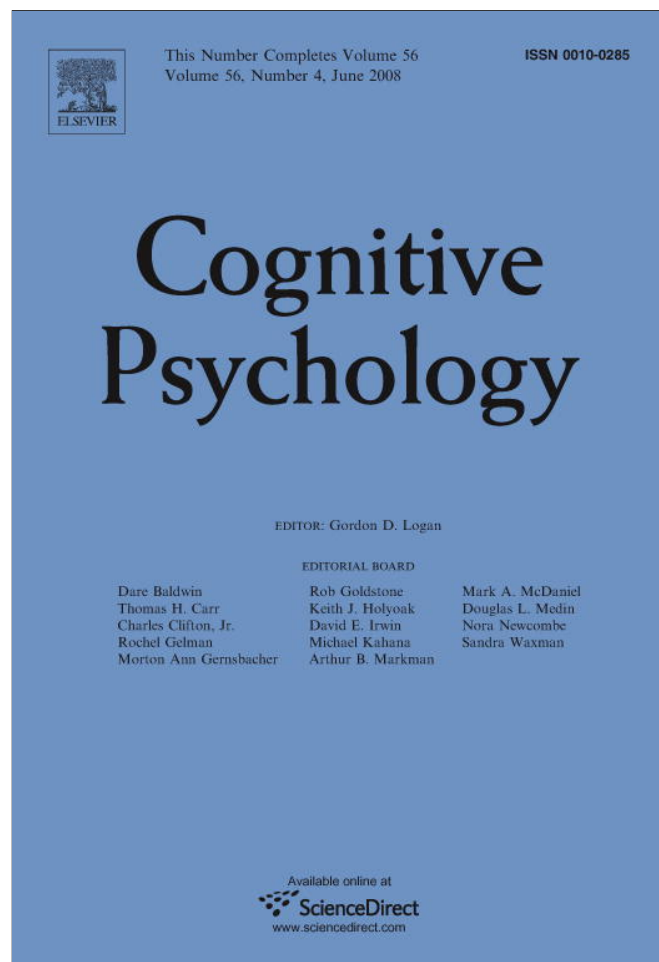


Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



# Dynamic binding of identity and location information: A serial model of multiple identity tracking

Lauri Oksama<sup>a,b,\*</sup>, Jukka Hyönä<sup>b</sup>

<sup>a</sup> *Department of Behavioral Sciences, National Defence College, University of Turku, FIN-20014 Turku, Finland*

<sup>b</sup> *Department of Psychology, University of Turku, FIN-20014 Turku, Finland*

Accepted 6 March 2007

Available online 23 April 2007

---

## Abstract

Tracking of multiple moving objects is commonly assumed to be carried out by a fixed-capacity parallel mechanism. The present study proposes a serial model (MOMIT) to explain performance accuracy in the maintenance of multiple moving objects with distinct identities. A serial refresh mechanism is postulated, which makes recourse to continuous attention switching, a capacity-limited episodic buffer for identity-location bindings, indexed location information stored in the visuospatial short-term memory, and an active role of long-term memory. As identity-location bindings are refreshed serially, a location error is inherent for all other targets except the focally attended one. The magnitude of this location error is a key factor in predicting tracking accuracy. MOMIT's predictions were supported by the data of five experiments: performance accuracy decreased as a function of target set-size, speed, and familiarity. A mathematical version of MOMIT fitted nicely to the observed data with plausible parameter estimates for the binding capacity and refresh time.

© 2007 Elsevier Inc. All rights reserved.

*Keywords:* Multiple identity tracking; Dynamic visual scenes; Identity-location binding; Episodic buffer; Visuospatial short-term memory; Long-term memory; Change detection; Multiple object tracking

---

---

\* Corresponding author. Fax: +358 2 3335060.

*E-mail addresses:* [loksama@utu.fi](mailto:loksama@utu.fi) (L. Oksama), [hyona@utu.fi](mailto:hyona@utu.fi) (J. Hyönä).

## 1. Introduction

Our perceptual-attentional system faces a challenging task when attempting to simultaneously track multiple moving elements in a visual scene. Tasks of this kind are common in many real-life visual environments, such as in traffic and sports, for example, when a player keeps track of other moving players during a soccer game, or when an air traffic controller monitors aircraft on a radar screen. In order to make quick and sensible decisions in tasks like these (e.g., passing the ball to a team mate, or giving ATC clearance to aircraft), observers need to keep track of where each visual element is located at any given time.

Computationally, in this kind of task observers must be able to dynamically bind correct identities to continuously changing spatiotemporal locations: observers face a dynamic what–where binding problem (for the binding problem, see e.g., Treisman, 1996; the special issue of *Visual Cognition* 3–5/8, 2001). In the real world, our perceptual–attentional system must solve this problem in very different environments and stimulus conditions. The target stimuli may vary in many ways with respect to, for example, their spatiotemporal and semantic properties. The number of objects-to-be-tracked may vary and so may speed, movement direction, and movement trajectory of the target objects, as well as the number of distracter items present. The objects may temporarily disappear due to external (occlusion) or internal (saccadic suppression) reasons. The objects may be familiar or unfamiliar to the observer. Finally, the tracked objects may be visually identical to or distinct from each other.

There is empirical evidence demonstrating that binding of ‘what’ to ‘where’ in a visual scene is not an easy task for the attentional-perceptual system even when the observed scenes are static. For example, data suggest that observers may be prone to illusory conjunctions, that is, to erroneously bind visual feature combinations (e.g., binding the red color with a triangle when it should have been attached to a circle; Treisman & Schmidt, 1982). Moreover, it seems that what–where bindings are not achieved automatically in an early stage of visual processing. Instead, neurophysiological evidence suggests that the visual input is functionally decomposed into separate dimensions in the ventral (what) and the dorsal (where) streams (e.g., Ungerleider & Mishkin, 1982). All this supports the idea that the dynamic binding problem described above is real and that it is a non-trivial endeavor for the perceptual-attentional system to accomplish the task. Clearly, a mechanism is needed to ensure that correct identity-location combinations are perceived and maintained.

Several fundamental questions can be asked about the workings of the binding mechanism. How efficiently does it carry out its task across different dynamic input situations? Is it vulnerable to changes in ‘where’ information but not in ‘what’ information, or vice versa, or both, or neither? How crucial is the availability of identity information for successful tracking? Does it matter whether the identities are known to the observer (*the role of semantics*)? How susceptible is the binding mechanism to changes in target velocity (*the effect of speed*)? Does the mechanism possess a large parallel capacity so that the observer is able to genuinely track several moving target identities simultaneously, or does it possess only a very limited capacity so that the observer needs to constantly shift the focal attention from one target to another to check out the identities (*the effect of set-size*)?

Previous research on dynamic visual attention suggests that the tracking mechanism carries out its task very efficiently without recourse to featural or semantic properties of the objects (e.g., Kahneman, Treisman, & Gibbs, 1992; Pylyshyn & Storm, 1988, see Section 9 for a more detailed discussion for different versions of the fixed capacity parallel account). According to theories of this kind, tracking is parallel and semantic properties of the objects play no role in the process, but dynamic what–where bindings are made solely on the basis of low-level physical or spatiotemporal properties of the moving objects.

In this article, we propose an alternative serial account where semantic properties can influence dynamic tracking. The starting point for the model were the findings reported by Oksama and Hyönä (2004). Among other things, they observed notable individual variation in the tracking capacity, which significantly correlated with visuospatial working memory and task-switching capacity. Oksama and Hyönä sketched a serial model based on temporary visuospatial memory in order to explain their results. This serial model is here developed further and also fully formalized. Its predictions investigated in dynamic binding experiments where set-size, object speed and object type were manipulated. We demonstrate that our serial model can quantitatively account for observed set-size, speed, and semantic effects.

In what follows, we introduce our model of dynamic binding, called MOMIT (Model Of Multiple Identity Tracking), which makes recourse to continuous attention switching, episodic buffer, visuo-spatial short-term memory (VSTM), and an active role of long-term memory in dynamic binding. We first lay out and motivate the basic assumptions of the model and outline the general architecture of MOMIT. We then describe how the postulated mechanisms are assumed to function in practice (a functional model). The mathematical model and its fit to experimental data are described at the end of the Results section.

### *1.1. Model of multiple identity tracking (MOMIT)*

MOMIT is based on the following five tenets:

- (1) Efficient maintenance of multiple moving objects requires continuous serial (re)activation and refreshing of the dynamic identity-location bindings. If bindings are not periodically refreshed, they are eventually lost. The refreshing of existing bindings is assumed to be a non-automatic and effortful process based on continuous shifts of attention between targets.
- (2) There is a capacity limitation as to the number of bindings that can be simultaneously kept active in the episodic buffer where the identity-location bindings are assumed to be created and temporarily stored. Moreover, there is significant individual variation in the tracking capacity.
- (3) Long-term memory (LTM) representations are utilized in creating temporary bindings. Thus, binding is more readily made for familiar than unfamiliar visual objects. As a result, tracking performance is better for familiar than unfamiliar targets.
- (4) Spatial indexes (or location pointers) for the tracked targets are temporarily stored in VSTM. These indexes are then utilized by the mechanism that programs shifts of visual attention between targets. As targets move continuously, there is a location error involved in the spatial indexes.

- (5) The system responsible for shifting of attention during tracking also obtains location information of moving objects in parallel. This information is provided by the peripheral vision. However, unlike the location information provided by VSTM (see Principle 4), this spatial information is not indexed; in other words, targets are not differentiated from distracters on the basis of this information.

Next, we motivate in more detail each of the five tenets of MOMIT.

The first assumption of MOMIT is that targets are tracked by continuously refreshing the identity-location bindings one at a time. Underlying this assumption is the view that focal attention is required (1) to identify a target and (2) to (re)bind the identity to a spatial index. Several parallels with this assumption can be found in the relevant prior literature. First, [Moray \(1984\)](#) and [Logan \(2002a\)](#) have argued that when task-relevant visual episodes are spatially spread around a large visual scene, due to visual acuity reasons, it is highly likely that simultaneous tracking of multiple objects becomes impossible (in the present set of experiments, target objects moved wide apart from each other). Moray develops this idea further by introducing uncertainty as a critical factor in dynamic environments. Uncertainty can be either exogenous (due to characteristics of the visual input, such as poor visibility) or endogenous (due to psychological reasons, such as forgetting) in nature. Uncertainty increases as a function of time. To reduce uncertainty, the observer is assumed to periodically update the constantly changing situation by taking visual samples of the current state of the dynamic environment. Second, the first assumption of MOMIT is also in keeping with the studies of Treisman and colleagues, who have argued that focused attention is required to both create feature conjunctions ([Treisman & Gelade, 1980](#)) and to temporarily maintain bindings ([Wheeler & Treisman, 2002](#)). Third, it is in line with the view advocated by [Logan and Zbrodoff \(1999\)](#), according to which visuospatial attention is intimately involved in constructing compositional representations, such as identity-location bindings. Finally, the assumption bears analogy to the findings in reading research where it is shown that words (a specific type of visual objects) are primarily identified only when they are focally attended (see [Rayner, 1998](#), for a review). (We hasten to add that word recognition in reading undoubtedly differs from multiple identity tracking in many important respects.) The binding with the lowest activation (or highest uncertainty) is assumed to have the highest priority when a target object is chosen to be refreshed (see [Moray, 1984](#)). It should be noted the first assumption of MOMIT stands in contrast to [Cavanagh and Alvarez \(2005\)](#), [Pylyshyn and Storm \(1988\)](#), and [Yantis \(1992\)](#), who all argue in favor of parallel models and against serial accounts.

By efficient maintenance of identity-location bindings we mean that they can be accessed rapidly and accurately if needed, for example, in order to make a rapid decision, such as giving clearance to an aircraft or passing the ball to a fellow player. More direct evidence supporting the assumption of continuous refreshing comes from [Oksama and Hyönä \(2004\)](#). They observed that tracking deteriorated linearly as a function of set-size ([Pylyshyn & Storm, 1988](#) also observed that tracking deteriorated as a function of set-size but they interpret the finding differently). This implies that the tracking performance involves a serial component. Evidence supporting the quickly decaying nature of target location information is also provided by [Oksama and Hyönä \(2004\)](#). They conducted a signal detection analysis of the performance of tracking multiple identical objects and showed that the tracking performance deteriorated as a function of time (see also [Pylyshyn, 2004](#)). This is taken as evidence that visual tracking is not completely automatic

and parallel but effortful serial attention is also needed. Finally, Oksama and Hyönä (2004) demonstrated that individual differences in the attention switching capacity significantly correlate with the tracking performance.

We assume that there exists a tight coupling between attention shifts and eye movements (see e.g., Deubel & Schneider, 1996; Findlay & Gilchrist, 2003): attention drives the eyes in that a shift of attention is followed by a saccade to the attended object location (Henderson, 1992). That eye movements are needed in tracking objects in real-life type of visual environments is also proposed by Logan (2002a) and Moray (1984). However, this specific assumption is not put to a test in the present study (see Landry, Sheridan, & Yufik, 2001, for an attempt to use eye movements to study tracking of multiple moving objects).

The second building block of MOMIT is a limited-capacity episodic buffer where identity-location bindings are assumed to be created and temporarily stored. For example, Luck and Vogel (1997) have shown that about four static identity-location bindings may be kept active at a time. The average capacity is about 4 even when the objects are moving (Horowitz et al., 2007; Oksama & Hyönä, 2004). The notion of episodic buffer is borrowed from Baddeley (2000), who posits such a temporary store for the service of integrating different fleeting representations into one unified whole (e.g., binding identity and location information together). We like to add that we are not committed to the notion of episodic buffer, but any temporary store would do. The fact that working memory is indeed intimately involved in multiple identity tracking (MIT) is demonstrated by Oksama and Hyönä (2004), who observed a strong correlation between MIT performance and working memory capacity measured using visuospatial stimuli. Finally, the claim that significant individual differences exist in the tracking capacity is also based on the results of Oksama and Hyönä (2004).

The third assumption of MOMIT is that LTM is involved in the creation of temporary identity-location bindings. If available, identity information is activated during MIT (i.e., when creating bindings). Thus, attention is assumed to be disengaged sooner from familiar than unfamiliar objects, and the latency of revisiting a target object is thus shorter for familiar than unfamiliar objects. Hence, MOMIT predicts better tracking performance for familiar than unfamiliar objects. This assumption is in line with theories that consider LTM representations to be pertinent in perception and short-term maintenance of visual objects and other type of stimuli (see e.g., Chun & Potter, 1995; Cowan, 1995; Kanwisher, 1991; Ruchkin, Grafman, Cameron, & Berndt, 2003; Shapiro, Raymond, & Arnell, 1994). This principle is also in keeping with the spirit of the late selection theories of attention, which argue that all object properties, including semantic properties, are activated and utilized in the early stages of visual processing and thus influence the binding process. To our knowledge, this view has not yet been applied to dynamic visual attention.<sup>1</sup>

The fourth assumption of MOMIT is that the location information necessary to create bindings is available in VSTM (perhaps in the form of the visuospatial sketchpad proposed by Baddeley, 1986, see also Logie, 1995) as spatial indexes (Logan & Zbrodoff, 1999; Pylyshyn & Storm, 1988). Spatial indexes are deictic pointers that specify the location of the target objects in space. They act as an ‘address’ for the perceptual object but they do not specify the properties of objects (see Logan & Zbrodoff, 1999). A key feature

---

<sup>1</sup> Henderson and Anes (1994) and Henderson (1994) argue that both episodic spatiotemporal and LTM representations are activated in the early stages of visual processing, but they argue that these representations are independent. In their view, LTM does not influence the binding process.

of this subsystem is that the spatial coordinates are not assumed to be perfectly up-to-date, but they provide only an approximate location for the targets that are currently not focally attended. As target location information is updated only when the target is focally attended, in dynamic visual environments the location information provided by spatial indexes is always old by nature. Thus, there is a difference between the stored VSTM coordinates and the present location of the targets (except for the focally attended one). We call this difference the *VSTM coordinate error*. In MOMIT, the magnitude of this VSTM error is one of the key factors determining the success of tracking multiple moving objects. The assumption that spatial indexes are provided by VSTM is at odds with the models of multiple object tracking, which posit that no memory is needed to perform the task (e.g., Pylyshyn, 1989; Yantis, 1992). It is also inconsistent with the FINST theory of Pylyshyn (1989, 1994, 2001), which posits that spatial indexes are constantly updated; once the pointers are aligned with the targets, they are assumed to move along with the targets without the need of attention (they are like ‘sticky fingers’ placed on targets). To sum up, in the MOMIT architecture semantic identity-location bindings are assumed to be created and stored in the episodic buffer, but the targets’ indexed location coordinates are stored in VSTM.

