

# **Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach**

Lauri Oksama

*Finnish Defence Forces Education Development Centre, Järvenpää, Finland*

Jukka Hyönä

*University of Turku, Finland*

Existing theories of multiple object tracking (MOT) offer different predictions concerning the role of higher level cognitive processes, individual differences, effortful attention and parallel processing in MOT. Pylyshyn's model (1989) argues for an automatic parallel processing mechanism separate from other cognition, whereas alternative models (e.g., Kahneman & Treisman, 1984 or spotlight models) are based on higher level cognition such as spatial short-term memory and/or effortful attention switching. These predictions were examined in Experiment 1 where identical objects and in Experiment 2 where visually and semantically distinct objects were tracked. Both experiments demonstrated a substantial individual variation in the estimated tracking capacity. Tasks measuring visuospatial short-term memory and attention switching proved to be significant predictors of MOT. In addition, tracking performance deteriorated as a function of tracking time and set size. Our results are in contrast to Pylyshyn's model. A mechanism with both parallel and serial processing and temporary spatial memory is outlined to accommodate the observed pattern of results.

---

Please address all correspondence to Lauri Oksama, Finnish Defence Forces, Education Development Centre, Behavioral Sciences Division, PO Box 5, FIN-04401 Järvenpää, Finland. Email: lauri.oksama@pvkk.inet.fi

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 thank Marja-Leena Haavisto, Krista Oinonen and Juhani Lehto, who participated in the planning of the test battery, especially in constructing the visuospatial complex span task. We also thank Krista Oinonen and Maija Seppänen for collecting the data. Juhani Sinivuo supported the work in many ways. Finally, we are grateful to Glyn Humphreys and two anonymous reviewers for their constructive and thoughtful criticism on an earlier version of this paper, and to Jari Laarni and Mika Koivisto for their valuable comments.

Visual tracking of moving targets is an essential part of many real world tasks. For example, air traffic controllers and military pilots using up-to-date technology need to keep track and integrate different kinds of visual information. In many sports, such as football and ice hockey, a player has to keep track of other players in a constantly changing situation. But how good is our visuoperceptual system at doing this kind of tracking? Can we genuinely track several moving targets simultaneously and effortlessly, or do we need to shift effortfully our focal attention from one target to another during tracking? Are some people better than others in tracking tasks like these? Several interesting questions about multiple object tracking (MOT) can be raised from both a practical and theoretical point of view. They concern the capacity of simultaneous tracking, the automatic versus the effortful nature of tracking, and the role of individual differences in tracking. These questions were examined in the present study.

Theoretically, perhaps the most relevant question concerns the psychological mechanisms behind tracking performance. Is the mechanism based on an independent module separate from other cognitive processes with its own operation and memory systems? Or does it make use of processes also shared by other visuospatial tasks? Four plausible models of MOT are first described and discussed and we outline their predictions related to these questions (tracking capacity, automatic versus effortful nature of MOT, and the nature of individual differences). The models differ from each other in many important respects. Pylyshyn's FINST (fingers of instantiation) theory (e.g., Pylyshyn, 1989, 1994; Pylyshyn, Burkell, Fisher, Sears, Schmidt, & Trick, 1994; Pylyshyn & Storm, 1988) capitalizes on low-level early vision processes, where tracking is carried out preattentively without recourse to memory representations. In contrast, Yantis' model (1992) and that of Kahneman and Treisman (1984; Kahneman, Treisman, & Gibbs, 1992) are based on higher level attentive processes, object representations, and spatial relations between the tracked objects. The fourth model to be discussed, an attention switching model, assumes that tracking is carried out by shifting a "spotlight" of attention continuously from one target to another (cf. Posner, 1980).

