendix). With the longer refresh times, bindings stored in the episodic buffer run the danger of becoming outdated, which in turn will result in an accurate and efficient shifting of focal attention between targets becoming more difficult to carry out leading to some bindings to be lost either permanently or temporarily (Pinto et al., 2010, note that when tracking objects with unique identities a missing target can in principle be added on the fly to the tracked target set).

With respect to the effect of stimulus familiarity for long textual stimuli, we proposed two contrasting predictions regarding the tracking of targets whose identities take particularly long to access, as is the case with long pseudowords. According to the first alternative, such complex stimuli would result in the poorest performance of all the tested conditions. According to the second alternative, full access of the long items may not be possible in a dynamic environment resulting in a lack of a word type effect for long items. The results of Experiment 2 turned out to be consistent with the second alternative.

According to this view, the stringent time constraints induced by the MIT task allow insufficient time to fully access the identity of long items, leading to the tracking performance being no better for familiar than unfamiliar targets. Recall that the long words were presented on two lines. In order to fully access the identity information of long words, it is possible that two eye fixations are needed (one on each line). Although this was not tested directly, the MOMIT simulations computed to estimate refresh time for long items yielded estimates that were significantly shorter (about 300 ms) than the time needed to make two fixations (about 450-500 ms). Moreover, the estimates were almost identical for familiar words and pseudowords – a finding consistent with the null result of familiarity in tracking accuracy. Thus, it seems that with long words observers are required to trade off between full target identification and maximum accuracy in tracking. The observers appear to behave as if they get a better pay-off by keeping the refresh times short even if this happens at the expense of not fully accessing the target identities. In other words, they do not slow down their refresh pace to about 500 ms/target, which would be needed to fully access the identity of long words. It might mean that they only attend to what is written on the first line. As this does not reveal the word’s full identity (it appears similar to a short pseudoword), observers cannot make use of LTM representations in keeping bindings active. Thus, no effect of familiarity is observed for long words and tracking of long words is poorer than that of short words.

As may be noted from Table 2, the performance for short pseudowords tends to be better than that for familiar long words. This suggests that the observers may not just attend to the first line of long words (i.e., trying to treat them all as short pseudowords) but periodically make an additional effort to more fully attend to them, which somewhat lengthens refresh times (from 250 to 300 ms as indicated by the simulations) and consequently deteriorates the tracking performance for existing long words.

General Discussion

The present study investigated the role of identity information in tracking multiple moving verbal stimuli. Our experimental task simulated dynamic visual environments where the observer needs to keep track of where objects of interest are located from moment to moment. What is crucial in this task is that the identity of the target objects is conveyed by a letter string (making up either a real word or a pseudoword). Thus, the observer needs to bind this identity information with its current spatial position. As the target objects constantly move, the identity-location bindings need to be frequently updated to maintain adequate situation awareness (SA; Endsley, 1995). This type of task environment bears close resemblance to the tasks of air traffic controllers. They typically must maintain good SA of moving aircraft so that potential collisions will be prevented in a timely manner and all aircraft will be safely guided to their intended destinations. The identity of each aircraft is marked by call signals that are a combination of letters and numbers (e.g., SAS6609).

We examined in two experiments the human capacity to dynamically track textual objects that varied in number, familiarity and complexity. Contrary to the claims made in the seminal research on MOT (Pylyshyn, 2004) but in line with more recent evidence (e.g., Botterill et al., 2011; Horowitz et al., 2007; Liu et al., 2012; Makovski & Jiang, 2009; Ren et al., 2009), the present study demonstrated that identity information significantly influences the tracking performance. First, we found that the complexity of textual identity information affects tracking: short words are easier to track than long words. Second, we observed that familiarity of identity information affects tracking of short words; in other words, short familiar words were easier to track than short pseudowords. However, such a familiarity effect (see also Oksama & Hyönä, 2004, 2008; Pinto et al., 2010) was not observed for long words; long familiar words and long pseudowords were equally difficult to track (see Figure 1).