The fifth assumption of MOMIT is that peripheral vision provides non-indexed location information (cf. the indexed location information of VSTM) about all moving objects in parallel. This information is accurate, but it does not differentiate targets from distracters. The serial shifting of attention is controlled partly endogenously with the help of VSTM (see above) and partly exogenously with the help of peripheral vision. VSTM provides a rough spatial location of the to-be-attended target; the object to be focally attended next is the closest object around the area determined by VSTM. Thus, the attended object is not necessarily the intended one (due to the VSTM error)—it could also be a distracter or a recently refreshed target. If the attended object is not the intended one, it is attended anyway; if it is a distracter (or another target), a search for the intended target item is initiated.

### *1.2. Maintenance of what and where bindings: the general architecture of MOMIT*

Here we outline the general architecture of MOMIT. A graphic description of the architecture of MOMIT is shown in Fig. 1. The model provides a functional description of the process of how *what* and *where* information are linked together and maintained in a dynamic visual environment. The model’s architecture consists of five components: (1) a component responsible for the analysis of what information, (2) a component responsible for the analysis of where information, (3) a temporary memory buffer (VSTM) for maintaining indexed location information, (4) a control system for attention switching, and (5) a temporary episodic buffer for maintaining what–where bindings. The model is designed to simulate a situation where multiple, constantly moving objects are tracked in a visual scene (possibly, but not necessarily in the presence of distracters). The situation where objects are static or only one object is moving are special cases of the kind of visual environment we are interested in here.<sup>2</sup> In the following, we describe the different components of the model.

---

<sup>2</sup> It is also possible that features specific to static environments may emerge. At least it may be difficult to distinguish object-specific effects from location-specific effects (see Experiment 1 of Gordon & Irwin, 1996).

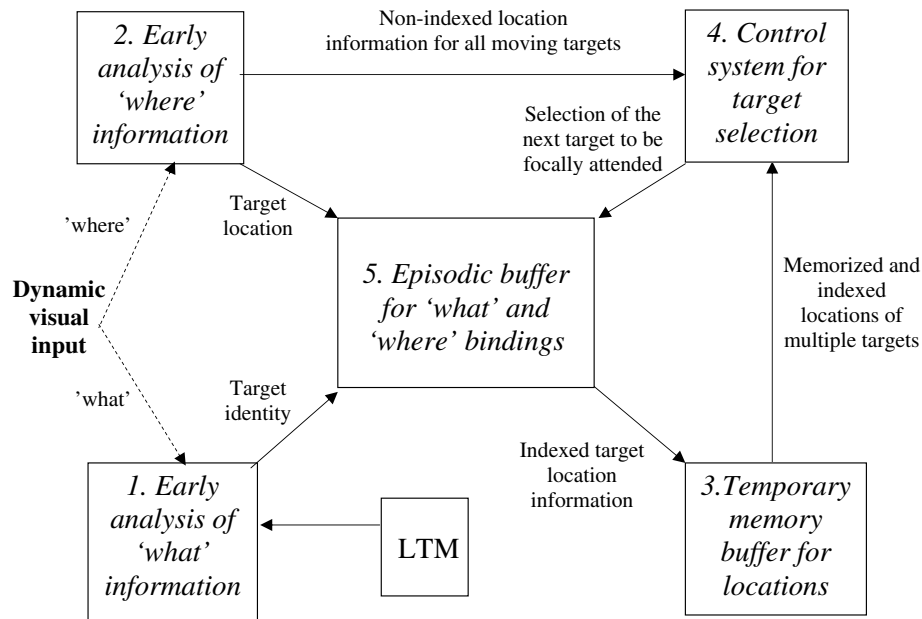


Fig. 1. The general architecture of MOMIT (model of multiple identity tracking).

- (1) *Analysis of 'what'*. This component provides access to the semantic identity of a focally attended object. It is assumed that identity information is activated early and independently of the 'where' analysis. Identity information is transmitted to the episodic buffer (Component 5), where temporary bindings are formed and maintained. The model assumes that the faster identity information is accessed for the attended object, the more readily a binding will be formed and, consequently, the more efficiently multiple bindings will be maintained (i.e., the refresh rate is fast and the probability of losing previous bindings is small). Thus, factors that influence the speed of access of identity information will influence the efficiency to maintain multiple bindings. One such factor is the familiarity of the identity information stored in LMT. Familiar objects will be identified faster and consequently tracked more efficiently. A second factor by which LTM may influence the binding process is related to discriminability among target objects (not tested in the present study). It may be presumed that a low discriminability will slow down the *what*-analysis and thus hamper with the dynamic maintenance of multiple bindings (cf. Chun & Potter, 1995; Duncan & Humphreys, 1989).
- (2) *Analysis of 'where'*. This component yields parallel location information for those moving objects (both targets and distracters) that are positioned within an area constrained by peripheral vision. This spatial information is transmitted to Component 4 that is responsible for serial shifting of attention. It is also transmitted to the episodic buffer (Component 5) where the target-relevant location information is used in creating or updating an identity-location binding for the currently attended target. The resulting indexed location information is then temporarily stored in VSTM (Component 3). Obviously, all factors that impoverish the peripheral perception of moving objects (e.g., occlusions) hamper with the workings of this parallel *where* system.

- (3) *Temporary memory buffer for indexed location information.* The core assumption here is that in order to maintain of multiple bindings, temporary storage of the targets' former locations is needed. This information is used when programming attention shifts between the tracked objects. Thus, MOMIT assumes that there is a component that saves temporal location information for the tracked objects when they are focally attended. This storage function is presumed to be carried out by VSTM. Because the target location information is updated only when targets are focally attended, it means that with constantly moving targets this indexed location information is not completely accurate (cf. the VSTM coordinate error in the mathematical formulation of MOMIT presented below). This short-term memory component plays a central role in the predictions derived from MOMIT. The more accurate the location information is in the buffer, the more accurately attention will be shifted between the to-be-tracked targets. In dynamic visual environments, the accuracy of indexed location information is modulated by target speed and set-size, as discussed above. Apart from the indexed locations becoming inaccurate as a function of time, they are also vulnerable to endogenous uncertainty (Moray, 1984), such as forgetting and interference.
- (4) *Control system for attention switching.* Any model that makes recourse to serial attention switching requires an attention control mechanism. This is because during tracking a decision of which target to select next is frequently made. A random selection does not work, as non-optimal selection results in inefficient maintenance of multiple identities. For example, by attending a recently refreshed target another target is put into danger (i.e., the binding may be lost). It is assumed that this control system receives input from two sources, from VSTM (an endogenous component) and from the parallel *where* system (an exogenous component), and the object to be attended is determined by the joint interplay between these two information sources. VSTM provides a rough spatial location of the to-be-attended target and the parallel component determines the specific object to be focally attended next, which is the closest object around the area determined by VSTM. The above argumentation leads to a prediction that attentional resources available for time-sharing and smooth serial allocation of attention between items-to-be-tracked contribute to the overall tracking capacity and to the effectiveness with which bindings are maintained in an active state. Individual differences in the executive capacity (an ability to serially allocate attention to multiple tasks) constitute one such factor as demonstrated by Oksama and Hyönä (2004), experience with the task another, as demonstrated by Allen, McGeorge, Pearson, and Milne (2004). MOMIT also predicts that when MIT is performed concurrently with another (non-visual) task that also calls for executive resources, the MIT performance will be hampered.
- (5) *Temporary episodic buffer for what–where bindings.* MOMIT assumes that temporary episodic memory representations of the formed bindings are constructed and maintained during tracking.<sup>3</sup> The obvious function of these representations is to retain bindings for a short period of time to overcome temporary disappearance of visual

---

<sup>3</sup> The “precious child” goes by many names. Kahneman and Treisman (1984) call it object file, Kanwisher (1991) and Chun (1997) object token (in contrast to spatiotemporal token), and Rensink (2000) calls it nexus.

input (e.g., during saccades, blinks, or occlusions). These representations may then be consulted if needed ('the lion is behind the tree'). As argued above, the episodic buffer consists on average of four bindings.

### 1.3. *A functional description of MOMIT*

Based on the principles described above, we next provide a functional description of the mechanisms assumed to be responsible for maintaining multiple dynamic identity-location bindings. The starting point in the description is that the system tries to maintain three moving targets: T1, T2, and T3.

1. Target T1 is focally attended and a semantic identity-location binding is formed for it (or updated in case a binding has already been created during a previous cycle) in the episodic buffer. The present location of Target T1 is stored in VSTM.
2. The next target-to-be-attended is endogenously selected among the alternatives (T2 or T3). Target T2 is selected on the basis of the activation level of the bindings (the binding with the lowest activation has the highest priority).
3. When attention is disengaged from Target T1, focal attention is shifted to an exogenously determined object in the vicinity of the endogenously selected target (i.e., to an object nearest to the endogenously selected target location).
- 4a. If the new attended object is the intended one (T2 in this case), then an identity-location binding is (re)constructed for T2 (see Cycle 1), and in Cycle 2 T3 is selected as the next to-be-attended moving object. Cycles 1–4 are repeated as long as necessary. Or,
- 4b. if the attended object is not the intended one, a corrective attention shift is carried out from the wrong object to a new one located in the vicinity. When the right one is found, the process is analogous to what is described in (4a).

### 1.4. *Testing the goodness of fit of MOMIT*

In order to test MOMIT's goodness of fit, we quantitatively fitted the mathematically formalized MOMIT to the data (see Section 8). In addition, we also derived MOMIT's specific predictions as regards main effects and interactions in the analyses of variance (see [Appendix A](#) for a mathematical proof of the predictions and [Appendix B](#) for a detailed comparison of the observed and predicted effects). In short, MOMIT predicts that maintenance of dynamic bindings deteriorates as a function of target set-size, speed, and familiarity. It also predicts that all interactions between these three factors are significant. These predictions arise from the model's serial architecture. As bindings have to be refreshed serially, the effects of these factors that influence the time it takes to refresh target bindings add up in a multiplicative fashion—hence the predictions for interactions.

### 1.5. *Overview of experimental procedures*

To study dynamic identity tracking, i.e., the observer's awareness of the location and identity of visual elements at any given time, we developed two multiple identity tracking

(MIT) tasks. In the first one (used in Experiments 1 and 3A), several non-identical elements move around for a while, after which movement is stopped and elements are immediately masked (to eliminate iconic memory), one element is probed, and the participant is asked to identify the probed element (in the variant used in Experiment 2A objects continue to move also when masked and probed). In order to minimize verbalization and memory requirements during response selection, the probed item is chosen among all the moving objects that are presented in a separate response screen. We call this the partial report probe recognition (PRPR) task (it was also used by Oksama & Hyönä, 2004).

The second variant of MIT makes use of the change detection paradigm (e.g., Rensink, O'Regan, & Clark, 1997; Simons & Levin, 1998). In this variant (used in Experiments 2B and 3B), the movement phase is followed by a short flicker, during which the moving objects are erased and the participants see only empty framed squares moving (each object is surrounded by a frame throughout the trial). When objects reappear to the empty frames, they appear either in their original frames, or in half of the time two of the targets swap position. The participant's task is to respond whether or not a change took place (i.e., to respond either 'yes' or 'no'). Thus, for the response, no semantic identity information needs to be necessarily consulted (for another variant of the change detection procedure in multiple object tracking, see Bahrami, 2003).

Notice that distracter items are not needed in MIT because the tracked objects are visually distinct from each other. Notice also that in order to be able to respond accurately to the masked probes some kind of temporary, short-lived (but not iconic) memory representation of the what–where binding must be consulted. If there is no memory trace for the binding, the accuracy in responding to the masked probes is at a chance level. Thus, in a sense we test whether an 'occlusion-tolerant' memory representation is created during dynamic identity tracking (the issue of temporary memory representations is discussed in more detail in Section 9).

## 2. Experiment 1

Effects of set-size, object type, and object speed on dynamic identity tracking was studied in Experiment 1. By including all these factors in the same experimental design it was possible to examine the interactions between the factors and thus test the predictions of MOMIT put forth in Section 1. Set-size was varied from 2 to 6. Three speed conditions were created that are called slow, medium, and fast. To study possible effects of semantic identity in MIT we compared highly familiar objects to pseudo-objects. By definition, a LTM representation is available for familiar objects but not for pseudo-objects (note that pseudo-objects have distinct featural identity). The PRPR technique was employed.

### 2.1. Method

#### 2.1.1. Participants

A total of 168 participants were recruited for the experiment. They were randomly divided to three groups of 56 participants (the three speed conditions). 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. All participants had normal, uncorrected vision. A pre-selection was made on the basis of previous school achievement (in mathematics and English). This screening was done as a part of the air pilot recruitment process.

### 2.1.2. Apparatus

The stimuli were presented on 19-inch Eizo FlexScan F730 monitors with a resolution of 1280 by 1024 pixels controlled by Matrox G400 cards, Pentium 3, 500 MHz, 128 Mt RAM computers, and the E-prime software (Schneider, Eschman, & Zuccolotto, 2002a, 2002b). The software that generated the motion sequences was written in Visual Basic.

### 2.1.3. Stimuli

Two sets of six stimuli were used, six familiar objects and six pseudo-objects. As familiar objects, vertically oriented line drawings of real objects (flower, coat, lobster, rocking chair, rooster, and watch) were used (see Fig. 2). The objects were selected to represent different semantic categories. The pictures were selected from a standardized set of black-and-white line drawings (see Snodgrass & Vanderwart, 1980). The visual complexity of the chosen pictures in terms of Snodgrass and Vanderwart's 5-point rating scale ranged from 3.25 to 4.48; thus the visual complexity of all the chosen pictures was above average. The pseudo-objects were vertically oriented, visually rather complex, black-and-white line drawings that were selected from the pictures provided by Kroll and Potter (1984). They were visually similar to each other and had an object-like appearance (see Fig. 3). A visual complexity estimation was done subjectively by the authors, because the visual complexity of the non-objects was not rated by Kroll and Potter (1984).

The stimuli (75 pixels in height and 41–69 pixels in width) were black outline drawings on a white background subtending a visual angle of  $1.9 \times 1.1$ – $1.8$  deg. The computer screen subtended a visual angle of 25 deg horizontally and 32 deg vertically. A subset of two to six objects was designated as targets. To keep the total number of objects and the probability of guessing constant, the non-targets were also visible during each trial; thus the animated stimuli always consisted of six moving pictures. The picture combinations within a target set were constructed so that each picture was selected equally often as a target (each picture appeared eight times as a target in the target-set two, 12 times in the target-set three, etc.). After the movement phase, target probing was carried out by flashing a black frame ( $75 \times 75$  pixels, 2 pixels in width,  $1.9 \times 1.9$  deg) around the target. The frames were not visible during the movement phase. Visual masks ( $75 \times 75$  pixels) of

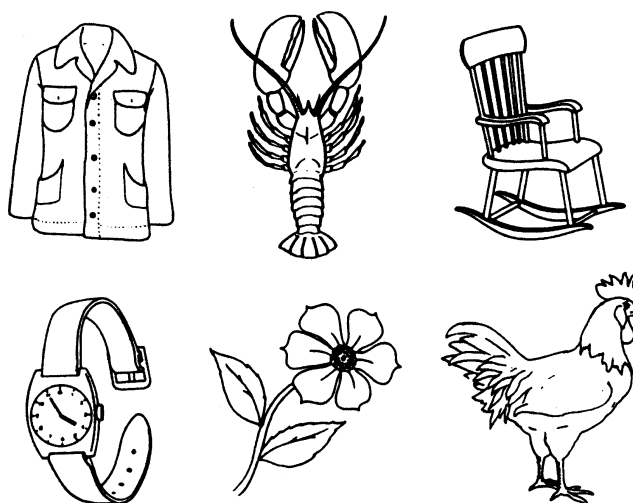


Fig. 2. The target pictures of the familiar objects used in Experiments 1, 2A, and 2B. They were reprinted from Snodgrass and Vanderwart (1980), © 2007 Life Science Associates.

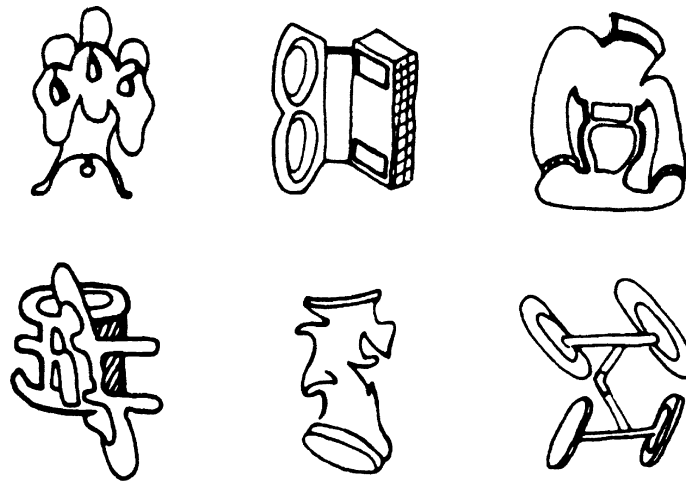


Fig. 3. The target pictures of the pseudo-objects used in Experiments 1, 2A, and 2B. Pictures were reprinted from Kroll and Potter (1984) with permission from Elsevier.

variable kind that replaced the pictures at the end of movement phase were created for different stimulus versions by copying, rotating and combining parts of the pictures used in a given version.