In cognitive psychology, the research on dynamic visual attention started only about a decade ago. The first multiple object tracking experiment was designed by Pylyshyn (Pylyshyn & Storm, 1988; for a new development of the paradigm, see Saiki, 2002). Before that, visual attention was mainly studied in the context of static stimuli. In the tracking task, the participant has to track randomly moving targets among several moving distracters. The seminal study of Pylyshyn and Storm indicated that the participants were able to track between four and five targets with an 85% accuracy.

## PYLYSHYN'S EARLY VISION MODEL

The original FINST theory contends that the performance in MOT (and in other visual tasks involving multiple objects) is based on a very primitive mechanism provided by an “early vision” system (recently Pylyshyn, 2001, has added new assumptions and slightly changed the terminology; however, it seems that the basic claims of the original model remain the same). The assumed mechanism provides four or five indexes or pointers, which can be assigned to visual objects. These pointers or reference tokens do not encode or represent anything about the objects they refer to (i.e., in that sense they are “feature-blind”); they just make a primitive mechanical reference relation possible between the visual system and the outside world, similarly to physical fingers touching physical objects. Thus, the FINST mechanism is separate from higher cognition; it is an encapsulated, “cognitively impenetrable” system with perhaps its own memory system (cf. Pylyshyn, 1999). Higher order serial cognitive processes can access the output of this mechanism or select input to this system, but they cannot influence the internal workings of the encapsulated system. Indexes work independently from each other and make it genuinely possible to track several objects simultaneously. The mechanism has limited capacity, about four indexes, due to the architectural limitations in the visual nervous system (its localization is not known; the posterior parietal cortex is one potential candidate; cf. Culham, Brandt, Cavanagh, Kanwisher, Dale, & Tootell, 1998; Scholl & Pylyshyn, 1999).

After indexes have been assigned to the to-be-tracked objects they continue to remain with the same objects despite the fact that the moving objects change their locations continuously on the retina and despite any temporary occlusions (i.e., indexes are “sticky”). Originally, this notion of “stickiness” was introduced to suggest that index maintenance is automatically carried out by the FINST mechanism without a need of allocating attentional effort (Pylyshyn & Storm, 1988). Later, however, Pylyshyn (Pylyshyn et al., 1994; Pylyshyn, 2001) has admitted that the MOT task as a whole is attentionally demanding and that successful index maintenance during the tracking task may perhaps require periodic attentional effort to reactivate index binding (but he still continues to assume that the mechanism is in principal automatic; see Pylyshyn, 2001, p. 149). Yet, to our knowledge, the question about the nature of tracking (whether automatic versus effortful) has not been systematically studied in the context of the Pylyshyn paradigm.

Pylyshyn and Storm (1988) introduced the MOT paradigm as a test of the FINST theory. In the original version of the task, subjects were required to track a prespecified subset (1–5) of identical randomly moving objects among of identical distracters. At various moments during movement, one items was flashed, and the subject was to indicate whether the flashed element (the probe)

was a member of the target set or not. If the flash occurred on a target, the subjects were instructed to press a response key as quickly and accurately as possible.

The FINST model with an automatic and parallel tracking mechanism predicts an accurate and flawless tracking performance in this task as a function of tracking time and target set size (when the capacity limit is not exceeded). According to Pylyshyn, relatively errorless performance (85% accuracy with five targets) supports the argument for parallel processing in tracking. In addition, a simulation of a serial strategy showed that it could not account for the observed level of performance in the task (Pylyshyn & Storm, 1988). However, it was also found that tracking performance deteriorated (i.e., accuracy decreased and response latency increased) as the number of targets increased, as reflected in a significant set-size effect (Pylyshyn & Storm, 1988). This finding is not fully consistent with a purely parallel model but shows that either some serial processing takes place or that parallel processing is to some extent resource limited (see the “Serial attention-switching and mixed models” section below).