MOMIT: Theoretical Framework to Explain the Results

According to MOMIT (Oksama & Hyönä, 2008), dynamic identity–location bindings are created and updated by using serial attention and temporary memory buffers. In order to maintain an up-to-date representation of the whereabouts of the target objects, focal attention is serially switched between the moving elements. The purpose of attention shifts is to continuously update and refresh identity-location bindings. When focal attention is directed to a target element, the attended target is identified and the identity-location binding is temporarily stored in the episodic buffer. Hence, the faster target identity is bound to a new location, the sooner attention is disengaged from one target and switched to the next one. Moreover, the faster attention shifts are made between targets, the shorter distance the targets have moved away from their previous locations (location information is assumed to be stored in the visuospatial working memory), when a new attempt is made to revisit them to update and refresh the identity-location binding. As stored location coordinates are assumed to be utilized in programming attention shifts, a smaller object movement means less location error between the stored and the present location of the target, which in turn leads to more accurate attention shifts between the targets. In other words, it is more likely that focal attention is directly shifted to the intended object rather than to a non-target or a non-intended target, which in turn means that a time-consuming target search may be avoided and the probability of loosing the binding is diminished. Hence, the model predicts that the time spent in serially refreshing and updating identity-location bindings is intimately related to the tracking efficiency. The faster the bindings are serially refreshed, the better the tracking performance is.

As becomes apparent from the above description, MOMIT assumes identity information and stimulus properties of objects to play an active role in dynamic binding. The influence of stimulus properties is mediated by the time taken to identify a target and bind location information to the perceived identity. MOMIT predicts that all the factors that influence the target identification and binding time, also influence the capacity to maintain multiple identity-location bindings for moving objects. Such factors are, for example, object familiarity and complexity. As more time is needed to identify an unfamiliar than a familiar target, familiar objects should be easier to track than unfamiliar ones. An analogous argument is put forth for object complexity.

As argued above, the observed word familiarity effect for short words is consistent with the functional architecture of MOMIT. This is also evident in the good quantitative fit of MOMIT to the data of Experiment 1. The observed word length (i.e., complexity) effect is also consistent with the underlying assumptions of MOMIT. As less complex stimuli (5-character words) are identified faster than more complex stimuli (9-10-character words), MOMIT predicts tracking to be better for short than long words. However, the absence of a familiarity effect for long words seems inconsistent with MOMIT. As long pseudowords take longer to identify than long familiar words, MOMIT predicts a difference in tracking accuracy between the two stimulus types. However, this prediction is based on an assumption that target identification/binding time can increase without any limit and full access to target identities is always attempted in dynamic situations. Instead, the pattern of data and the model simulations computed for the long word data suggest that this assumption may be wrong (at least with the rather fast moving stimuli used here). In addition, the simulations estimated the refresh time of identity-location bindings to be around 300 ms for long words, which is likely to be too short for their full identification. As they were presented on two text lines, their identification probably requires two eye fixations whose summed duration would be approximately 400-500 ms. Thus, we argue that observers could not fully access the target identities and were thus forced to treat the long words more like pseudowords, which in turn prevented them from using LTM representation to aid in the maintenance of identity-location bindings.

As noted above, the data for tracking long words and their simulation by MOMIT bring fore a new phenomenon in multiple identity tracking (see also Oksama & Hyönä, 2016). When identity information is complex, observers are faced with a dilemma between full access of target identities on one hand and maximum tracking efficiency (i.e., minimizing the refresh time of identity-location bindings) on the other hand. Such a trade-off has previously been identified by Moray (1984). When several channels of dynamic information have to be monitored, observers are assumed to tradeoff between sampling rate and observation time. Moray (1984) speculated that the trade-off can be solved either by shortening the observation time and increasing the sampling rate or by delaying the sampling rate and keeping the observation time constant.

Recently, Oksama and Hyönä (2016) registered observers' eye movements while they were tracking multiple objects with distinct identities (2-5 line-drawings of familiar objects). Thus, they were able to directly examine the trade-off between sampling rate and observation time. They found that the participants used both strategies outlined by Moray (1984). When the set-size increased from 2 to 4, participants both increased sampling rate and decreased observation time. At set-size 4 the observation time and sampling rate levelled off: fixation duration asymptoted at 220 ms and the visual sampling rate (fixations per second) reached its maximum value of 3.7 Hz. In other words, there seems to be a limit for the trade-off.