#### 2.1.4. Movement sequences

The experimental trials consisted of 167, 233, or 300 static frames presented one after another for 30 ms each. In the slow speed condition, items moved a minimum of 1 and a maximum of 5.7 pixels per frame. As each frame had a 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 item 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 item velocities were in the range of 6.0 to 15.7 deg/s (the average speed was 10.7 deg/s).

Initial object positions were generated at random. Movement direction for each object was chosen randomly from among the eight compass directions. Each object was assigned a movement duration that were 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; or 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 during the motion phase. Several actions were taken to avoid these collisions. First, an invisible cushion (which went through the vertices of the  $75 \times 75$  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

in case of potential collisions. 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.

Twenty-four animation sequences (trajectory files) were generated and stored offline and were divided into two sets of twelve files for the two experimental blocks. The same trajectory files were used for both object types (i.e., for familiar objects and pseudo-objects). One trajectory file was used five times in the block, one time in each target set and duration. Thus, all the trajectories within each target set and duration were different but the trajectories between the different target sets, durations, and stimulus blocks were similar. However, the chosen target objects were different in different target sets while the trajectories were the same. This technique ensured that any differences between different set-size and object type conditions were not due to differences in the trajectory patterns (see Scholl & Pylyshyn, 1999, for a similar procedure).

#### 2.1.5. Task

In Fig. 4 schematic illustration is presented of the partial-report probe recognition procedure. At the beginning of each trial the targets (2–6 objects) were designated by flashing a frame around them. The participants' task was to track the identity of these designated targets during the movement phase. After the movement stopped, all the objects were masked and one of the target objects was probed by flashing a frame around it. Finally the screen was cleared and a response screen appeared, where all the six pictures present during the tracking phase were arranged into an array with two rows and three columns. A computer mouse was used to collect the responses. A response frame (100 × 100 pixel) surrounded the pictures. The mouse pointer was initially positioned in the middle of the stimulus array. Participants were asked to select (point and click within a framed picture) the probed picture as accurately and quickly as possible. They were asked to guess if they did not know the answer.

#### 2.1.6. Procedure

Participants (a maximum of 10 at a time) were seated approximately 57 cm from the display; a chinrest was used to reduce head movements and to control the viewing distance. A screen was placed between the participants in order to prevent them from seeing or disturbing each other. Participants were given written instructions prior to the experiment, which outlined the general procedure and explained a trial sequence. They were to note the positions and identity of the flashing targets at the start of each trial and to keep track of them during the movement phase. The familiar objects and pseudo-objects were presented in separate blocks.

At the beginning of each trial, the items were displayed for 1s. After that, a black frame flashed on and off for 10 times (flashing duration was 150 ms; total flashing time was 3000 ms) around the designated targets (2–6 targets). The targets then began to move in a random and continuous fashion around the screen. The participants tracked them for 5, 7, or 9 s, after which the movement stopped, and the objects were simultaneously masked. This was followed by the flashing of a black frame for five times (flashing duration was 150 ms; total flashing time was 1500 ms) around one of the targets (i.e., the

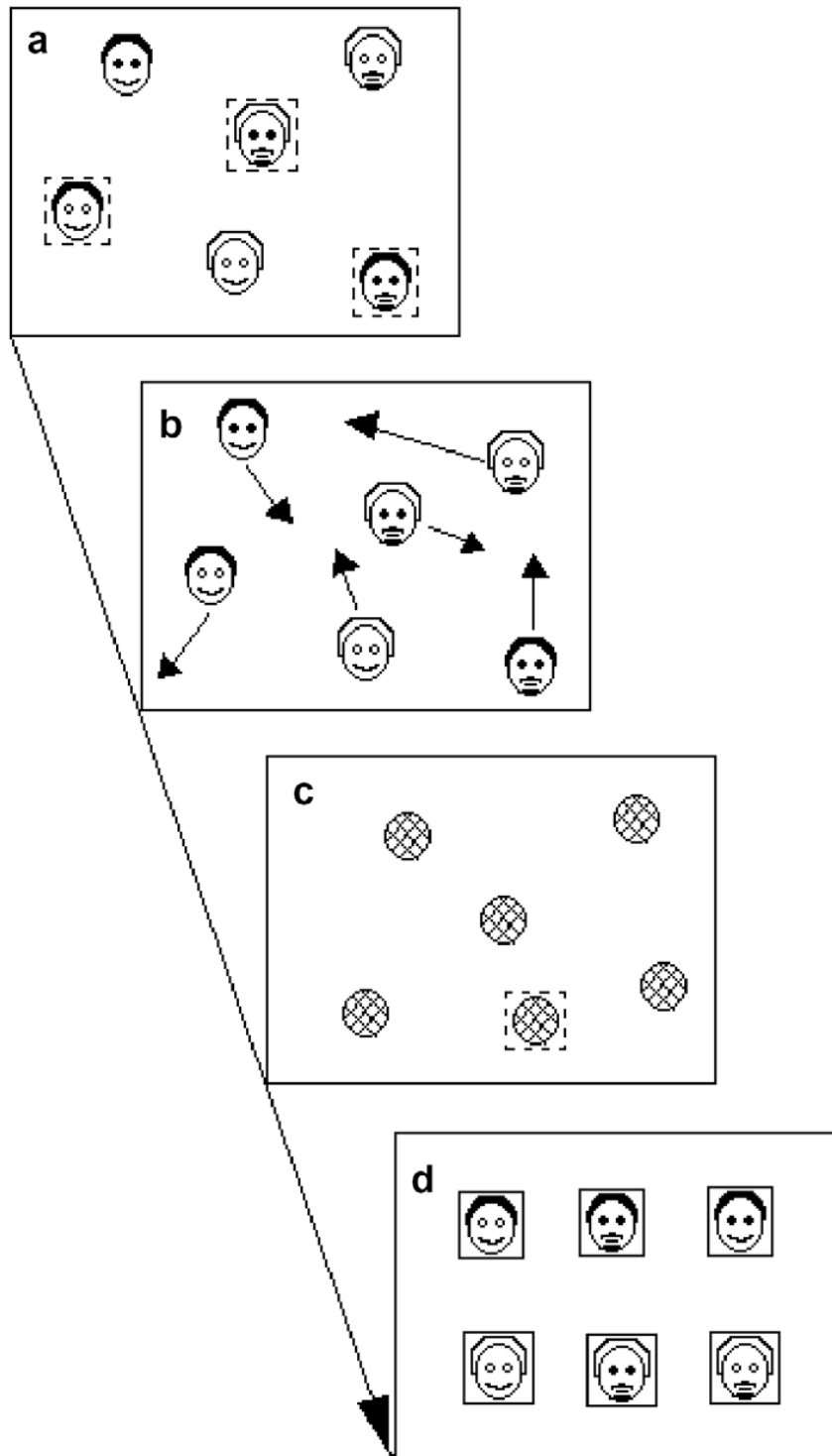


Fig. 4. A schematic depiction (exactly these stimuli were not used in any of the experiments reported) of the multiple-identity tracking task with the partial-report probe-recognition technique (not to scale) used in Experiments 1, 2A, and 3A. Display a: six different objects are presented, and 2–6 of them are designated as targets by flashing a frame around them. Display b: all of them begin to move randomly about the screen. Display c: after a 5, 7, or 9 s, the motion stops, the objects are masked, and one of the targets is marked by flashing a frame around it (the probe). Display d: a response screen appears where all the presented objects are presented and participants are to select the probed item by using a mouse.

probed item). After a response was given, 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. Each participant completed two blocks of 60 trials for each experimental set (i.e., the familiar objects and the pseudo-objects), altogether 240 trials. The order of trials was randomized separately for each participant within each block. The order of blocks was counterbalanced across participants. There was a short rest period between the blocks. The entire session took about 75 min.

### 2.1.7. Design

There were three manipulated factors in the experiment: the number of targets tracked (2–6), the type of object tracked (familiar object vs. pseudo-object), and the object speed (2.6, 6.3, 10.7 deg/s; coined ‘slow’, ‘medium’ and ‘fast’). Number of targets and type of object were within-subject variables, and object speed was a between-subject variable. There were 24 trials in each of the 10 conditions of the within-subject variables. The counterbalancing group (familiar objects first vs. pseudo-objects first) was used as a between-subject variable to partial out from the total variance a possible practice effect.

## 2.2. Results and discussion

### 2.2.1. Error rate in performance accuracy

The error data were submitted to a 5 (set-size)  $\times$  2 (object type)  $\times$  3 (speed)  $\times$  2 (counterbalancing group) mixed design analysis of variance (ANOVA). In this and subsequent analyses, a Greenhouse-Geisser correction was applied to the  $p$  values whenever needed. Table 1 shows the mean error rate in performance accuracy as a function of target set-size and object type in the three different speed conditions.

A significant main effect was found for the number of targets ( $F(4, 648) = 527.78$ ,  $p < .001$ ); performance deteriorated as the number of targets increased. The main effect of object type was significant ( $F(1, 162) = 18.95$ ,  $p < .001$ ); familiar objects produced less errors than pseudo-objects. Moreover, the interaction between the number of targets and the type of object was also significant ( $F(4, 648) = 5.72$ ,  $p < .01$ ). A trend analysis showed that this interaction was located mainly in the linear component (linear compo-

Table 1

Percent of trials on which the participant failed to correctly respond to the probe in Experiment 1, for the five set-sizes (2–6), the two object types (pseudo vs. familiar), and the three levels of object speed (slow, medium, fast)

Object speed	Type of object	Number of targets									
		2		3		4		5		6	
		<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Slow	Pseudo	1.93	0.41	3.20	0.69	6.18	1.14	13.47	1.72	19.42	1.77
	Familiar	1.49	0.34	2.46	0.50	3.50	0.68	9.75	1.41	16.07	1.92
Medium	Pseudo	1.79	0.37	2.98	0.59	8.26	1.19	20.61	1.82	28.65	1.93
	Familiar	1.86	0.41	3.50	0.53	6.99	0.88	14.88	1.53	26.26	1.95
Fast	Pseudo	1.71	0.41	5.65	0.77	19.64	1.75	34.15	2.10	44.35	2.06
	Familiar	1.41	0.32	5.58	0.73	18.23	1.71	32.89	1.87	39.88	1.90

ment accounts for 87.15% of the variance,  $F(1, 162) = 13.75$ ,  $p < .001$ ; a cubic component accounts for 12.8% of the variance,  $F(1, 162) = 3.94$ ,  $p < .05$ ). That is, performance deteriorated steeper for pseudo-objects than for familiar objects as a function target set-size. Post-hoc comparisons (paired  $t$  tests) indicated that performance was significantly better for familiar objects than for pseudo-objects when four, five or six targets were tracked (set-size 4:  $t(167) = 2.28$ ,  $p < .05$ ; set-size 5:  $t(167) = 3.24$ ,  $p < .01$ ; set-size 6:  $t(167) = 2.93$ ,  $p < .01$ ). Performance was at ceiling for set-size 2 and 3, so an effect of object type is difficult to establish in these conditions.

The main effect of object speed proved significant,  $F(2, 162) = 62.35$ ,  $p < .001$ ; performance deteriorated as the target velocity increased. Finally, the set-size  $\times$  object speed interaction was also significant ( $F(8, 648) = 41.59$ ,  $p < .001$ ). A trend analysis showed that this interaction was located almost entirely in the linear component (the linear component accounts for 93.2% of the variance,  $F(2, 162) = 66.83$ ,  $p < .001$ ; the quadratic component accounts for 2.1% of the variance,  $F(2, 162) = 4.83$ ,  $p < .01$ ; the cubic component accounts for 4.4% of the variance,  $F(2, 162) = 15.76$ ,  $p < .001$ ). That is, the faster the targets moved and the more targets there were to be tracked, the steeper the tracking performance deteriorated. The object type  $\times$  object speed interaction and the 3-way interaction were not significant ( $F < 1$ , and  $F(8, 648) = 1.13$ ,  $p > .34$ , respectively).

The set-size  $\times$  speed and set-size  $\times$  type interactions are correctly predicted by MOMIT. On the other hand, MOMIT also predicts a type  $\times$  speed interaction and a 3-way interaction between set-size, type, and speed, both of which remained non-significant (numerically there is a trend for a 3-way interaction). The observed and predicted effects are compared and discussed in more detail in [Appendix B](#). In short, the comparison indicates that the overall fit of MOMIT is very good. The failure to find two interactions involving object type as a factor is argued to be due to lack of statistical power, given the modest effect size of the type effect, and due to familiar objects being tracked more accurately in the slow speed condition than predicted by MOMIT.

To sum up, the following findings were obtained in Experiment 1: (1) the MIT performance deteriorated as a function of the number of targets. (2) Performance also deteriorated as the object speed increased. (3) The faster the targets moved and the more targets were to be tracked, the steeper the performance deteriorated. (4) Familiar objects were easier to track than pseudo-objects. (5) The performance deteriorated steeper for pseudo-objects than for familiar objects as a function of set-size. In all, the pattern of results is in good accordance with the predictions derived from MOMIT, except for two interactions that were predicted by MOMIT but not observed in the data.

### 3. Experiment 2A

In the PRPR technique used in Experiment 1, the movement of the elements stops at the time when objects are masked and one of the designated targets is probed by flashing a frame around it. It is possible that some specific phenomenon or a strategy related to static locations may emerge at this stage (e.g., binding is easier to maintain because there is no spatio-temporal processing cost). To eliminate this possibility we developed another variant of the probe recognition task for Experiment 2A. In this variant, elements do not stop moving when masked and when one of the targets is probed but continue to move also during that stage. To successfully cope with this task, the observer needs to maintain the what–where binding even when the target continues to move occluded for another 1.5 s.

### 3.1. Method

#### 3.1.1. Participants

Sixty participants took part in the experiment. None of them had participated in Experiment 1. 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. All participants had normal, uncorrected vision. A pre-selection was made on the basis of their previous school achievement (in mathematics and English).

#### 3.1.2. Apparatus and stimuli

Apparatus and stimuli were identical to those of Experiment 1.

#### 3.1.3. Procedure

Procedure and experimental task were identical to those of Experiment 1 except that the movement of the objects did not stop during the masking and probing phase, but they continued to move in a similar fashion as in the tracking phase (constrained by the similar rules to avoid collisions, etc.). As in Experiment 1, the moving objects were suddenly masked and one of them was flashed for 1500 ms during the probing stage. The movement sequences were generated similarly as in Experiment 1; the tracking phase consisted of 167, 233, or 300 static frames presented one after another for 30 ms each. In Experiment 2A, 50 extra frames were added to the movement files to continue the movement in the probing stage. Object speed was comparable to the medium-speed condition of Experiment 1.

#### 3.1.4. Design

Two factors were manipulated: the number of targets tracked (2–6) and the type of object (familiar object vs. pseudo-objects). Both variables were within-subject variables. There were 24 trials in each of the 10 conditions. The counterbalancing group (familiar objects first vs. pseudo-objects first) was used as between-subject variable to partial out from the total variance a possible practice effect.

### 3.2. Results

#### 3.2.1. Error rate in performance accuracy

The error data were submitted to a  $5 \times 2 \times 2$  mixed design ANOVA. Table 2 shows the mean error rate in performance accuracy as a function of target set-size and object type.