To accommodate the above-mentioned nonparallel aspect in performance, Pylyshyn and Storm (1988) postulated two processing stages: A parallel tracking stage and a serial response selection stage. Serial processing is invoked by response selection, when properties of the tracked objects need to be checked (i.e., outside the domain of the tracking itself). According to this two-stage model, the serial response stage is responsible for the set-size effect in the MOT paradigm. This might be true for the effect of the number of objects on the time taken to press the response key, because as the number of targets increases it takes longer to serially check whether or not the probed object belongs to the target set. However, it is not clear how the serial response stage can explain a set-size effect in response accuracy, because it is highly unlikely that the yes/no response is forgotten during the short response. Thus, it seems unlikely that a purely parallel model could fully explain tracking performance. At least it may be argued that the issue about the serial versus parallel nature of tracking has not yet been satisfactorily settled.

The question about the nature of individual differences in tracking has not been explicitly raised in the context of the FINST model. The model appears to predict that possible individual differences are distributed around some typical or universal value, say four, with little variance presumably caused by non-interesting random factors or by measurement error. In the original study, Pylyshyn and Storm (1988) do mention the existence of some individual variations in multiple tracking performance. Although their sample size ( $n = 7$ ) was too small to say anything conclusive about individual differences, this finding encourages further work with a larger sample.

### YANTIS' MODEL OF PERCEPTUAL GROUPING

According to Yantis (1992), when participants track multiple moving objects, they spontaneously group individual target elements into a single virtual object or into a higher order perceptual representation. Tracking performance is thus

based on the participant's ability to maintain a perceptual grouping during motion. Thus, performance is assumed to consist of two stages: A group formation and a group maintenance stage. The group formation stage is governed by Gestalt laws of grouping, and it has some similarity to Pylyshyn's initial indexing stage (i.e., it is preattentive, automatic, and stimulus driven; e.g., Pylyshyn & Storm, 1988). However, in contrast with Pylyshyn's idea of automatic sticky indexes, the group maintenance stage is assumed to be goal directed, effortful, and attention demanding. The effortful group maintenance is based on a continuous updating of the representation of the elements as they move around. Yantis provides experimental support for his claim by showing that factors influencing perceptual grouping (whether the objects conformed to the Gestalt law of common fate during motion or whether the objects formed a canonical polygon, e.g., a regular triangle, in their starting positions) also influence the success of maintaining a representation of the tracked objects. According to Yantis, group maintenance is assumed to be closely linked to the processes needed in mental rotation.

Yantis (1992) does not provide any exact prediction about how many objects could be tracked. The specific number may vary greatly depending on the efficiency of perceptual grouping. For instance, when the configuration of elements becomes more complex, the mental transformations needed to update the representation also become more difficult and slower. As a consequence, tracking capacity is different for different configurations of elements. As regards the question about the automatic nature of tracking, Yantis' model provides a clear prediction: Group maintenance is an attention demanding, effortful, and nonautomatic process.

The assumption about continuous mental transformations being responsible for group maintenance links individual differences in MOT to visuospatial processing abilities (i.e., ability to make spatial transformations). The higher the visuospatial processing capacity is, the more efficiently tracking is performed. Thus, tracking is not assumed to rely on an encapsulated system separate from other cognitive processes but carried out using higher cognitive processes like mental rotation and attention.

### OBJECT FILE THEORY OF KAHNEMAN AND TREISMAN

The object-file theory of Kahneman and Treisman (1984; Kahneman et al., 1992) posits that temporary memory representations are necessary when perceiving a dynamically changing visual scene. Unlike the feature-blind visual indexes of Pylyshyn (1989, 1994, 2001), these representations, metaphorically called "object files", collect different types of information (location, feature, semantic) about objects. Although object files gather information about object properties, they are addressed only by spatiotemporal properties of the objects, and not by featural or semantic information (thus, "a frog can turn into a

prince’’). It is further assumed that the number of object files that can be maintained open at the same time is limited.