The present data and their model simulations suggest that observers trade off by minimizing the observation times (approximately one fixation per target) at the expense of full access to complex target identities. This strategy makes sense in dynamic situations: by using it observers are able to maintain an accurate spatial representation of multiple targets while sacrificing accurate identity information. A full access to target identities may not be necessary as long as target identities can be kept separate from each other. This can be achieved by partial access, for example by accessing only the first part of long words (i.e., with the present stimuli, what is written on the upper line). A slower refresh pace would probably yield full access to two or three target identities with a danger of loosing the identity-location binding for the other targets. It is also noteworthy that the poor performance for tracking 5 or 6 verbal items (see Figure 1) may be partly explained by the limit in trade-off observed by Oksama and Hyönä (2016). Yet, it should be noted that this kind of trade-off between full identification and maximum tracking efficiency is not implemented in the present version of MOMIT.

We have recently updated some of MOMIT's general principles (Li, Oksama, & Hyönä, 2019). However, unlike the original model, it has not been mathematically formulated. Thus, the present data cannot be simulated by the new version. The most relevant modification in the present context is that

MOMIT 2.0 does not assume the dynamic identity tracking to rely on temporarily stored identity-location bindings but rather on proto-objects that are formed in a course-to-fine manner. Low-resolution proto-objects are formed in parallel from location and basic featural information (for using surface feature information in multiple object tracking, see e.g. Papenmeier, Meyerhoff, Jahn, & Huff, 2014). Such representations might be sufficient for tracking objects whose identities are distinguishable via elementary visual features (e.g. disks of different color). On the other hand, for stimuli such as written words that are combinations of elementary visual forms (letters formed from curved and straight lines in different angles), to keep them separate from each other it is highly likely that the resolution of the formed proto-objects needs to be enhanced by serially attending to each target in turn. Yet, the construction of high-resolution representations does not necessarily imply that word meanings are activated. For successful tracking, it suffices that the formed representations are distinguishable from each other.

Alternative Account Based on Short-Term Memory Capacity

An alternative theoretical account of the present results is to consider them as a reflection of capacity limitations in short-term storage. In Baddeley's (1986) theory of working memory, the phonological loop keeps temporarily active task-relevant verbal stimuli. The capacity of the phonological store is constrained by the amount of verbal information that can be subvocally articulated within two seconds. Specifically, word length constrains its capacity (Baddeley, Thomson & Buchanan, 1975): more short than long words can be temporarily retained in the phonological store. This finding may be readily applied to the present results. When using their full identities, the articulation time for six words or pseudowords exceeds the capacity limit. In order to compensate for this capacity limitation, observers may use abbreviations of long stimuli (e.g., what is written on the first line). This may partly explain why there was no difference in tracking capacity between long words and pseudowords. However, this account can only explain the results for large set-sizes (5 or 6 targets), but it cannot explain why there is no difference with smaller set-sizes of 2 or 3 targets whose identities when articulated would not exceed the capacity of the phonological store.

Storage capacity also features in MOMIT. In the model's mathematical simulations, Oksama and Hyönä (2008) tried out fixing the target refresh time and allowed storage capacity to vary. The simulation yielded a good fit to the data with a 0.2 difference in tracking familiar objects and pseudo-objects. Thus, the storage capacity account cannot be ruled out. Yet, it cannot explain target set-size effects within capacity limits, nor object speed effects observed, for example in Experiment 1. Moreover, as pointed out above, it cannot explain the lack of target familiarity effect for long words even with small set-sizes. All in all, a plausible suggestion is that the tracking capacity of multiple moving words is jointly determined by refresh time and storage capacity.