A significant main effect was observed for the number of targets ( $F(4, 232) = 336.15$ ,  $p < .001$ ); performance deteriorated as the number of targets increased. The main effect of object type proved significant, ( $F(1, 58) = 7.68$ ,  $p < .01$ ); familiar objects produced less errors than pseudo-objects. Moreover, the interaction between the number of targets and the type of object was also significant ( $F(4, 232) = 4.62$ ,  $p < .01$ ). A trend analysis showed that this interaction was located in the linear component (linear component accounts for 91.8% of the variance,  $F(1, 58) = 16.35$ ,  $p < .001$ ; other trend components were not significant,  $ps > .25$ ). That is, performance deteriorated steeper for pseudo-objects than for familiar objects as the number of targets increased. Post-hoc comparisons (paired  $t$  tests) indicated that the performance was significantly better for the familiar objects than for the pseudo-objects when five or six targets were tracked (set-size 5:  $t(59) = 1.95$ ,  $p = .059$ ;

Table 2

Percent of trials in which participant made an error in Experiments 2A, 2B, 3A, and 3B, for the five set-sizes (2–6) and the two object types (familiar object vs. pseudo-object)

Experiment	Stimulus type	Set-size									
		2		3		4		5		6	
		<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
2A	Pseudo-object	1.53	0.33	4.03	0.63	11.46	1.51	27.57	1.94	37.08	1.80
	Familiar object	2.43	0.63	3.68	0.73	10.56	1.25	23.19	1.66	32.43	1.73
2B	Pseudo-object	4.54	0.90	8.76	1.14	16.67	1.34	24.20	1.52	29.06	1.43
	Familiar object	2.51	0.53	4.86	0.74	14.96	1.08	22.86	1.36	28.79	1.27
3A	Pseudo-face	2.08	1.19	3.75	1.74	10.52	1.98	19.27	2.35	25.10	2.38
	Familiar face	0.83	0.27	1.15	0.37	7.19	1.22	15.00	1.71	19.69	2.35
3B	Pseudo-face	3.29	0.71	7.02	1.06	16.56	1.53	24.12	1.36	27.08	1.43
	Familiar face	2.85	0.72	5.04	0.96	11.40	1.32	21.71	1.66	25.00	1.31

set-size 6:  $t(59) = 2.36, p < .05$ ). At set-sizes 2–4, performance was at ceiling or near ceiling so that an effect of object type could not be established.

#### 4. Experiment 2B

In the PRPR task used in Experiments 1 and 2A, the participants have to keep active in short-term memory the identity of the probed object during the response selection stage in order to be able to respond accurately to the probe. While our partial report recognition procedure was designed minimize memory requirements it may still be argued that familiar stimuli may be easier than unfamiliar stimuli to keep active in short-term memory during this short response selection stage (the set-size and speed effects are not subject to this criticism). To eliminate this possibility (we think it is not very likely, see Section 4.3), we created another version of the MIT task that further minimized the memory requirement and the need to consult or maintain identity information in providing the response. Our other MIT task is a variant of the change detection paradigm (e.g., Rensink et al., 1997; Simons & Levin, 1998). In this task, the movement phase is identical to the PRPR task except that the moving objects are surrounded by frames that move along with the objects. At some point in time the frames go blank for 120 ms, but the empty frames continue to move. When the objects reappear, they appear either in their original frames, or in half of the time two of the targets swap position. Objects continue to move for another second before the response screen is presented. The participant's task is to respond whether or not a change took place (see Fig. 5 for a schematic illustration of the MIT task with the change

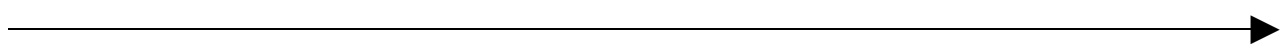
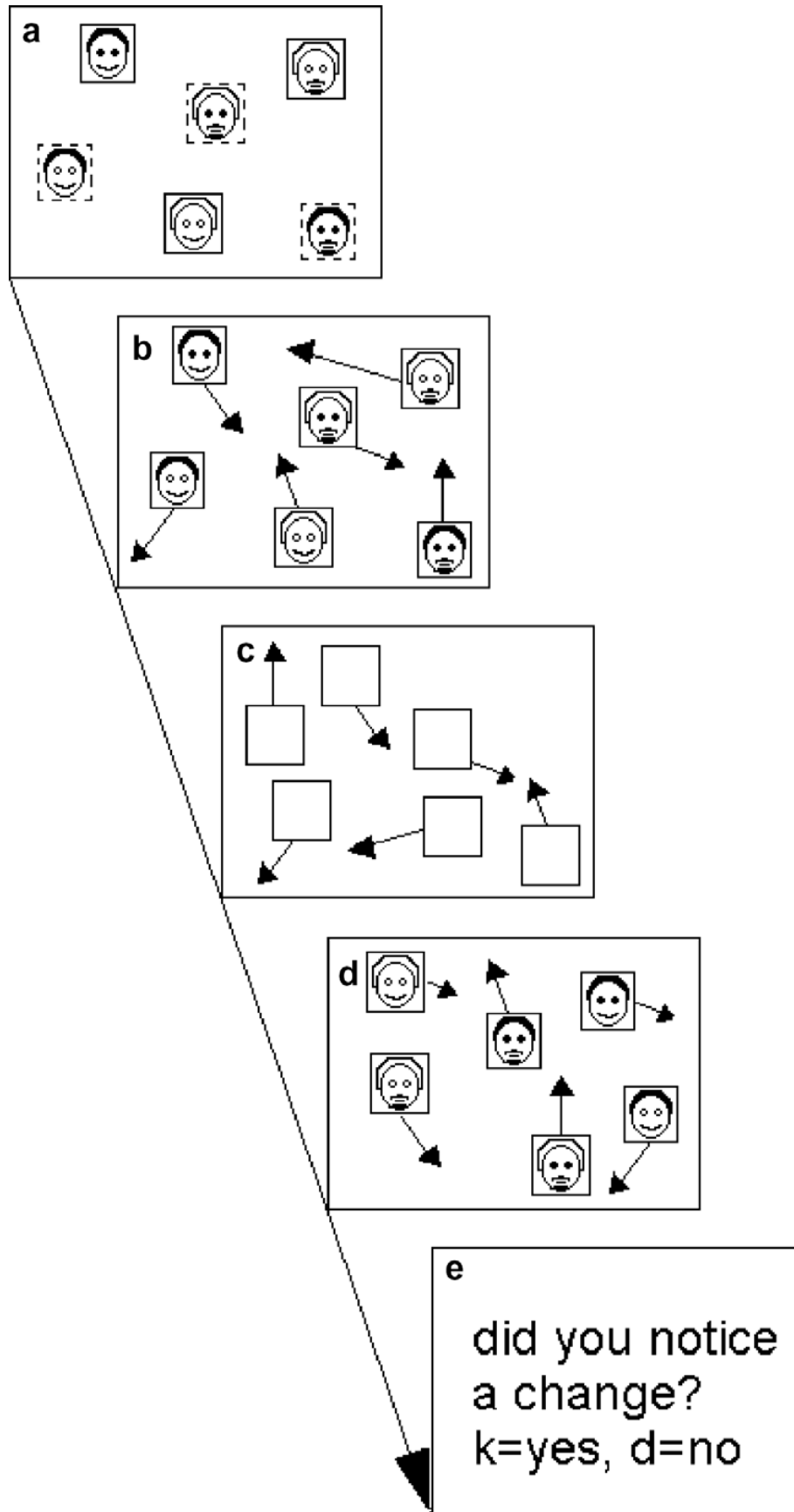


Fig. 5. A schematic depiction (exactly these stimuli were not used in any of the experiments reported) of the multiple-identity tracking task with the change detection technique (not to scale) used in Experiments 2B and 3B. Display a: six different objects are presented, and 2–6 of them are designated as targets by flashing a frame around them. Display b: all of them begin to move randomly about the screen. Display c: a flicker takes place during which pictures disappear and participants see empty moving framed squares. Two of the designated targets may swap positions. Display d: the pictures reappear in the frames and move still about a second. Display e: a response screen appears and participants are to decide if a change (a swap of two pictures) took place or not among designated targets by pressing an appropriate key in the keyboard.



detection procedure). Thus, the procedure is similar to the one in Experiment 2A in that objects continue their movement during the probing phase. On the other hand, the change detection procedure differs from the PRPR task in the response phase: only a ‘yes’ or ‘no’ response is required—no identity information needs to be consulted for the response.

We like to emphasize, that although change detection procedure may prevent the need to maintain identity information in the response phase, the change detection task may be highly difficult and cognitively demanding task before the response phase. During the blank period, identity of the tracked objects have to be effectively maintained in short-term memory or episodic buffer in order to carry out the comparison and decision process whether change took place or not. This comparison/decision process may be cognitively very demanding and subject to decision bias (see discussion of the difficulty of comparison process in change detection, [Simons & Rensink, 2005](#)).

#### 4.1. Method

##### 4.1.1. Participants

Seventy-eight participants took part in the experiment. None of them participated in Experiment 1 or 2A. 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. All participants had normal, uncorrected vision. A pre-selection was made on the basis of their previous school achievement (in mathematics and English).

##### 4.1.2. Apparatus

The apparatus was the same as in Experiment 1.

##### 4.1.3. Stimuli

The stimuli were the same as in Experiment 1 with two exceptions: no masks were used at the end of trial, and black frames ( $75 \times 75$  pixels, 2 pixels in width) surrounded the objects throughout the trial. The frames were needed because the change detection technique involves a flicker event where pictures disappear for a short moment (120 ms). To keep the task dynamic, the empty framed squares continued moving during the flicker period.

##### 4.1.4. Task

[Fig. 5](#) provides a schematic illustration of the MIT task with the change detection procedure. The first task phase was identical to that of Experiment 1. At the beginning of each trial the targets (2–6) were designated by flashing a frame around each of them. The participants’ task was to track these designated targets during the movement phase. However, in distinction to Experiment 1, participants were instructed to also pay attention to a flicker during the movement phase, during which two targets may swap position. Their task was to respond whether or not two targets swapped position during this flicker, by pressing the K or D key in the computer keyboard (K appeared in green and D in red). They were encouraged to guess if they did not know the correct answer. The index finger of the right and left hand was used for pressing the response keys.

#### 4.1.5. Procedure

The procedure was identical to that of Experiment 1 with the following exceptions. As in Experiment 1, in the beginning of each trial the items were displayed for 1 s, but unlike in Experiment 1, now a black frame surrounded each object. A flicker (lasting for 120 ms) took place 1110 ms before the end of the trajectory duration. During the flicker, the participants saw empty framed squares moving. After that the pictures reappeared inside the frames and moved still for 990 ms. The movement phase after the flicker was needed to give the participants time to detect whether or not a change had taken place. In half of the trials all pictures reappeared within their own frames, in the other half of the trials two targets swapped position. Object speed was comparable to the medium-speed condition of Experiment 1.

#### 4.1.6. Design

The design was identical to that of Experiment 2A.

### 4.2. Results

#### 4.2.1. Error rate in performance accuracy

The error data were submitted to a similar  $5 \times 2 \times 2$  mixed-design ANOVA as in Experiment 2A. Table 2 shows the means for the error rate in performance accuracy as a function of target set-size and object type. A significant main effect was found for the number of targets ( $F(4, 304) = 242.20, p < .001$ ) and for the type of object ( $F(1, 76) = 7.35, p < .01$ ). The interaction between the number of targets and the type of object was not significant ( $F(4, 304) = 1.54, p > .10$ ). The absence of this interaction and the observed main effect for the type of object indicates that, in contrast to Experiment 1, the semantic effect was now present even in smaller set-sizes, but it did not increase as a function of set-size.

#### 4.2.2. Sensitivity and response bias

The parametric sensitivity index  $d'$  was computed from the hit rates and false-alarm rates. The  $d'$  values are shown in Table 3.

The  $d'$  values were submitted to a similar  $5 \times 2 \times 2$  mixed-design ANOVA as above. The ANOVA yielded a significant main effect for the number of targets ( $F(4, 304) = 305.11, p < .001$ ); sensitivity deteriorated as the number of targets increased. The main effect of target type was also significant ( $F(1, 76) = 8.66, p < .01$ ); sensitivity was better for familiar objects than for pseudo-objects. Moreover, the interaction between the number of targets

Table 3

Mean value of  $d'$  for the five set-sizes (2–6) and the two object types (familiar object vs. pseudo-object) in Experiments 2B and 3B

Experiment	Stimulus type	Set-size									
		2		3		4		5		6	
		<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
2B	Pseudo-object	3.90	0.11	3.36	0.14	2.47	0.13	1.77	0.13	1.38	0.11
	Familiar object	4.19	0.08	3.84	0.11	2.56	0.12	1.84	0.11	1.40	0.10
3B	Pseudo-face	4.04	0.12	3.47	0.15	2.36	0.16	1.69	0.13	1.56	0.12
	Familiar face	4.15	0.12	3.76	0.14	2.94	0.17	1.92	0.15	1.63	0.11

and the type of object also proved significant ( $F(4, 304) = 2.79, p < .05$ ). A trend analysis showed that this interaction was primarily located in the linear component (the linear component accounts for 61.4% of the variance,  $F(1, 76) = 3.56, p < .05$ ; other trend components were not significant,  $ps > .10$ ). Post-hoc comparisons (paired  $t$  tests) indicated that sensitivity was significantly better for the familiar objects than for the pseudo-objects only when two or three targets were tracked (set-size 2:  $t(77) = -2.05, p < .05$ ; set-size 3:  $t(77) = -3.19, p < .01$ ).

The final ANOVA was performed on the nonparametric index of response bias, the criterion cutoff  $C$  that was computed from the hit and false-alarm rates. The  $C$  values are shown in Table 4.

In the ANOVA, a significant main effect was found only for the number of targets ( $F(4, 304) = 15.75, p < .001$ ), but the main effect of target type and the set-size  $\times$  type interaction were not significant (both  $F < 1$ ). The set-size effect suggests a criterion shift in the task performance. As is evident from Table 4, the  $C$  value is neutral (i.e., close to 0) for set-size 2 and 3, but in the larger set-sizes the response criterion shifts to positive (i.e., a bias toward a 'no' response). That is, the participants became more conservative in responding (i.e., if in doubt, a 'no' response was given).

In sum, the set-size effect and the object type effect observed using the PRPR task were replicated in Experiment 2B with a change detection procedure. Now the semantic effect in error rates was significant also in the smaller set-sizes. The signal detection analysis of sensitivity demonstrated that the participants were susceptible to a conservative response bias in the larger set-sizes.

#### 4.3. Discussion of Experiments 2A and 2B

Experiment 2A differed from Experiment 1 in that object movement continued also in the probing stage. Despite this difference, the main effects of target-set size and object type were replicated. These effects were also observed in Experiment 2B, where a change detection procedure was used to measure tracking performance. In all, the pattern of results is largely consistent with the predictions of MOMIT, with one exception. The nature of the object type  $\times$  set-size interaction in  $d'$  calculated for the change detection accuracy was not the kind predicted by MOMIT. This result is discussed in more detail in Section 9.

The partial report recognition technique (used in Experiments 1 and 2A) can be criticized by pointing out that it requires a response to be made on the basis of identity information. According to this criticism, a semantic effect may arise due to the fact that familiar

Table 4

Mean value of the criterion cutoff  $C$ , for the five set-sizes (2–6) and the two object types (familiar object vs. pseudo-object) in Experiments 2B and 3B

Experiment	Stimulus type	Set-size									
		2		3		4		5		6	
		<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
2B	Pseudo-object	0.07	0.04	0.10	0.04	0.29	0.05	0.23	0.06	0.31	0.05
	Familiar object	0.03	0.03	0.04	0.04	0.27	0.05	0.24	0.05	0.31	0.06
3B	Pseudo-face	0.06	0.06	0.09	0.07	0.19	0.08	0.37	0.07	0.37	0.09
	Familiar face	0.02	0.06	0.10	0.06	0.20	0.07	0.29	0.07	0.32	0.08

items are easier to maintain in short-term memory than unfamiliar items during the response stage and thus the task would not measure object tracking but instead would primarily reflect short-term memory constraints. However, the pattern of our results does not support this criticism: (1) if performance in the PRPR task only reflected different memory requirements present during the response stage, then a speed manipulation should not have any effect on performance. The significant speed effect observed in Experiment 1 demonstrates that a spatiotemporal manipulation during the movement phase significantly affected performance and that the PRPR task is indeed sensitive to index dynamic tracking. (2) A clear semantic effect was observed also using the change detection technique where no identity information is required for the response. (3) The set-size effect observed with the PRPR technique is not readily interpretable as a STM effect.

As the next step, we investigated whether the results of Experiments 2A and 2B may be replicated using other type of visual stimuli. In Experiments 3A and 3B, we contrasted the tracking performance of familiar faces to that of ‘pseudo-faces’. In Experiment 3A, we used the partial-report probe recognition task, while the change detection procedure was employed in Experiment 3B.

## 5. Experiment 3A

In Experiment 3A, familiar faces and ‘pseudo-faces’ were used as stimuli. Pseudo-faces were created from familiar faces by deconstructing them and then rearranging and recombining the pieces of familiar faces. The face-like appearance of the pseudo-faces was preserved (pseudo-faces have eyes, nose, hair, etc.). Nevertheless, they give an odd appearance, which is why we named them ‘frankensteins’. What is crucial in the present context, the pseudo-faces lack a known identity (analogously to pseudo-objects used in Experiments 1, 2A, and 2B). It should also be noted that the familiar faces and pseudo-faces are visually identical at the pixel level, because the pseudo-faces were created from the familiar faces. In Experiment 3A, we used the same partial report technique as in Experiment 1.