Kahneman and Treisman (1984; Kahneman et al., 1992) have provided empirical evidence for the temporary object-specific representations by examining priming effects in perception in relation to individual objects (in distinction to nonspecific conceptual effects due to long-term memory). Moreover, they also found evidence for capacity limitations. The object-specific priming benefit decreased as a number of objects to be perceived increased, and the effect disappeared somewhere after four objects (it was statistically significant for four objects but not for eight objects). They assumed that the observed capacity effect was due to a visuospatial, relatively long-lived post-categorical memory, as they were able to rule out an explanation based on iconic or long-term memory. In the standard terminology of cognitive psychology, this kind of memory system often goes as the visuospatial short-term or working memory. Object file representations are thus located in an intermediate level between primitive feature registration and object identification. In their view, the assumed mechanism is then different from both the early vision and identification (long-term memory) systems.

The object file theory does not provide any explicit predictions about how effectively or successfully objects are maintained during tracking. On the other hand, the assumption of a limited visuospatial short-term memory as the basis for maintaining object files appears to lead to a prediction that the number of object files successfully kept open should correlate with the individual capacity differences in visuospatial short-term memory. The larger the visuospatial memory capacity is, the more object files can be kept open during perception of a dynamic scene. Unlike Yantis’ (1992) model, which relates tracking performance to visuospatial processing capacity, the object-file model relates tracking performance to visuospatial working memory capacity. Therefore, on the assumption that visuospatial working memory capacity is normally distributed in the population, a normal distribution would be observed for the individual tracking capacity around an average value, say four or five.

### SERIAL ATTENTION-SWITCHING AND MIXED MODELS

An extreme alternative to Pylyshyn’s early vision FINST model (Pylyshyn, 1989, 1994) and object-based attention models by Yantis (1992) and Kahneman and Treisman (1984; Kahneman et al., 1992) is a serial attention-switching model (described and simulated in Pylyshyn & Storm, 1988; Yantis, 1992). In this type of space-based model (i.e., where only one single spatial region can be attended at any one time, in contrast to object-based models in which attention operates on perceptual objects, not on specific regions of space), successful performance is not achieved by simultaneously tracking several objects but by a

continuous serial and effortful switching of focal attention between the tracked objects. The purpose of these continuous attention shifts is to update coordinates for target items in memory as the targets move about in the display. It may be noted that different kinds of serial models with different spotlight velocity parameters may be postulated. Moreover, it is possible to put forth mixed models that would combine parallel and serial processing elements. One example of mixed models is a limited-capacity parallel model (Pylyshyn & Storm, 1988, discuss this kind of model; for more information about this kind of models, see, e.g., Townsend, 1990). The limited-capacity parallel model assumes that the system has a fixed reserve of resources that is allocated between the tracked objects. The increase in the global processing load is assumed to result in a decrease in the amount of (parallel) resources available per object per unit of time and in difficulties in the allocation of the parallel resources between the target objects. Thus, the serial and mixed models predict that visual tracking is prone to errors as a result of demanding effortful allocation of attention. The exact predictions about the magnitude of set-size effects would depend on the assumed parameters (attention velocity, memory requirements, etc.) of the model. As regards the question of the automaticity in tracking, these models clearly argue that tracking is nonautomatic and highly effortful.

Some versions of a serial model were described and tested in the simulations of Pylyshyn (Pylyshyn & Storm, 1988) and Yantis (1992). They provided evidence that the observed performance in multiple object tracking tasks with, say, four objects (with typical target velocities) is not possible to achieve by a serial switching strategy unless unrealistically fast attention movement or spotlight velocities were assumed. On the other hand, the set-size effect observed by Pylyshyn and Storm (1988) suggests that there must be at least some serial involvement present in MOT. Moreover, it may be noted that although four objects could not be tracked with realistic attentional spotlight velocities, it might be possible to track two objects using a serial strategy (the fewer objects, the slower attention velocities are needed to accomplish successful tracking by continuous attention switching). This means that an errorless tracking performance with two targets may be accounted for both by a purely serial, a purely parallel, or a mixed model, and it would thus be difficult to differentiate between the models on the basis of accuracy data when only a couple of items are tracked.