Caveats

In comparing tracking of short and long words in Experiment 2, we decided to equate the stimuli for their horizontal extent. This way we were able to use identical movement trajectories for the short and long stimuli. This resulted in the long words being presented in two lines. Above, we have argued that given the time constraints of the tracking task observers may have only had time to attend to the first line in the long words, which presumably wiped out the word type effect for long words that was observed for short words. Even though we consider it quite plausible, the conclusion should only we considered tentative, as the present experimental design is insufficient for proving it. It would have required a design where word length and presentation type would have been orthogonally manipulated and combined with the eye movement registration. All in all, at present it is not clear whether the inability to access the full identity of long words during dynamic tracking was due to time constraints, stimulus presentation, or a combination of the two.

Applications

The present results demonstrate that a manipulation of identity information can affect the efficiency of visual tracking and the maintenance of dynamic SA. We found that simple textual identities are easier to track than complex ones. We also found that verbal stimuli that have an LTM representation are somewhat easier to track than those that lack a representation in LTM. Finally, written words, particularly long words are more difficult to track compared to pictorial stimuli used in prior MIT research. These results have clear implications on human performance in MIT tasks, such ATC. Controllers' tracking—and consequently their Level 1 situation awareness (Endsley, 1995)—may be impeded by complex stimuli (e.g., long verbal codes that are presented on multiple lines, such as aircraft data blocks on ATC displays) due to the observer not having sufficient time to encode target identities and due to the danger of losing targets by the lengthened refresh times. In ATC, inaccurate SA would lead to yielding clearance to a wrong aircraft. Presently, call sign confusion is a real problem in radio communications in ATC, resulting in repeated communications and potentially very serious consequences (e.g., near-midair collisions) (UK Civil Aviation Authority, 2000), but problems of controller confusing visually presented call signs on displays have not been investigated (see however, Schneider, Healy & Barshi, 2004, for effects related to navigational commands). It seems reasonable to assume that call sign similarity significantly contributes to controller's mental workload. In military aviation, target identity confusion would mean severe problems to maintain awareness of the whereabouts of the other aircraft in the fleet as well as potential enemy aircraft. With respect to sport games like soccer or ice hockey, what –where -binding errors of this kind mean that passes go occasionally to opponent team members instead of the player's team mate. This kind of passes can be observed time-to-time even in professional sport games (a defender passes a ball straight to his opponent at his/her defense area).

Notice also that recent developments in international air traffic have made the task of the controller even more difficult. Traditionally, a large proportion of flights generally have operated on constant, recurrent routes with little variation (in terminal areas, depending on runway configurations in use). Recently however, pilots have become less dependent on ATC, so they are also exposed to a dynamic tracking task similar to the one of controllers. Conversely, increased autonomy of pilots in choosing their routes and maneuvers to avoid conflicts with other traffic results in increasing uncertainty of traffic flows for the controller, effectively undermining the foundation of controllers' SA. Our task of tracking unfamiliar complex target identities on unpredictable trajectories may therefore be closer to the reality of controllers' tasks in the near future than might at first appear.

Our results suggest that the potential of human error is substantial in tracking complex identities. Based on the simulations of the MOMIT model, we speculate that observers may not have enough time to both fully process all the complex identities and keep track of their locations at the same time, when the speed of movement is rather high. The observer has to choose between the full identification and accurate spatial representation of the targets. Our data suggest that observers may choose accurate spatial representation at the expense of full identification of complex stimuli. As discussed above, this strategy makes sense in some dynamic environments, but it means that the observer's temporary mental model may not fully correspond to the real situation, that is, operators may not know exactly what object is where. As a consequence, the observer is susceptible to an identity-location binding error, or an action that is meant for target A is instead subjected to object B. It is important to understand and model such deficiencies in human performance in complex, dynamic, and high-risk tasks.

Finally, the account for maintenance of dynamic identity–location bindings by serial attention and temporary memory buffers as provided by MOMIT may offer important means for further study of SA and discovery of specific mechanisms behind this ill-defined but critical construct (Sarter & Woods, 1991). An equally critical as well as equally elusive construct in many complex and dynamic tasks and task environments is mental workload. One of the primary drivers of mental workload is time pressure (Hendy, 1995; Hendy, Liao, & Milgram, 1997, Hancock & Chignell, 1988; Laudeman & Palmer, 1995). Time pressure can be defined as the ratio of time required to time available to perform a task (e.g., Hendy, 1995), or task load imposed on the operator. The MOMIT accounts for the time required for refreshing dynamic identity-location information can therefore be used to both measure task load in MIT tasks as well as establish limits for human tracking capabilities. Application of MOMIT in increasingly realistic task settings offers exciting future avenues of research.