### 5.1. Method

#### 5.1.1. Participants

Forty participants were tested. None of them took part in any of the previous experiments. Their age ranged from 20 to 36, the median was 22 years (93% was under 30 years of age). The experiment was conducted as a part of a large test battery that was administered to civilian air company applicants. All participants had normal or corrected-to-normal vision (with maximum of  $\pm 1.5$  diopter). A pre-selection was made on the basis of the applicants’ previous school achievement (in mathematics and English), standardized reasoning tests, and clinical personality tests. This screening was done as a part of the air pilot recruitment process.

#### 5.1.2. Stimuli

Two sets of six stimuli were used, familiar faces and pseudo-faces. As familiar faces, six vertically oriented line drawings of faces of familiar persons (Elvis Presley, Mauno Koivisto, Bill Clinton, Sylvester Stallone, Saddam Hussein, and Albert Einstein) were used (see Fig. 6). The faces were selected to be (1) as recognizable as possible as smallish line draw-

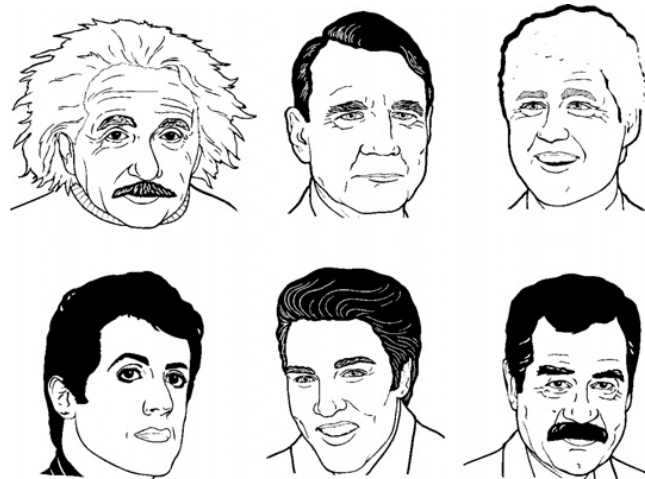


Fig. 6. The target pictures of the familiar faces used in Experiments 3A and 3B (top-row from left to right: Albert Einstein, Mauno Koivisto, and Bill Clinton; bottom-row from left to right: Sylvester Stallone, Elvis Presley, and Saddam Hussein). The faces were selected from a commercial package of digitized black-and-white drawings (Corel Mega Gallery, 1996, © Corel Corporation, all rights reserved).

ings, and (2) to represent different semantic categories (politicians, actor, singer, scientist). When piloting with familiar faces, it appeared that it was difficult to fully meet both objectives (e.g., we did not find clearly recognizable athletes or artists from our picture gallery). Thus, we gave familiarity first priority and chose three faces belonging to the category of international politicians (Koivisto, Clinton, and Hussein). However, despite belonging to the same category, these three persons are semantically highly distinguishable, for example in terms of nationality (former presidents of Finland, USA, and Iraq). The faces were selected from a commercial package of digitized black-and-white drawings (Corel Mega Gallery, 1996). Slight adjustments were made to the bitmaps of the drawings to make them comparable to each other. In a pretest given to the study participants, 75% was able to give a correct name to all the familiar faces.

Pseudo-faces (we coined them ‘frankensteins’) were created from familiar faces by deconstructing them (i.e., each face was cut into an upper and lower half, and facial components were separated) and then rearranging and recombining the parts (see Fig. 7). For



Fig. 7. The target pictures of the pseudo-faces used in Experiments 3A and 3B.

instance, the top-left frankenstein in Fig. 7 comprised the upper half of Stallone's face and the lower half of Saddam's face, Clinton's left eye, Elvis' right eye, Koivisto's nose, and Einstein's mouth and moustaches. The face-like appearance of the pseudo-faces was preserved (pseudo-faces have eyes, nose, hair, etc.). The other stimulus details were the same as in Experiment 1 except that the masks were created from facial stimuli and the width of the pictures ranged from 56 to 75 pixels.

### 5.1.3. Apparatus, task, procedure, and design

The experimental task, procedure, trajectory files, and apparatus were the same as in Experiment 1. Object speed was comparable to the medium-speed condition of Experiment 1. The design was identical to that of Experiments 2A and 2B.

## 5.2. Results and discussion

### 5.2.1. Error rate in performance accuracy

The error data were submitted to a similar  $5 \times 2 \times 2$  mixed-design ANOVA as in Experiments 2A and 2B. Table 2 shows the means for the error rate in performance accuracy as a function of target set-size and target type.

A significant main effect was observed for the number of targets ( $F(4, 152) = 102.19$ ,  $p < .001$ ). In addition, a main effect of target type proved significant ( $F(1, 38) = 7.65$ ,  $p < .01$ ), but set-size  $\times$  target type interaction remained non-significant,  $F(4, 152) = 1.41$ ,  $p > .10$  (however, the linear component of the interaction just missed significance,  $F(1, 38) = 3.82$ ,  $p = .058$ , and accounts for 99.2% of the variance). The lack of an interaction indicates that the semantic effect (i.e., a better performance for familiar faces than for pseudo-faces) was present regardless of the number of targets.

## 6. Experiment 3B

In Experiment 3B, tracking of familiar faces and pseudo-faces was examined using the change detection paradigm introduced in Experiment 2B. The facial stimuli were the same that were used in Experiment 3A.

### 6.1. Method

#### 6.1.1. Participants

Thirty-eight participants took part in the experiment. None had participated in any of the previous experiments. Their age ranged from 20 to 44, the median was 23 years (76% was under 30 years). The experiment was conducted as a part of a large test battery that was administered to civilian air company applicants. All participants had normal or corrected-to-normal vision (with maximum of  $\pm 1.5$  diopter). The same pre-selection procedure was followed as in Experiment 3A.

#### 6.1.2. Stimuli

The facial stimuli were identical to those of Experiment 3A. In a pretest given to the study participants, 83% was able to name correctly all the familiar faces.

### 6.1.3. Apparatus, task, procedure, and design

The apparatus and trajectory files were the same as in Experiment 1. Object speed was comparable to the medium-speed condition of Experiment 1. The experimental task and procedure were the same as in Experiment 2B. The design was identical to that of Experiment 2A.

## 6.2. Results and discussion

### 6.2.1. Error rate in performance accuracy

The error data were again submitted to a  $5 \times 2 \times 2$  mixed-design ANOVA. Table 2 shows the means for the error rate in performance accuracy as a function of target set-size and target type.

The pattern of results replicated those observed in Experiment 2B, where the change detection technique was used with different stimuli. A significant main effect was found for the number of targets ( $F(4, 144) = 149.86, p < .001$ ) and for the type of target ( $F(1, 36) = 10.37, p < .01$ ), but their interaction was not significant ( $F(4, 144) = 1.68, p > .10$ ). The absence of an interaction and the observed main effect of target type indicate that the semantic effect (i.e., a better performance for familiar faces than for pseudo-faces) was present regardless of the number of targets.

### 6.2.2. Sensitivity and response bias

The  $d'$  and  $C$  values were submitted to a  $5 \times 2 \times 2$  mixed-design ANOVA. The  $d'$  values are shown in Table 3 and the  $C$  values in Table 4.

The ANOVA of the  $d'$  scores yielded a significant main effect for the number of targets ( $F(4, 144) = 161.93, p < .001$ ) and for the target type ( $F(1, 36) = 8.66, p < .01$ ), but their interaction was not significant ( $F(4, 144) = 1.75, p > .10$ ). The absence of an interaction and the observed main effect of target type indicate that a semantic effect was present regardless of the number of targets.

In the ANOVA of the  $C$  scores, a significant main effect was found for the number of targets ( $F(4, 144) = 9.09, p < .001$ ), but the main effect of target type and the set-size  $\times$  target type interaction remained clearly non-significant (both  $F < 1$ ). The significant main effect of the number of targets suggests that the response criterion shifted from neutral to conservative as the set-size increased, analogously to Experiment 2B (see Table 4).

To sum up, Experiment 3B demonstrated a clear semantic effect across the set-sizes; familiar faces were easier to track than pseudo-faces. Second, a effect of set-size was obtained; the tracking performance deteriorated as the number of tracked targets increased. A conservative response bias was found in the larger set-sizes as in Experiment 2B. In all, the change detection data of Experiment 3B closely replicated those obtained for objects and pseudo-objects in Experiment 2B and those obtained for facial stimuli using the PRPR task.

## 7. A mathematical formulation of MOMIT

In the following we provide a mathematical formulation of MOMIT. The aim is to model performance accuracy of dynamic identity-location binding as a function of target set-size, object speed, and object familiarity. The mathematical model consists of three

components: binding capacity, probability of guessing, and dynamic processing cost. We first describe each component separately, after which the three components are integrated into a single formula that allows us to fit the model to the empirical data.

### 7.1. Binding capacity $m$

The binding capacity  $m$  is an estimate of the number of *static*, non-moving identity-location bindings that can be held temporarily active. If the number of to-be-remembered items is  $n$ , the probability of having access to a binding is  $m/n$ , when guessing probability is not considered. Thus, when  $m \geq n$  the probability of reporting a probed object is 1; when  $m < n$  the probability of correct response is  $m/n$  plus probability of guessing. Here we have made use of the formula introduced by Scholl, Pylyshyn, and Feldman (2001); note, however, that they interpret  $m$  differently from us.

### 7.2. Probability of guessing ( $P_{\text{guess}}$ )

In addition to the binding capacity, the probability of guessing is also assumed to affect performance accuracy.  $P_{\text{guess}}$  is influenced by the number of response alternatives and by possible strategies adopted by the participant. The probability of correctly reporting identity-location bindings is assumed to vary as a function of correctly remembering  $m$  items and guessing not explicitly remembered items. By adapting the formula proposed by Scholl et al. (2001), the performance accuracy with static objects may be estimated as

$$P_{\text{static}} = \frac{m}{n} + \left(1 - \frac{m}{n}\right) * P_{\text{guess}}. \quad (1)$$

In other words, performance accuracy varies as a function of binding capacity  $m$ , the number of to-be-remembered items  $n$ , and the probability of guessing ( $P_{\text{guess}}$ ) non-remembered items  $1 - m/n$ . Eq. (1) in fact expresses a version of a fixed-capacity parallel model; we fitted also this model to the data (see below).

### 7.3. Dynamic processing cost ( $P_e$ )

The third component affecting dynamic binding performance is the processing cost related to creating and reactivating serially identity-location bindings. Recall that the target coordinates stored in VSTM are only updated when the target is focally attended. Thus, in a dynamic environment these stored coordinates are completely correct only at the time of updating. All other times they are incorrect, in other words, there is *coordinate error* ( $e$ ). In the following, we derive a formula that estimates the size of this coordinate error, which in turn contributes to the overall performance accuracy.

Let's assume that the observer is exposed to a moving stimulus set  $A$ , which consists of  $n$  visually distinct targets that move around with an average speed of  $\bar{v}_A$  (degrees per second).<sup>4</sup> The average processing time (including the time elapsed in shifting the attention from one target to another) required to create or refresh a binding among the stimulus set  $A$  is denoted by  $\bar{s}_A$ . The time needed to refresh all targets among the set is thus

<sup>4</sup> Visually identical objects constitute a special case that is not dealt with here.

$n * \bar{s}_A$ . The model assumes that bindings are created and refreshed serially one at a time. Thus, after creating a binding for target  $k$  and temporarily storing its current location coordinates in VSTM the same is done for other targets. During this time target  $k$  moves a distance estimated by  $(n - 1) * \bar{s}_A * \bar{v}_A$ . This is the average coordinate error for any target among the stimulus set. Thus,

$$e = (n - 1) * \bar{s}_A * \bar{v}_A, \tag{2}$$

where  $e$  is the distance target  $k$  has moved from the spatial location it occupied when an identity-location was created for it until it is again focally attended. If objects do not move (i.e.,  $\bar{v}_A = 0$ ), then  $e$  is of course zero.  $e$  varies as a function of target set-size, object speed, and object familiarity, as all these factors affect the time it takes to complete a refresh cycle (see above for the functional description of MOMIT). The more targets there are, the faster they move, and the less familiar they are, the more time it takes to complete a refresh cycle. It is assumed that when  $e$  is small performance is not much deteriorated, while a larger  $e$  is associated with marked performance decrements. This is because with a small  $e$  the target is either correctly attended at once or if an erroneous object is attended first, the intended one is readily perceived when it is still within the foveal vision (or very close to it). On the other hand, performance is assumed to deteriorate quadratically as  $e$  increases. A quadratic relation is implicated by the fact that the searched space expands as a second-order function of distance (analogously to the circle's area being determined by its radius,  $\Pi r^2$ ). This probability function  $P_e$  of the processing cost  $e$  should be 1 when  $e$  is less than 2 deg of visual angle (e.g., the foveal area). On other hand, it reaches 0 when  $e$  is sufficiently large.

$$P_e = a * e^2 + b * e + c. \tag{3}$$

In Formula 3,  $a$  and  $b$  are coefficients and  $c$  is a constant. As  $P_e = 1$  when  $e = 0$ , then  $c = 1$ . In addition, it is theoretically assumed that  $P_e$  will become 0 when  $e$  is sufficiently large. Let's denote this value with  $x$ . We can then solve  $b$ :  $0 = a * x^2 + b * x + 1$ ;  $b = -\frac{ax^2+1}{x}$

Thus, Formula 3 may be reformulated as

$$P_e = a * e^2 - \frac{ax^2 + 1}{x} * e + 1. \tag{4}$$

Finally, the three components are brought together in the following two formulas:

$$P_{\text{dynamic}} = P_{\text{static}} * P_e \quad \text{or} \tag{5}$$

$$P_{\text{dynamic}} = \left( \frac{m}{n} + \left( 1 - \frac{m}{n} \right) * P_{\text{guess}} \right) * \left( a * e^2 - \frac{ax^2 + 1}{x} * e + 1 \right), \tag{6}$$

where  $e = (n - 1) * \bar{s}_A * \bar{v}_A$ .

The Formula 6 can be reformulated with regard to  $n$ , then

$$P_{\text{dynamic}} = k_1 + k_2 * n + k_3 * n^2 + k_4 * \frac{1}{n}, \tag{7}$$

where

$$\begin{aligned}
k_1 &= -2 * a * \bar{s}_A^2 * \bar{v}_A^2 * m * (1 - P_{\text{guess}}) - \frac{(a * x^2 + 1) * \bar{s}_A * \bar{v}_A * m * (1 - P_{\text{guess}})}{x} \\
&\quad + P_{\text{guess}} * a * \bar{s}_A^2 * \bar{v}_A^2 + \frac{P_{\text{guess}} * \bar{s}_A * \bar{v}_A * (a * x^2 + 1)}{x} + P_{\text{guess}}, \\
k_2 &= - \left\{ (2 * P_{\text{guess}} * a * \bar{s}_A^2 * \bar{v}_A^2) + \frac{(a * x^2 + 1) * \bar{s}_A * \bar{v}_A * P_{\text{guess}}}{x} \right. \\
&\quad \left. - a * \bar{s}_A^2 * \bar{v}_A^2 * m * (1 - P_{\text{guess}}) \right\}, \\
k_3 &= P_{\text{guess}} * a * \bar{s}_A^2 * \bar{v}_A^2, \\
k_4 &= m * (1 - P_{\text{guess}}) * \left( a * \bar{s}_A^2 * \bar{v}_A^2 + \frac{(a * x^2 + 1) * \bar{s}_A * \bar{v}_A}{x} + 1 \right).
\end{aligned}$$

As can be seen from Formula 7, Formula 6 does not reduce to a polynomial equation (it includes a component  $1/n$ ).

As is evident from Formula 6, dynamic binding performance is affected by

$m$  = binding capacity,

$n$  = number of tracked objects,

$P_{\text{guess}}$  = probability of guessing,

$a$  = convexity of the probability function  $P_e$  (visual acuity function regarding  $e$ ),

$x$  = the value when  $P_e$  reaches 0,

$e$  = coordinate error (in visual angle),

$\bar{s}_A$  = average time required to refresh a binding among the stimulus set  $A$ ,

$\bar{v}_A$  = average object speed among the stimulus set  $A$ .