References

Alvarez, G. A. Franconeri, S. L. (2007). How many objects can you track Evidence for a resource-limited attentive tracking mechanisms. *Journal of Vision*, 7, (13):14, 1–10, http://journalofvision.org/7/13/14/, doi: 10.1167/7.13.14.

Baddeley, A.D. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A.D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*, 575-589.

Botterill, K., Allen, R., & McGeorge, P. (2011). Multiple-object tracking: The binding of spatial location and featural information. *Experimental Psychology, 58*, 196-200.

Calvo, M. G., & Meseguer, E. (2002). Eye movements and processing stages in reading: Relative contribution of visual, lexical and contextual factors. *Spanish Journal of Psychology, 5*, 66–77.

Cohen, M.A., Pinto, Y., Howe, P.D.I., & Horowitz, T.S. (2011). The what-where trade-off in multiple-identity tracking. *Attention, Perception & Psychophysics, 73*, 1422-1434.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.

Ericson, J.M., & Beck, M.R. (2013). Changing target trajectories influences tracking performance. *Psychonomic Bulletin & Review, 20*, 951-956.

Haavisto, M., & Oksama, L. (2007). Kognitiivisen kuormituksen arviointi: Esimerkkinä hävittäjälentäjän tehtävä- ja kuormitusanalyysi [Cognitive workload assessment: A fighter pilot's task and workload analysis as an example]. *Työ ja ihminen,1*, 17-29.

Hancock, P. A., & Chignell, M. H. (1988). Mental workload dynamics in adaptive interface design. *IEEE Transactions on Systems, Man, and Cybernetics, 18*(4), 647-658.

Hendy, K. C. (1995). Situation awareness and workload: Birds of a feather? *AGARD AMP Symposium on Situation Awareness: Limitations and Enhancements in the Aviation Environment* (21-1–21-7). Brussels, April 24-28, 1995.

Hendy, K. C., Liao, J., & Milgram, P. (1997). Combining time and intensity effects in assessing operator information processing load. *Human Factors, 39*(1), 30-47.

Holcombe, A. O., & Chen, W. Y. (2012). Exhausting attentional tracking resources with a single fast-moving object. *Cognition, 123*, 218–228.

Horowitz, T. S., Klieger, S. B., Fencsik, D. E., Yang, K. K., Alvarez, G.A., & Wolfe, J.M. (2007). Tracking unique objects. *Perception & Psychophysics, 69*, 172-184.

Howe, P.D.L., Cohen, M.A. Pinto, Y., & Horowitz, T.S. (2010). Distinguishing between parallel and serial accounts of multiple object tracking. *Journal of Vision, 10, 11*. doi:10.1167/10.8.11

Hummel, J. E. (2003). The complementary properties of holistic and analytic representations of object shape. In G. Rhodes & M. Peterson (Eds.), *Perception of faces, objects, and scenes: Analytic and holistic processes* (pp. 212-234). New York: Oxford University Press.

Hyönä, J. (1995). Do irregular letter combinations attract readers' attention? Evidence from fixation locations in words. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 68-81.

Hyönä, J., & Olson, R.K. (1995). Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1430-1440.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329–354.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of Attention* (pp. 29–61). New York: Academic Press.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of the object files: Object-specific integration of information. *Cognitive Psychology, 24,* 174–219.

Kirchner, H., & Thorpe, S. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research, 46*, 1762-1776.

Kliegl, R. Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*, 262-284.

Laudeman, I. V., & Palmer, E. A. (1995). Quantitative measurement of observed workload in the analysis of aircrew performance. *International Journal of Aviation Psychology, 5*(2), 187-197.

Li, J., Oksama, L., & Hyönä, J. (2019). Model of multiple identity tracking (MOMIT) 2.0: Resolving the serial vs. parallel controversy in tracking. *Cognition, 182*, 260-274.

Liu, T., Chen, W., Liu, C.H., & Fu, X. (2012). Benefits and costs of uniqueness in multiple object tracking: The role of object complexity. *Vision Research, 66*, 31-38.

Makovski, T., & Jiang, Y.V. (2009). Feature binding in attentative tracking of distinct objects. *Visual Cognition, 17*, 180-194.

Madigan, S. (1983). Picture memory. In J. C. Yuille (Ed.), *Imagery, memory, and cognition: Essays in honor of Allan Paivio* (pp. 65-89). Hillsdale, NJ: Erlbaum.

McDonald, S.A. (2006). Effects of number-of-letters on eye movements during reading are independent from effects of spatial word length. *Visual Cognition, 13*, 89-98.

Meyerhoff, H. S., Papenmeier, F., Jahn, G., & Huff, M. (2016). A single unexpected change in target – but not distractor motion impairs multiple object tracking. *i-Perception, 4*, 81-83.

Meyerhoff, H. S., Papenmeier, F., Jahn, G., & Huff, M. (2016). Not FLEXible enough: Exploring the temporal dynamics of attentional reallocations with the multiple object tracking paradigm. *Journal of Experimental Psychology: Human Perception and Performance, 42*, 776–787.

Moray, N. (1984). Attention to dynamic visual displays in man-machine systems. In R. Parasuraman & D. Davies (Eds*.), Varieties of attention* (pp. 485–513). Orlando: Academic Press.

Oksama, L., & Hyönä, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Visual Cognition, 11*, 631–671.

Oksama, L., & Hyönä, J. (2008.). Dynamic binding of identity and location information: A serial model of multiple identity tracking. *Cognitive Psychology, 56*, 237-283.

Oksama, L., & Hyönä, J. (2016). Position tracking and identity tracking are separate systems: Evidence from eye movements. *Cognition, 146*, 393-409.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, & Winston.

Paivio, A. (1991). *Images in mind: The evolution of a theory*. New York: Harvester Wheatsheaf.

Papenmeier, F., Meyerhoff, H.S., Jahn, G., & Huff, G. (2014). Tracking by location and features: Object correspondence across spatiotemporal discontinuities during multiple object tracking. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 159-171.

Pinto, Y., Howe, P.D.L., Cohen, M.A., & Horowitz, T.S. (2010). The more often you see an object, the easier it becomes to track it. *Journal of Vision, 10(10),* 1-15.

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition, 32*, 65–97.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80*, 127–158.

Pylyshyn, Z.W. (2004). Some puzzling findings in multiple object tracking: I. Tracking without keeping track of object identities. *Visual Cognition, 11*, 801-822.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3*, 1–19.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.

Ren, D., Chen, W., Liu, C.H., & Fu, X. (2009). Identity processing in multiple-face tracking. *Journal of Vision, 9(5),* 1-15.

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *International Journal of Aviation Psychology, 1*(1), 45-57.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002a). *E-prime user's guide*. Psychology Software Tools Inc: Pittsburgh.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002b). *E-prime reference guide*. Psychology Software Tools Inc: Pittsburgh.

Schneider, V., Healy, A. & Barshi, I. (2004). Effects of instruction modality and readback on accuracy in following navigation commands. *Journal of experimental psychology: Applied, 10*(4), 245–257.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology, 38*, 259–290.

Scholl, B.J., Pylyshyn, Z.W., & Franconeri, S.L. (1999). When are featural and spatiotemporal properties encoded as a result of attentional allocation? *Investigative Ophthalmology and Visual Science, 40,* 797.

Smith, M.C., & Magee, L.E. (1980). Tracing the time course of picture-word prceossing. *Journal of Experimental Psychology: General, 109,* 373-92.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 174–215.

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology, 18*, 813-847.

UK Civil Aviation Authority Safety Regulation Group (2000). *Aircraft call sign confusion evaluation safety study* (CAP704). CAA.

Watson A. B., & Pelli D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, 33, 113–120.