## 8. Model fitting

We fit MOMIT to the data of Experiment 1 using Eq. (6), which includes 7 parameters, two of which,  $s$  and  $m$ , are free parameters. The parameters that were not free were fixed on the basis of empirical data ( $a, x$ ) or were derived from the structure of the experiment ( $n, v, p$ ). Parameters  $a$  and  $x$  determine the probability function  $P_{\text{ce}}$  in relation to visual acuity. Parameter  $a$  determines the curvature of the function, while parameter  $x$  determines the location where  $P_{\text{ce}}$  reaches 0. For  $a$ , we aimed for a value which yields flawless performance for a coordinate error of less than 2 deg of visual angle (i.e., the span of foveal vision);  $a = -0.0022$  was found to yield such a function. To find a reasonable estimate for  $x$ , we conducted a small peripheral object recognition experiment with the stimuli of Experiment 1. We observed that object recognition was at a chance level, when objects were more than 20 deg away from the center fixation point. Next we ran simulations with values of 20–25. The simulations showed that all these values gave reasonable fits to the data. Thus,  $x$  was set to 22 deg. The parameter values for  $n$  and  $v$  were derived from the design of Experiment 1:  $n$  varied from 2 to 6;  $v$  was set to 2.6, 6.3 or 10.7. The guessing probability  $p$  was set to reflect a rather effective guessing strategy,  $P_{\text{guess}} = 1/(6 - m)$ . The use of an effective guessing strategy is possible by remembering a subset of objects that are tracked correctly and then using this information to narrow down the choice for guessing. For instance, let us assume that there are six designated targets and the participant can accurately track three of them. If the participant notices that the probed item is not among the accurately tracked items and (s)he is able to remember the identities of

the correctly tracked targets, (s)he may then be able to limit her/his guessing to the three targets (s)he knows (s)he did not track. Thus, the guessing probability becomes 1/3. In general, this yields a guessing probability of  $1/(6 \text{ minus the number of accurately tracked and remembered objects})$ .

With respect to the free parameters, we assumed the static binding capacity  $m$  to be approximately 4 (see e.g., Oksama & Hyönä, 2004; Luck & Vogel, 1997) and a plausible range for the values of the refresh time parameter  $s$  to be around 250–500 ms, which is the time taken to fixate a static object for recognition (De Graef, Christiaens, & d'Ydewalle, 1990; Germeys, De Graef, & Verfaillie, 2002; Henderson, Pollatsek, & Rayner, 1989), or to serially shift attention between objects (Horowitz, Holcombe, Wolfe, Arsenio, & DiMase, 2004).

The MOMIT model with two free parameters was fit to 9 mean data points (three speed conditions and set-sizes 4, 5, and 6) using the SPSS 14 (SPSS, Inc. Chicago, IL) nonlinear regression program. Set-sizes 2 and 3 were excluded from the model fitting, because with the initial value of  $m = 4$ , for set-sizes 2 and 3  $m$  would be greater than  $n$  and the equation would in that case yield a performance accuracy estimate of over 1. 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 of familiar objects and pseudo-objects, because MOMIT assumes  $s$  to be different for the two object types.

Estimates of model fit and the best fitting parameters are presented in Table 5. As can be seen, MOMIT with two free parameters (Run 1 and 2) yields a very nice fit to the data of both stimulus types with small standard errors, sufficiently narrow confidence intervals, and over .9 coefficient of determinations. Perhaps most importantly, the best fitting values for  $s$  (238 and 241 ms) and  $m$  (3.9 and 3.7) are psychologically plausible. The estimate of  $m$  is very close to the aforementioned value of 4 proposed by Luck and Vogel (1997); the estimate of  $s$  is also within a plausible range (see above).

The next step was to experiment with the refresh time parameter  $s$ . MOMIT assumes a different  $s$  value for familiar and unfamiliar stimuli but a common  $m$ . Thus, we fixed  $m$  to 3.8 (an average of 3.7 and 3.9 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 5. The best fitting  $s$  for familiar objects was 230 ms and for pseudo-objects 250 ms—a difference of 20 ms. The model fit is very good with narrow confidence intervals, small standard errors, and a high  $R^2$ . This difference in  $s$  is assumed to reflect a semantic familiarity effect. It may be argued that as such the effect should be somewhat bigger. It may be noticed, however, that the confidence intervals allow the effect to be considerably bigger.

A semantic effect may also be materialized by assuming that  $m$  varies as a function of object familiarity but both stimulus types share a common  $s$ . According to this view, the semantic effect is not modulated by a dynamic refresh cost but by a different storage capacity for familiar and unfamiliar objects. We experimented with this alternative; the results are reported as Run 5 and 6 in Table 5 ( $s$  was fixed to 240 ms, which is the average of Run 1 and 2). The simulation yields a difference of 0.2 objects in  $m$  between the familiar and pseudo-objects, and the model fits well to the data. Although this version is derived from the mathematical formulation of MOMIT, it is not consistent with the functional architecture of the model, where long-term memory representations are assumed to be consulted in creating and refreshing serially identity-location bindings. At present we are not in a position to strongly argue for or against either alternative. Nevertheless, we consider it likely that at least  $s$  varies as a function of familiarity.

Table 5  
Parameter estimates and measures of goodness of fit of the different runs of the MOMIT model

Run #	Number of data points	Procedure	Stimuli	Mathematical model	Fixed parameter values	Best fitting free parameter	Std. error	95% Confidence Interval		R <sup>2</sup>
								Lower Bound	Upper Bound	
1	9	Probe recognition	Familiar objects	MOMIT	—	$m = 3.90$ $s = .238$	.12 .016	3.60 .200	4.19 .277	.91
2	9	Probe recognition	Pseudo-objects	MOMIT	—	$m = 3.69$ $s = .241$	.08 .012	3.50 .212	3.89 .270	.95
3	9	Probe recognition	Familiar objects	MOMIT	$m = 3.8$	$s = .230$	.013	.200	.260	.90
4	9	Probe recognition	Pseudo-objects	MOMIT	$m = 3.8$	$s = .250$	.009	.229	.271	.94
5	9	Probe recognition	Familiar objects	MOMIT	$s = .240$	$m = 3.90$	0.09	3.70	4.11	.91
6	9	Probe recognition	Pseudo-objects	MOMIT	$s = .240$	$m = 3.69$	0.06	3.55	3.83	.95
7	9	Probe recognition	Familiar objects	Parallel	—	$m = 3.39$	0.18	2.97	3.81	.40
8	9	Probe recognition	Pseudo-objects	Parallel	—	$m = 3.22$	0.17	2.83	3.62	.47
9	4	Change detection	Familiar objects	MOMIT	$s = .230$	$m = 2.73$	0.02	2.68	2.78	1.0
10	4	Change detection	Pseudo-objects	MOMIT	$s = .250$	$m = 2.58$	0.10	2.26	2.90	.95

The parallel model is based on the formula proposed by Scholl et al. (2001).

As the next step, we fitted the data of Experiment 1 to a one-parameter parallel model. This was easy to do because such a model is nested in MOMIT. By setting  $s = 0$  Eq. (6) reduces to Eq. (1). Eq. (1) can in turn be interpreted as a version of a fixed capacity parallel model with a single parameter  $m$ , where  $m$  is the parallel tracking capacity, i.e., the number of sticky fingers in the FINST theory (Pylyshyn, 1989), or the number of foci of attention in the theory of Cavanagh and Alvarez (2005). As the two models are nested, their statistical comparison is possible using the extra sum of squares test.

Estimates of the parallel model are presented in Run 7 and 8 (see Table 5). The parallel model does not provide an excellent fit to the data in terms of  $R^2$ . An obvious reason is that this simple fixed-capacity parallel model does not capture the speed effect. The extra sum of squares test shows that MOMIT with two free parameters yields a significantly better fit to the data than the one-parameter parallel model, both for familiar objects,  $F(1, 7) = 37.5$ , and for pseudo-objects,  $F(1, 7) = 68.8$ , both  $p < 0.001$ . In addition, the best fitting  $m$  values for familiar (3.4) and pseudo-objects (3.2) are somewhat lower than what is assumed in the literature for parallel tracking capacity (4 or 5).

As the final step, we examined MOMIT's capacity to predict the change detection data of Experiment 2B. This experiment had only one speed condition, so there are fewer data points to fit (in the following we use four data points, set-sizes 3–6). To avoid overfitting and too wide confidence intervals, we fixed parameter  $s$  to the values derived from Experiment 1 (we assume that  $s$  is same for both experimental procedures) and let only  $m$  to vary freely. We initially assumed that the probability of guessing is 0.5 in this two-choice task. However, there is a strategy by which participants may improve their change detection performance. Because a pair of targets swaps position in half of the trials, in principle participants need to track one fewer than the total number of designated targets to successfully detect a change. For example, with two targets, the participant only needs to track one target to perform the task successfully. Similarly, with three targets the participant only needs to track two, and so forth. If the availability of this strategy is discovered, it is very simple to use. If observers consistently used it, then the model's predictions with a guessing probability of 0.5 would underestimate the performance accuracy. We tried both 0.5 and 0.6 and found that 0.6 yielded a slightly better result in terms of confidence intervals. The results are presented as Run 9 and 10 in Table 5. The model fitted very well with the data with narrow confidence intervals, small standard errors, and very high  $R^2$ . However, the best fitting  $m$  values are now about one object smaller (2.6 and 2.7) than what was estimated using the probe recognition task. In Section 9, we provide an explanation for the difference in the binding capacity estimates between the two tasks employed to measure performance accuracy in MIT.

In sum, the mathematical formulation of MOMIT captures the effects of object speed, target set-size, and object familiarity very well using plausible estimates of refresh time and binding capacity. It was also demonstrated that MOMIT provided a significantly better fit to the data than a one-parameter parallel model.

## 9. General discussion

The present study investigated the observer's ability to track and maintain multiple identities in a dynamic visual environment. In five experiments, we examined effects of target set-size, semantic identity information, and object speed on performance accuracy. Two experimental techniques (partial report probe recognition and change detection) were

employed to study tracking of familiar objects, pseudo-objects, familiar faces, and pseudo-faces. The following findings were obtained: (1) a highly significant target set-size effect was observed in all experiments; performance deteriorated as a function of the number of targets. (2) A significant semantic effect was observed in all experiments; familiar objects or faces were easier to track than pseudo-objects or pseudo-faces. Moreover, this main effect was modulated by target set-size in two experiments where tracking was measured using the PRPR technique. That is, performance deteriorated steeper for pseudo-objects than familiar objects when the target set-size increased. (3) A highly significant speed effect was observed; performance deteriorated as target speed increased. Furthermore, the speed effect was modulated by target set-size; the faster the targets moved and the more targets were to be tracked, the more pronounced was the performance decrement.

The mathematical formulation of the MOMIT model of multiple identity tracking captured all these observed effects and produced an excellent quantitative fit to the data. The main tenets of the model are that continuous attention switching, spatial indexes stored in VSTM, bindings stored in episodic buffer, and LTM representations are intimately involved in creating and updating identity-location bindings. Based on these principles, MOMIT correctly predicted the effects of target set-size, target velocity, target familiarity, as well as most of their interactions (for further details concerning the predicted and observed main effects and interactions, see [Appendix B](#)). In addition to the observed pattern of results being in good agreement with the functional and mathematical formulations of MOMIT, MOMIT has the advantage over former models of being able to account for both known effects reported earlier (set-size and speed effects) and also new effects (type effect). It is noteworthy that some theoretical alternatives to MOMIT, which rely on the early selection view of attention and parallel processing ([Pylyshyn, 1989](#); [Kahneman et al., 1992](#)), cannot account for the semantic effect, and they have problems with the speed effect and also to some extent with the effect of set-size in performance accuracy, as argued below.

In what follows, we first discuss the generalizability of the results. We then describe in more detail how MOMIT can accommodate the main findings of the study. We close by a discussion of the compatibility of the present data with the previous theories.

### *9.1. The generalizability of the observed effects*

Our observation that the tracking performance deteriorates as a function of target set-size is a very robust phenomenon, as a highly significant, target set-size effect was consistently observed in all experiments of the present study, regardless of the type of stimuli or the experimental task. Similar findings have also been obtained in other studies with different techniques and tracking tasks (in MOT by [Pylyshyn & Storm, 1988](#); [Yantis, 1992](#); [Liu et al., 2005](#); in MIT by [Oksama & Hyönä, 2004](#); in the multiple object permanence tracking paradigm by [Saiki, 2002](#)).

The influence of object speed on tracking performance is a much less studied phenomenon. Nevertheless, the study of [Saiki \(2002\)](#) supports our findings. Saiki found a similar speed effect with a variant of the change detection paradigm and with extremely high speeds (see below for more details). Both of these studies indicate that the mechanism responsible for tracking moving objects is susceptible to changes in object speed. Furthermore, the present study shows that the mechanism is surprisingly prone to even smallish changes in object speed. On the other hand, our study differs from that of Saiki in that we found a highly significant interaction between object velocity and target set-size whereas

Saiki did not. We strongly suspect that the interaction would have reached significance, had Saiki used more participants (with his 6 participants the interaction approached significance). Thus, we do not think that Saiki's failure to find a reliable interaction should be interpreted to contradict our significant result. Note also that a recent study of Liu et al. (2005) observed an effect of speed and an interaction between speed and set-size.

Our observation that familiar objects were easier to track than unfamiliar objects is a new finding in the research on multiple object tracking. The finding was consistently observed with two different types of stimuli (objects and faces), with two different measurement techniques (partial report probe recognition and change detection), and using both error rate and detection sensitivity ( $d'$ ) as dependent variables.

The semantic effect was modulated by target set-size in experiments, where the partial-report probe-recognition technique was employed (i.e., Experiments 1 and 2A; in Experiment 3A there was a trend). This interaction was located in the linear component; that is, the performance deteriorated steeper for pseudo-objects than familiar objects as a function of target set-size. In the experiments that used the change detection technique (i.e., Experiments 2B and 3B) the interaction was either not significant or it was qualitatively different from the interaction described above (i.e., the semantic effect in  $d'$  in Experiment 2A was apparent only for the two smallest set-sizes). We argue that the different pattern of results reflects a differential sensitivity of the two dependent measures. The partial report probe recognition technique is more sensitive at the higher end and less sensitive (due to a ceiling effect) at the lower end. On the other hand, an opposite pattern is apparent in the change detection paradigm. When set-size increases, change detection becomes increasingly difficult (observers become increasingly conservative in responding) and the technique thus loses its sensitivity. The mean size of the semantic effect (i.e., the difference in error rate between pseudo-objects and familiar objects) is plotted in Fig. 8 separately for the experiments using the PRPR task (the mean of Experiments 1, 2A, and 3A) and for the experiments where the change detection paradigm was used (the mean of Experiments 2B and 3B). As is apparent from Fig. 8, in the change detection experiments the effect size is bigger than in the PRPR experiments for smaller set-sizes (2–4), whereas the opposite is true for the larger set-sizes (5–6).

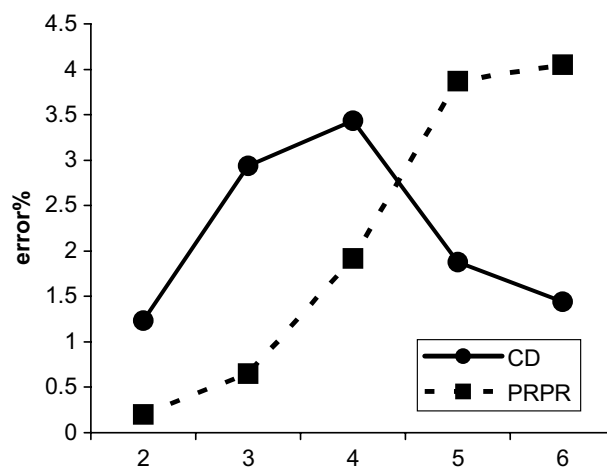


Fig. 8. The mean size of the semantic effect (the difference in error rate between pseudo-objects and familiar objects) is plotted separately for the experiments using the partial report probe recognition (PRPR) task (the mean of Experiments 1, 2A, and 3A) and for the experiments where the change detection (CD) paradigm was used (the mean of Experiments 2B and 3B).

This differential sensitivity between the two experimental tasks is also reflected in the results of model fitting. In the PRPR method the best fit with the data was obtained with a capacity estimate of approximately 4 static bindings while in the change detection technique the best fit was obtained using a capacity estimate of approximately 3 static bindings. We assume that the estimate taken from PRPR technique is closer to the right one and that the change detection procedure slightly underestimates this capacity. Following [Simons and Rensink \(2005\)](#) it is assumed that the change detection procedure includes a very demanding process where the post-change scene is compared to the memory-based representation. This comparison process may be difficult and/or slow; observers might not be able to compare all the materials in the STM buffer before it decays or they may become uncertain about their decision. Our finding of a conservative bias in the larger set-sizes supports this argument: observers became uncertain when a change had to be detected among four or more objects. On the other hand, probe recognition includes a forced choice procedure, which is known to activate material in a low-activation state. Thus, it is assumed to be a more sensitive measure of maximum capacity than the change detection procedure.

Our estimate of static binding capacity based on probe recognition is somewhat higher than those reported by [Horowitz et al. \(2007\)](#) for dynamic capacity. An obvious explanation is that the static binding capacity drops in dynamic situations as a function of the aforementioned cost factors. In their study, observers tracked 4 cartoon animals among 4 distracter animals. Horowitz et al. employed two probing methods: in one observers had to point to all targets that were occluded (whole report) after some period of time, in another they were asked to point to the location where a specific target (e.g., camel) was hiding (occluded behind a picture of a cactus). They found that the whole report yielded a dynamic capacity estimate of about 2.3–3.4, while the special report yielded an estimate of about 1.4–2.6. Similarly, [Saiki \(2003\)](#) obtained a smaller capacity estimate with a change detection procedure compared it to one obtained by a priming technique ([Kahneman et al., 1992](#)).