Footnotes

1. Another possibility to examine effects of stimulus type, set-size and object speed is to use a staircase procedure (Howe, Cohen, Pinto, & Horowitz, 2010; Watson & Pelli, 1983)**.** In this procedure, a performance criterion is predefined (e.g., 75% correct), and experimental conditions are then determined that yield that performance level. If short and long words would have been equated for tracking accuracy using the staircase procedure and a speed manipulation, MOMIT would then predict better tracking of words than pseudowords for both short and long words. This is because the model assumes the tracking efficiency to be in part governed by the speed of target identification.

2. The data of the two experiments can be obtained from the second author, Lauri Oksama (oksama@utu.fi), upon request.

Appendix: Model Fitting

In order to test MOMIT's goodness of fit, we quantitatively fitted the mathematically formalized MOMIT (see Oksama & Hyönä, 2008, for further details) to the tracking accuracy observed Experiment 1. The most important parameters of the model are *s* (average time required to refresh a binding) and *m* (binding capacity). We report three measures of goodness of fit: R^2, standard errors, and confidence intervals. Model fitting and parameter estimation were done separately for the data for familiar words and pseudowords, because MOMIT assumes *s* to be different for the two word types.

Estimates of model fit and the best fitting parameters are presented in Table A. As it appears from Table A, the MOMIT equation with two free parameters (Run 1 and 2) yields a very nice fit to the data of both stimulus types in Experiment 1 with small standard errors, sufficiently narrow confidence intervals, and over .9 coefficient of determinations. The best fitting values for both familiar words (*s* = 256 ms and *m* = 3.6) and pseudowords (*s* = 249ms and *m* = 3.3) are psychologically plausible. The estimate of *m* is close to 4 proposed by Luck and Vogel (1997); the estimate of *s* is also within a plausible range (a single eye fixation on a word typically lasts about 250 ms).

The next step was to experiment with the refresh time parameter s derived from the MOMIT's architecture, where a different value of *s*, but a common *m* is assumed for familiar words than pseudowords. Thus, we fixed *m* to 3.44 (an average of 3.58 and 3.30 from the two previous runs) for both stimulus types and let *s* vary freely. The results are presented as Run 3 and 4 in Table A. The best fitting *s* for familiar objects was 243 ms and for pseudowords 262 ms — a difference of 19 ms. The model fits are very good with narrow confidence intervals, small standard errors, and a high R^2. According to MOMIT, this difference reflects the relative ease of establishing an identity for familiar versus unfamiliar items.

Finally, we ran MOMIT simulations for the long word data of Experiment 2. This experiment had only one speed condition, so there are fewer data points to fit. To avoid overfitting and too wide confidence intervals, we fixed parameter *m* to the average value derived from Experiment 1 and let only *s* to vary freely. These simulations yielded a very nice fit to the data with an estimate of about 300 ms for the refresh time with only 1 ms difference between familiar and pseudowords 2 (see Run 5 and 6).