## 9.2. *The empirical support for the assumptions of MOMIT*

In Section 1 we laid out five principles that are the cornerstones of MOMIT. Next we discuss the compatibility of these assumptions with the present data.

- (1) *Efficient maintenance of multiple moving objects requires continuous serial (re)activation and refreshing of the dynamic identity–location bindings.* MOMIT as a serial account provides a natural explanation for why tracking performance deteriorates as a function of target set-size and speed, on the one hand, and why these effects combine in a multiplicative manner, on the other hand. Also our serial model based on the assumption of continuous refresh cycles yielded plausible estimates for refresh times and good fit with the observed accuracy data.
- (2) *There is a capacity limitation as to the number of bindings that can be simultaneously kept active in the episodic buffer, which affects performance accuracy.* Binding capacity was estimated by fitting the MOMIT model to the data. The best fitting estimate was approximately 4 static bindings for the partial-report probe recognition data, which is in line with previous research (e.g., [Luck & Vogel, 1997](#)). However, the estimated

capacity was approximately three items for the change detection data. As argued above, the difference is assumed to reflect a differential sensitivity of these two experimental procedures.

- (3) *Long-term memory (LTM) representations are utilized in creating temporary bindings.* The observed semantic effect (familiar objects are easier to track than unfamiliar objects) supports this claim. The effect was observed using both the PRPR and the change detection paradigms. MOMIT accounts for this effect in performance accuracy by assuming that the more time is spent in refreshing an identity-location binding, the more likely it is that another target is lost (the VSTM coordinate error becomes large before the binding is refreshed). According to MOMIT, the size of the semantic effect should increase as a function of set-size. As discussed above, the PRPR method provides data consistent with this prediction, whereas the change detection paradigm yields data that demonstrate a somewhat different pattern (i.e., the semantic effect is greater for small set sizes; see above for more).
- (4) *Spatial indexes (or location pointers) for the tracked targets are temporarily stored in VSTM. These indexes are then utilized by the mechanism that programs shifts of visual attention between targets.* As identity-location binding is assumed to be serial in nature, the spatial indexes provided by VSTM are not constantly updated, which leads to the VSTM coordinate error (i.e., the difference between the stored and the actual coordinates). The magnitude of this VSTM error is one of the key concepts underlying MOMIT's success in correctly predicting the effects of speed, set-size, and object type. The size of the VSTM error increases with an increase in target speed and set-size, which results in non-efficient switching of attention between targets. Similarly, the VSTM error becomes larger when the observer is required to spend extra time in refreshing an identity-location binding for unfamiliar objects.
- (5) *The system responsible for shifting of attention during tracking also obtains location information of moving objects in parallel. This information is provided by the peripheral vision.* Although MOMIT assumes that identity-location bindings are maintained by constantly updating and refreshing them with the help of serial attention, it does not completely preclude a parallel component in the form of peripheral vision. This kind of mechanism is needed, because otherwise we would need to postulate that the attention is directed purely endogenously on the basis of VSTM information, which we think is highly unlikely. However, the present study does not provide any direct evidence in favor or against this assumption. One possibility to test this is to track observers' gaze shifts when tracking targets among multiple distracters. MOMIT predicts erroneous attention shifts to moving objects (either target or distracter; however, see Pylyshyn, 2004) that occupy a position previously occupied by a target that has not been refreshed recently.

In sum, the observed results are largely consistent with the predictions derived from the principles of MOMIT. In the next sections, we compare MOMIT with other models of multiple object tracking and with general models of visual attention.

### 9.3. Comparison of MOMIT with parallel fixed capacity theories of dynamic visual attention

Strong and viable competitors of MOMIT are models of visual attention that are based on fixed- or limited-capacity parallel processing (unlimited capacity parallel models are

not psychologically viable and also clearly inconsistent with the present set of data). Fixed-capacity parallel models have recently become popular both as general models of visual attention (Bundesen, 1990; Bundesen, Habekost, & Kyllingsbaek, 2005; Logan, 2002b) and specifically in the area of dynamic visual attention (Cavanagh & Alvarez, 2005; Kahneman et al., 1992; Pylyshyn & Storm, 1988). We first discuss these specific models followed by a discussion of Bundesen's model, Rensink's (2000) coherence theory, and a recent mathematical model of Kazanovich and Borisyuk (2006).

The fixed-capacity parallel models assume that multiple visual objects can be selected and spatially tracked in parallel. This is done either attentively (Cavanagh & Alvarez, 2005; Kahneman et al., 1992), preattentively (Pylyshyn & Storm, 1988), or as a joint effort of early visual and attentional processes (Bundesen, 1990; Bundesen et al., 2005; Logan, 2002b). The size of the fixed capacity is most explicitly defined by Pylyshyn and Storm (1988), who posit 4–5 “fingers of instantiation” (visual indexes that move along with the moving objects, as if these pointers were glued to the tracked objects). The theories also differ in how access is achieved to identity information of the tracked targets. According to Pylyshyn's (1989) FINST theory, even though tracking itself is parallel (within the capacity limits), accessing identity information is done one object at a time with the help of serial attention (in that sense his model is a mixed model involving both a parallel and serial component). On the other hand, the object file theory of Kahneman and Treisman (1984; Kahneman et al. 1992) assumes that identity information is transmitted of and available for all visual objects selected as targets (again, within capacity limits). In other words, it is possible to temporarily retain multiple object files constructed for the target objects, each of which contains an identity-location binding. Finally, the multifocal model of Cavanagh and Alvarez (2005) posits multiple attentional foci (four; two of which track objects located in the left visual field and another two track objects in the right visual field; see also Alvarez & Cavanagh, 2005) that make possible to simultaneously track multiple moving objects. However, identity information and identity-location bindings are not necessarily transmitted automatically of the tracked targets. Thus, this model does not seem to assume a parallel existence of multiple bindings. Note that regardless of their stance on the binding issue, all these models predict a flawless tracking performance within the capacity limits. This is because location information is simultaneously available for all tracked targets. For example, according to the FINST theory (Pylyshyn, 1989), serial attentional switching required to extract identity information can be readily done between the indexed locations without the attentional focus being misdirected (e.g., to a distracter). Unlike MOMIT, the models described above do not specify underlying mechanisms that may lead to failure in performance (except for Pylyshyn et al., 1994, who suggest that the preattentive filter is vulnerable to “leaking”). It is thus understandable that the predictions of these alternative models do not differ from each other when it comes to performance accuracy (but they do differ with respect to response time).

On the basis of the present results, we cannot rule out fixed capacity parallel models as a viable alternative. As demonstrated by Logan (2002a) and Townsend and Wenger (2004), it is in principle possible to simulate the same data pattern using either a serial or a parallel model (at least in response times). Although none of the parallel models discussed above predict a priori the observed pattern of results or postulate a plausible mechanism, it is nevertheless possible to post-hoc postulate a parallel model with a tracking capacity of three items (in all experimental conditions tracking performance for three items was at the 90% level or better) that is also sensitive to object speed and object familiarity. In order

to be a viable alternative to MOMIT, such a model should also be able to address the following issues:

- (1) Why does performance accuracy deteriorate as a function of object velocity? An effect of object speed seems at odds with at least with the FINST theory (Pylyshyn, 1989), which assumes that preattentive pointers effectively move along with the moving targets, at least when object velocity is within reasonable limits. Our object speed manipulation was quite modest (certainly much smaller than in many previous studies), nevertheless we obtained a robust effect of speed. On the other hand, a model based on parallel allocation of attention to multiple targets (Cavanagh & Alvarez, 2005) may be able to account for speed effects (at least in response times) as a difficulty effect in allocating attention to fast moving objects. Another way of construing a speed effect is in terms of exogenous uncertainty (Moray, 1984): exogenous uncertainty of visual input becomes greater as object speed increases (i.e., it becomes more difficult to predict a future state of a dynamic visual environment on the basis of the current state). According to this view, exogenous uncertainty affects performance regardless of whether targets are attended serially or parallelly (but see issue 3 below).
- (2) Why does object familiarity affect tracking accuracy? The observed familiarity effect seems clearly at odds with those parallel accounts of dynamic visual attention, which rely on early selection view of attention (Pylyshyn, 1989; Kahneman et al., 1992). On the other hand, a version of parallel attentional models in which target properties influence processing speed may be able to account for at least changes in response times (see below our discussion of Bundesen, 1990; Bundesen et al., 2005).
- (3) Why do the main effects of set-size and speed combine in a multiplicative manner resulting in significant linear interactions? Our serial account predicts that as processing becomes more difficult (due to increase in set-size and speed) performance accuracy will indeed decline in a multiplicative manner. This is because the time needed to locate fast moving targets and refresh their bindings will increase cumulatively as a function of set-size, as every item is visited in a serial fashion. Parallel accounts predict no such interactions (within capacity limits; e.g., Pashler, 1995). At most, they predict an additive or constant increase in reaction times and/or performance errors as a function of speed (due to a general increase in exogenous uncertainty), regardless of set-size. If that were the case, then this additive effect should also be evident in the smaller set-sizes. Clearly, there is no evidence for a speed effect in set-size 2 (performance accuracy is highly similar for all speed conditions, see Table 1).
- (4) How is it possible to track parallelly visual objects and their characteristics in real-world environments where the to-be-tracked targets are often located wide apart from each other (targets may appear tens of degrees of visual angle from each other, e.g., when a lifesaver tracks swimmers)? Given the poor visual acuity of the peripheral vision it appears implausible that visual features and identities of multiple objects may be refreshed and maintained in parallel.

In sum, although we cannot rule out parallel accounts, we have demonstrated that MOMIT offers a coherent serial account of the observed pattern of results. It also offers psychologically plausible mechanisms underlying the variation in performance accuracy

in a variety of different visual and stimulus environments. We challenge parallel theorists to provide a detailed description of a parallel mechanism, which is capable of producing the kind of set-size, speed, familiarity effects and their interactions observed in the present study.

#### *9.4. Comparison of MOMIT with Bundesen's (1990; Bundesen et al., 2005) computational theory of visual attention*

On a general level, the results of the present study compare favorably with Bundesen's (1990; Bundesen et al., 2005) computational theory of visual attention. His theory is a fixed-capacity parallel model that is designed to simulate processing speed, while MOMIT is geared to simulate performance accuracy (although processing speed also features in the equation). Even though his theory is not tailored to model dynamic visual attention, and hence its architecture significantly differs from that of MOMIT, one can also find clear commonalities between them.

In Bundesen's (1990; Bundesen et al., 2005) theory, the process of categorizing a visual object is determined by three factors: sensory evidence related to the object, behavioural importance of the object, and decision bias associated with the relevant feature category. Sensory evidence and importance constrain what objects are selected as attentional targets, while decision bias regulates the categorization process. The objects present in a visual environment compete with each other to be categorized; the winner obtains access to the limited capacity VSTM. The aforementioned three factors determine what objects and categories win the race.

As attention is equipped with limited capacity, the theory predicts that the speed of object categorization slows down as the number of selected objects increases. This is because with multiple targets the fixed capacity needs to be distributed among several behaviorally pertinent objects. Applied to a dynamic visual environment, this would mean that updating identity information for objects retained in VSTM is slowed down. This becomes close to the mechanism posited in the MOMIT architecture (note, however, that in Bundesen's theory the mechanism is assumed to be parallel whereas in MOMIT it is serial in nature).

Bunden's (1990; Bundesen et al. 2005) theory can easily accommodate the familiarity effect by making recourse to decision bias. It may be assumed that familiar objects have a stronger decision bias than unfamiliar objects (familiar objects have a LTM representation that may be utilized in categorization). Thus, categorization speed is predicted to be faster for familiar than unfamiliar objects. This prediction is similar to the one derived from MOMIT, which assumes that creating and updating bindings for familiar objects is more readily achieved than for unfamiliar objects.

The speed effect appears to be more difficult to explain by Bundesen's (1990; Bundesen et al. 2005) model. One possibility is to suggest that the sensory evidence extracted from targets deteriorates as object speed increases, which in turn would slow down object categorization and their coding to VSTM. However, no explicit mechanism is offered in the model to account for object speed effects (the issue is merely rephrased in terms of the quality of the available sensory evidence).

In sum, although clear parallels can be found between MOMIT and the Bundesen (1990; Bundesen et al. 2005) model, we think that Bundesen's theory suffers from a plausibility problem when it is applied to tracking moving objects in real-life type of visual

environments where the objects are located wide apart from each other. As we have argued throughout the paper, in such environments it seems highly unlikely that a purely parallel system could handle the simultaneous refreshing and maintenance of multiple bindings (see Logan, 2002a; Moray, 1984).

#### 9.5. Comparison of MOMIT with the coherence theory of Rensink (2000)

The observed semantic, speed, and set-size effects may also be considered consistent with the coherence theory of Rensink (2000). The theory posits that a ‘stabilization’ process is needed when attempting to maintain a coherent representation of a visual scene. According to the coherence theory, only one object at a time can be represented so that it has a high degree of coherence over space and time. Other objects or their parts in a scene, so-called *proto-objects*, are volatile and non-stable. A coherent representation of one object, the nonvolatile higher level representational structure called *nexus*, makes recourse to focused attention that selects a small number of proto-objects and stabilizes them. This stabilization process makes use of a feedback circuit (*coherence field*) with bi-directional links between the attended proto-objects and the higher-level nexus. The upward links in the circuit transmit property information (featural and semantic) about the attended proto-objects; the downward links from the nexus provide coherence and stability over time to the attended proto-objects.

By using these assumptions of the coherence theory it seems possible to derive a plausible mechanism by which LTM representations could affect dynamic perception. It may be argued that familiar objects bring about more coherent higher-level representations than pseudo-objects. Moreover, the more coherent the representations are, the more stability they provide to lower-level proto-objects. Thus, maintaining identity-location bindings for familiar objects would be more effective in a dynamic situation than that for unfamiliar objects. Although such an interpretation seems to fall from his theory, it may not be what Rensink (2000) himself has intended, as he writes (p. 26) that the coherence theory “views attention as concerned with the formation of immediate spatiotemporal structures (or tokens) rather than the activation of long-term categories (or types)”.

The object speed effect seems consistent with Rensink’s (2000) coherence theory, according to which movement may hamper the stabilization of the output provided by lower level vision and hence the perception of coherent objects (see Rensink, 2002, for further discussion of speed effects). Furthermore, unlike the FINST (Pylyshyn, 1989) and the object file theory (Kahneman et al., 1992), the coherence theory does not postulate any parallel access mechanism that would provide access to multiple objects in a scene. Rensink claims that only one stable representation can be formed at a time and that the observer’s simultaneous perception of other objects in a visual scene is in fact an illusion. Similarly to MOMIT, other objects in the scene are perceived by rapidly switching attention back and forth between the objects.

#### 9.6. Comparison of MOMIT with the oscillatory neural model of Kazanovich and Borisyuk (2006)

Some of the observed effects are consistent with Kazanovich and Borisyuk’s (2006) neural model of multiple object tracking. This parallel model specifies a mechanism to

explain errors in MOT performance. Their attentional model is based on an oscillatory neural network, where local oscillatory elements are controlled by a central oscillator (or several central oscillators). The central oscillator plays the role of a central executive, which controls the dynamics of local, peripheral oscillators through feedforward and feedback connections. Thus, the model includes several peripheral components capable of tracking objects in parallel and a central component(s) capable of maintaining the attentional focus. An important property of their network is that its architecture contains only a small number of connections in comparison to connectionist models, which makes it technically feasible to simulate the tracking of moving objects.

According to the model, tracking of multiple identical objects (e.g., circles) performance is error prone due to crossing of the trajectories of targets and distracters. In such occasions the oscillator switch its focus from the target to the distracter, which results in a tracking error. When object speed increases, such switches are assumed to take place with a greater likelihood. The model's performance was compared to the tracking data of Oksama and Hyönä (2004) of identical moving objects (the classic MOT task). The simulation yielded a similar set-size effect in error rates as the data. Kazanovich and Borisyuk suggest that the model is also geared to simulate speed effects (not yet done). Although the model is parallel, it predicts that error rate will increase as a function of set-size and speed. The model is a serious competitor to MOMIT. However, at present it still has problems with the trial duration effect observed by Oksama and Hyönä (2004) with identical objects. Moreover, the semantic effect observed in the present study poses also problems to the model.