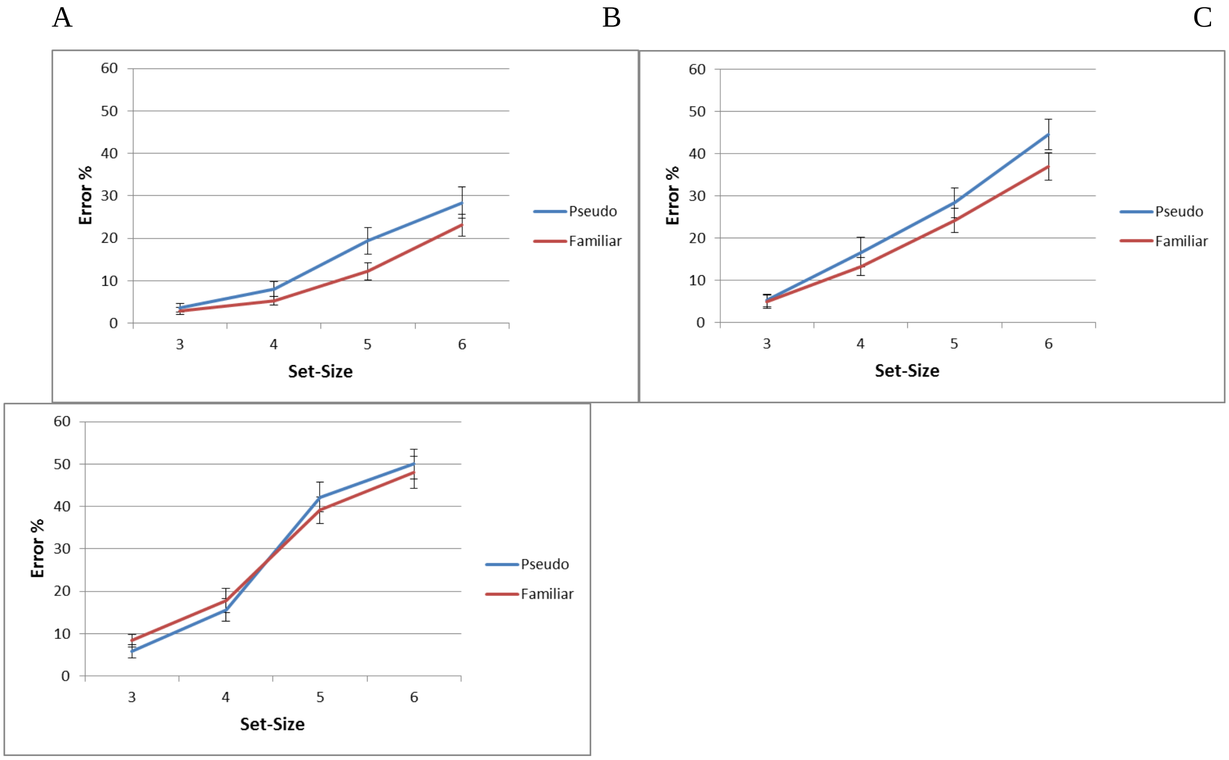

Figure 1. Percentage of errors in tracking short existing and pseudowords in Experiment 1, as a function of speed (A=slow; B=medium; C=fast) and target set-size (error bars = ± 1 SE).

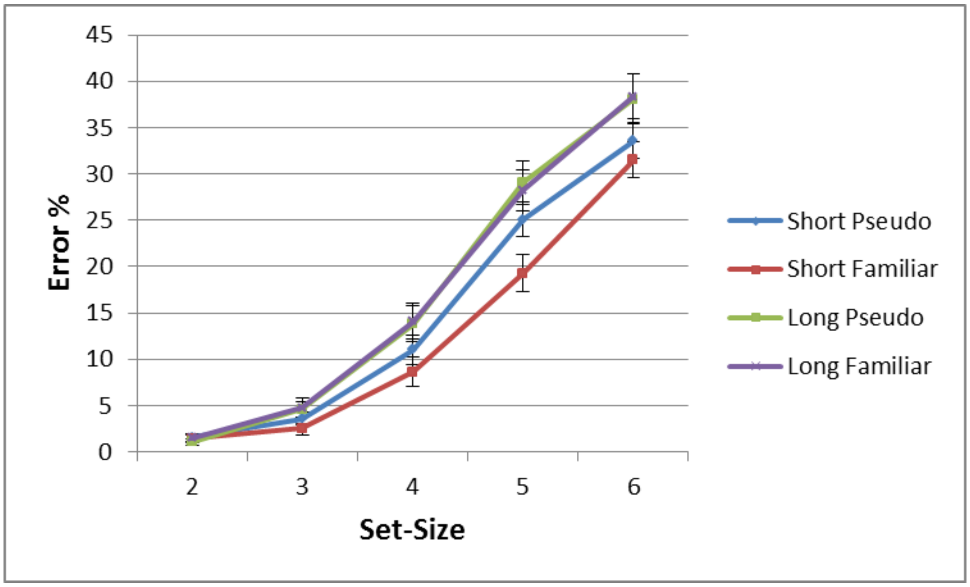

Figure 2. Percentage of errors in tracking short and long existing and pseudowords in Experiment 2, as a function of target set-size (error bars = ± 1 SE).