## 10. Conclusions

In the present article, we have proposed a functional and formal model for the maintenance of multiple bindings in a dynamic visual environment. The proposed serial MOMIT model can account for most of the results reported in the present study; in addition, it is also consistent with the results of Oksama and Hyönä (2004). It is sufficiently detailed to generate empirically testable predictions (some of which are mentioned above). It provides a theoretical alternative to parallel models popular in the area of dynamic visual attention. In the future, some of the model's predictions will be further tested by eye movement studies of multiple object tracking.

## Acknowledgment

The completion of this study was made possible by a grant provided by Finnish Scientific Advisory Board of Defence to the first author. The second author acknowledges the support of Suomen Akatemia (the Academy of Finland). We are grateful to Gordon Logan and three anonymous referees for their highly useful comments on a previous version of this article. We also thank Juhani Sinivuo, Marja-Leena Haavisto and Krista Oinonen, who supported the work in many ways, and Maija Seppänen and Jari Lipponen for their help in collecting the data. Finally, we are grateful to Jari Laarni and Virpi Kalakoski for their valuable comments during the project and to Sirkka Oksama for her mathematical

expertise. Portions of the data were presented at the 13th Conference of the European Society of Cognitive Psychology, Granada, Spain, September 17–20, 2003, and at the 29 European Conference on Visual Perception, Saint Petersburg (Russia), August 20–25, 2006.

### Appendix A. Mathematically derived predictions of MOMIT

In the following, the empirical predictions of MOMIT are mathematically derived on the basis of Formula 6. We first consider the prediction concerning the effect of target set-size  $n$ . By replacing all other factors with a constant (constants are different combinations of the terms appearing in Formula 6), the formula may be expressed as follows:

$$C1 * n^2 + C2 * n + C3 + \frac{C4}{n},$$

where  $C1$ ,  $C2$ ,  $C3$ , and  $C4$  are constants ( $C1$  obtains a negative value, see the model simulations). Thus, according to MOMIT, deterioration in performance accuracy as a function target set-size is depicted as a quadratic downward trend.

The predictions for the main effect of target speed and familiarity can be derived in a similar fashion. When all other factors are replaced with a constant, the effect of object speed  $v$  on performance accuracy is estimated as follows:

$$C1 * v^2 + C2 * v + C3$$

and the effect of object familiarity  $s$  as follows:

$$C4 * s^2 + C5 * s + C6,$$

where  $C1$ ,  $C2$ ,  $C3$ ,  $C4$ ,  $C5$ , and  $C6$  are all constants. As may be seen from these formulas, MOMIT predicts that accuracy drops quadratically as a function of object speed and familiarity (in order to thoroughly test these predictions one would need to vary speed and familiarity using several speed and familiarity levels, not just 2 or 3 as in the present study).

We next consider the model prediction with respect to a set-size  $\times$  speed interaction. Let's consider the difference between two speed conditions  $v$  and  $v + k$  ( $v$  and  $k$  obtain positive values, thus  $v + k > v$ ). Again, all other terms are replaced with a constant. We can formulate the difference between two speed conditions as a function of  $n$  as follows:

$$C1 * n^2 + C2 * n + C3,$$

where  $C1$ ,  $C2$ , and  $C3$  are constants. Thus, the model predicts an interaction between set-size and speed, where the difference between two speed conditions vary as a function of set-size following a quadratic trend. Note that this does mean that MOMIT would predict a significant quadratic component in the interaction term, but the interaction term is assumed to be linear. A quadratic interaction component denotes that the curves would have different convexity values. However, MOMIT assumes a single value (see Formula 6).

Predictions for set-size  $\times$  familiarity and speed  $\times$  familiarity interactions may be derived in a similar vein. Let's consider the difference in accuracy between two types of objects differing in familiarity (the refresh times are denoted by  $s$  and  $s + r$ ; both obtain positive values, thus  $s + r > s$ ), as a function of speed or set-size. All other terms are considered

constant. The difference in two familiarity conditions is predicted to vary quadratically as a function of set-size  $n$ :

$$C1 * n^2 + C2 * n + C3$$

and as a function of speed  $v$

$$C4 * v^2 + C5 * v + C6$$

Finally, one way to construe the 3-way interaction between set-size, speed, and familiarity is to examine whether the set-size  $\times$  familiarity interaction is different for different object speed conditions  $v$  and  $v + h$ . Let's denote the difference between two object types as a function of set-size with speed  $v$  as  $C1*n^2 + C2*n + C3$  (as above) and call it Difference 1. Let's then call Difference 2 the set-size  $\times$  familiarity difference with speed  $v + h$ . We can then examine whether the difference ( $D$ ) between Difference 1 and 2 is a constant (no 3-way interaction) or increases as a function of speed increase  $h$ . By solving the equation we find that indeed  $D$  varies a function of speed increase  $h$  ( $C1*h^2 + C2*h$ ). In other words, MOMIT predicts a 3-way interaction where the increase in the familiarity effect as a function of set-size is exacerbated by an increase in object speed.

### Appendix B. Modeling the observed main effects and interactions of Experiment 1

In the following we go over the predicted and observed effects for Experiment 1 that included all the manipulated variables, set-size, object speed, and object type. As mentioned earlier, MOMIT predicts all possible main effects and interactions between the manipulated variables (see Appendix A for the mathematical proof). As a further test

Table B.1

Error rates (%) predicted by MOMIT and the differences between the observed (Experiment 1) and predicted error rates (observed–predicted)

Object speed	Type of object	Number of targets						Marginal means	
		4		5		6		Predicted	Observed
		Predicted	Observed	Predicted	Observed	Predicted	Observed		
		– predicted		– predicted		– predicted		– predicted	
Slow	Pseudo	3.0	3.1	13.8	–0.4	21.3	–1.8	12.7	0.3
	Familiar	2.9	0.5	13.7	–3.9	21.0	–4.9	12.5	–2.8
	Semantic effect	0.1	2.6	0.2	3.6	0.3	3.1	0.2	3.1
Medium	Pseudo	6.2	2.0	19.2	1.4	29.2	–0.6	18.2	1.0
	Familiar	5.6	1.4	18.1	–3.2	27.6	–1.4	17.1	–1.1
	Semantic effect	0.6	0.6	1.1	4.7	1.6	0.8	1.1	2.0
Fast	Pseudo	14.2	5.4	32.2	1.9	48.3	–4.0	31.6	1.1
	Familiar	12.3	6.0	29.1	3.8	43.7	–3.8	28.4	2.0
	Semantic effect	1.9	–0.5	3.2	–1.9	4.6	–0.1	3.2	–0.9
	Marginal means	7.4	3.1	21.0	–0.1	31.9	–2.8	20.1	0.1

of the model, we substituted the parameters  $s$  and  $m$  in Eq. (6) with their best estimates ( $s = 230$  ms for the familiar objects;  $s = 250$  ms for the pseudo-objects;  $m = 3.8$ ) and compared the predicted condition means of each main effect and interaction to the observed ones. Table B.1 shows the predicted means and their differences from the observed ones. As in the model fitting section, only set-sizes 4–6 were used in the comparison (see that section for the motivation).

As is evident from Table B.1, the overall fit of MOMIT is very good: the overall mean difference in error rates between the predicted and observed values is only 0.1%. There is also a very close fit between the predicted and observed condition means for all main effects and the set-size  $\times$  speed and set-size  $\times$  type interactions (i.e., the patterns of the predicted effects and interactions closely resemble the observed ones). It is also noticeable that both the predicted and observed effect sizes related to set-size and speed are robust. Performance accuracy drops from a practically flawless performance with 2 objects (see Table 1) down to about 70% accuracy with 6 objects (see Table B.1). An effect size of similar magnitude is also predicted and observed for object speed. Analogously, the predicted and observed set-size  $\times$  speed interaction is robust: performance accuracy drops off dramatically when 6 fast moving objects are tracked. The robustness of the observed effects is also apparent in the power estimates, which were .98–1 for the aforementioned effects.

On the other hand, the predicted and observed effect of object type is relatively modest. The observed semantic effect for set-sizes 4–6 was 2.9%, while MOMIT predicts a 1.5% effect. Given the small effect size, it may not come as a surprise that two out of three interactions involving object type as a factor did not reach significance in Experiment 1. As regards the type  $\times$  speed interaction, MOMIT predicts that the effect of object type becomes more pronounced when speed is increased (the predicted error rates are 0.2, 1.1, and 3.2 for the slow, medium, and fast speed, respectively). Interestingly, this means that MOMIT predicts virtually no object type effect with slowly moving (or static) stimuli. The size of the observed type effect was somewhat bigger in the slow (3.3%) and medium speed (3.1%) condition than predicted by MOMIT, but they were quite similar in the fast condition (observed: 2.4% vs. predicted: 3.2%). As is apparent from Table B.1, the lack of interaction is primarily due to familiar objects being tracked more accurately in the slow and medium speed conditions than predicted by MOMIT. On the other hand, MOMIT predicts almost flawlessly the condition means of the pseudo-objects. Thus, MOMIT slightly underestimates the semantic effect in the slow speed condition.

As regards the 3-way interaction, MOMIT predicts no set-size  $\times$  type interaction with a slow speed (see Table B.1 for the semantic effect in the slow speed condition). On the other hand, an interaction is predicted for the medium speed condition, and it is predicted to be even more pronounced with a fast speed. Separate ANOVAs computed for each speed condition (set-sizes 2–6) partly confirmed these predictions. The set-size  $\times$  type interaction was not significant in the slow speed condition ( $F(4, 216) = 2.1, p = .082$ ), but it proved significant in the medium speed condition ( $F(4, 216) = 4.6, p = .001$ ). However, contrary to MOMIT's prediction, the interaction was non-significant in the high speed condition ( $F(4, 216) = 1.6, p = .175$ ). Pairwise comparisons of the predicted and observed condition means show that in the high speed condition the results diverged slightly for only one data point: tracking of 5 familiar objects was observed to be 1.9% worse than predicted. For set-sizes 4 and 6 in the high speed condition the size of the observed and predicted type effect was quite comparable, the differences were 0.5% and 0.1%, respectively. In sum, it seems that the failure to find a 3-way interaction is mainly due to lack of power (the power estimate was .44).

All in all, MOMIT provides an excellent overall fit to the main effects and interactions. The only exception is that the observed semantic effect is greater with a slow speed than predicted by MOMIT. However, the model and the data are in agreement with the generally modest size of the semantic effect.

## References

- Allen, R., McGeorge, D., Pearson, D., & Milne, A. (2004). Attention and expertise in multiple target tracking. *Applied Cognitive Psychology, 18*, 337–347.
- Alvarez, G. A., & Cavanagh, P. (2005). Independent resources for attentional tracking in the left and right visual hemifields. *Psychological Science, 16*, 637–643.
- Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford University Press.
- Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417–423.
- Bahrami, B. (2003). Object property encoding and change blindness in multiple object tracking. *Visual Cognition, 10*, 949–963.
- Bundesen, C. (1990). A theory of visual attention. *Psychological Review, 97*, 523–547.
- Bundesen, C., Habekost, T., & Kyllingsbaek, S. (2005). A neural theory of visual attention: Bridging cognition and neurophysiology. *Psychological Review, 112*, 291–328.
- Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences, 9*, 349–354.
- Chun, M. M. (1997). Types and tokens in visual processing: A double dissociation between the attentional blink and repetition blindness. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 738–755.
- Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 109–127.
- Corel Mega Gallery (1996). Corel Corporation.
- Cowan, N. (1995). *Attention and memory: An integrated framework*. New York: Oxford University Press.
- De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research, 52*, 317–329.
- Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*, 1827–1837.
- Duncan, J. N., & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review, 96*, 422–458.
- Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford: Oxford University Press.
- Germeys, F., De Graef, P., & Verfaillie, K. (2002). Transsaccadic perception of saccade target and flanker objects. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 868–883.
- Gordon, R. D., & Irwin, D. E. (1996). What's in an object file? Evidence from priming studies. *Perception & Psychophysics, 58*, 1260–1277.
- Henderson, J. M. (1992). Visual attention and eye movement control in reading and picture viewing. In K. Rayner (Ed.), *Eye movements and visual cognition* (pp. 260–283). Berlin: Springer.
- Henderson, J. M. (1994). Two representational systems in dynamic visual identification. *Journal of Experimental Psychology: General, 123*, 410–426.
- Henderson, J. M., & Anes, M. D. (1994). Roles of object-file review and type priming in visual identification within and across eye fixations. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 826–839.
- Henderson, J. M., Pollatsek, A., & Rayner, K. (1989). Covert visual attention and extrafoveal information use during object identification. *Perception & Psychophysics, 45*, 196–208.
- Horowitz, T. S., Holcombe, A. O., Wolfe, J. M., Arsenio, H. C., & DiMase, J. S. (2004). Attention pursuit is faster than attentional saccade. *Journal of Vision, 4*, 585–603.
- Horowitz, T. S., Klieger, S. B., Fencsik, D. E., Yang, K. K., Alvarez, G.A., & Wolfe, J.M. (2007). Tracking unique objects. *Perception & Psychophysics*, in press.
- 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.
- Kanwisher, N. (1991). Repetition blindness and illusory conjunctions: Errors in binding visual types with visual tokens. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 404–421.
- Kazanovich, Y. B., & Borisyuk, R. M. (2006). An oscillatory neural model of multiple object tracking. *Neural Computation*, *18*, 1413–1440.
- Kroll, J. F., & Potter, M. C. (1984). Recognizing words, pictures, and concepts: A comparison of lexical, object and reality decisions. *Journal of Verbal Learning and Verbal Behavior*, *23*, 39–66.
- Landry, S. J., Sheridan, T. B., & Yufik, Y. M. (2001). A methodology for studying cognitive groupings in a target-tracking task. *IEEE Transactions on Intelligent Transportation Systems*, *2*, 92–100.
- Liu, G., Austen, E., Booth, K., Fisher, B., Argue, R., Rempel, M., et al. (2005). Multiple-object tracking is based on scene, not retinal, coordinates. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 235–247.
- Logan, D. G. (2002a). Parallel and serial processing. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology. methodology in experimental psychology* (Vol. 4, pp. 271–300). NJ: John Wiley & Sons.
- Logan, D. G. (2002b). An instance theory of attention and memory. *Psychological Review*, *109*, 376–400.
- Logan, D. G., & Zbrodoff, J. N. (1999). Selection for cognition: Cognitive constraints on visual spatial attention. *Visual Cognition*, *6*, 55–81.
- Logie, R. H. (1995). *Visuo-spatial working memory*. Hove, UK: Psychology Press.
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.
- 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.
- Pashler, H. (1995). Attention and visual perception: Analyzing divided attention. In S. Kosslyn & D. Osherson (Eds.), *Visual cognition: An invitation to cognitive science* (Vol. 2, pp. 71–100). Cambridge, MA: MIT Press.
- 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. (1994). Some primitive mechanisms of spatial attention. *Cognition*, *50*, 363–384.
- 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., Burkell, J., Fisher, B., Sears, C., Schmidt, W., & Trick, L. (1994). Multiple parallel access in visual attention. *Canadian Journal of Experimental Psychology*, *48*, 260–283.
- 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.
- Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, *7*, 17–42.
- Rensink, R. A. (2002). Changes. In J. Hyönä, D. P. Munoz, W. Heide, & R. Radach (Eds.), *The brain's eye: Neurobiological and clinical aspects of oculomotor research* (pp. 133–148). Amsterdam: Elsevier Science.
- Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*, 368–373.
- Ruchkin, D. S., Grafman, J., Cameron, K., & Berndt, R. S. (2003). Working memory retention systems: A state of activated long-term memory. *Behavioral and Brain Sciences*, *26*, 709–777.
- Saiki, J. (2002). Multiple-object permanence tracking: Limitation in maintenance and transformation of perceptual objects. In J. Hyönä, D. P. Munoz, W. Heide, & R. Radach (Eds.), *The brain's eye: Neurobiological and clinical aspects of oculomotor research* (pp. 133–148). Elsevier Science: Amsterdam.
- Saiki, J. (2003). Feature binding in object-file representations of multiple moving items. *Journal of Vision*, *3*, 6.21.
- 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., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition*, *80*, 159–177.

- 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.
- Shapiro, K. L., Raymond, J. E., & Arnell, K. M. (1994). Attention to visual pattern information produces the attentional blink in RSVP. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 357–371.
- Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, 5, 644–649.
- Simons, D., & Rensink, R. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, 9, 16–20.
- 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.
- Townsend, J. T., & Wenger, M. J. (2004). The serial-parallel dilemma: A case study in a linkage of theory and method. *Psychonomic Bulletin & Review*, 11, 391–418.
- Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*, 6, 171–178.
- Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97–136.
- Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14, 107–141.
- Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.
- Wheeler, M. E., & Treisman, A. M. (2002). Binding is short-term visual memory. *Journal of Experimental Psychology: General*, 131, 48–64.
- Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology*, 24, 295–340.