In the attention-switching models, individual variations in tracking performance are readily predicted and explained by differences in the capacity to serially allocate focal attention and the capacity to update a temporary memory for the target coordinates. Thus, this leads to a prediction that tracking performance should correlate with performance in tasks that demand effortful attentional processing (e.g., continuous task switching) and visuospatial short-term memory. In other words, individual differences are assumed to reflect dif-

ferences in the general attentional capacity and visuospatial short-term memory, but not capacity differences in any specific mechanism unique to tracking. These models also predict a low average tracking capacity with quite large individual variation due to factors related to higher cognition (e.g., attention, processing strategies, visuospatial short-term memory).

Finally, Baddeley's working memory model (see Baddeley, 2000, for the most recent version) would assume that both general attentional resources (provided by a central executive) and the visuospatial sketchpad would be in operation in MOT. The visuospatial sketchpad consists of a temporary storage of visuospatial information and a mechanism that refreshes visuospatial information in the store before it is lost. However, the nature of the rehearsal mechanism is not quite clear (one candidate is implicit motor activity similar to eye movements, Baddeley, 1986; another is shifts of spatial attention, Smyth & Scholey, 1994; and still another is some form of amodal rehearsal mechanism that makes recourse to general attentional resources, Phillips & Christie, 1977).

## THE PRESENT STUDY

The four different types of models of MOT described above differ from each other in many important respects regarding issues of tracking capacity, automatic versus effortful nature of tracking, and the nature of individual differences. To date, only tracking capacity has been studied empirically. The purpose of the present study was to shed more light on these issues by using both an experimental and an individual difference approach.

The key question as to whether the mechanism underlying MOT is an independent encapsulated system separate from higher cognition or, alternatively, a system inherently utilizing other cognitive processes, was examined by means of predicting individual differences in MOT with measures of visuospatial working memory, attentional switching, and mental rotation (Experiment 2). A null or negligible correlation with cognitive task measures would support the claim about the encapsulated system separate from higher cognition. In contrast, if significant correlations (or a significant regression model) can be found relating MOT to cognitive measures, this would support the idea that the tracking mechanism shares processes with other "high level" tasks.

Possible individual differences in tracking capacity are interesting and informative as such. If the variance in the tracking capacity is large among participants, this significant variation should be taken into account theoretically and practically. On the other hand, a unimodal distribution with a narrow variation presumably peaking around four items would be evidence for negligible individual differences (maybe due to noninteresting random factors such as measurement error). In the present study, individual estimates of tracking capacity were calculated and their distributions were assessed from this point of

view. The capacity issue was examined by varying the number of objects to be tracked from two to six. If the set-size effect proved nonsignificant, it would support the idea that the mechanism is fully parallel with a large tracking capacity. In contrast, if a significant linear trend were observed, this would be taken as evidence for a serial process. Significant set-size effects with quadratic trends (with a good performance up to some number of objects, presumably four) would be evidence for parallel capacity with capacity limitations.

The issue of the automatic versus effortful maintenance of objects during tracking was studied by comparing different tracking durations. By studying tracking performance over time we can examine how vulnerable the maintenance process is to temporal performance decrements. In the research on vigilance, a decrement in the signal detection rate has been attributed to limitations in effortful attention (Parasuraman, 1985; Parasuraman & Mouloua, 1987). On the other hand, no sensitivity decrement over time is taken to support the view that signals can be sustained with automatic processes (Fisk & Schneider, 1981). If maintenance is automatic (or indexes are “sticky”), we should not find any large performance or sensitivity (in terms of the signal detection theory) decrements as a function of time, at least within the capacity limits. However, if maintenance of objects demands effortful attention, we should find large performance decrements as a function of time (cf. Fisk & Schneider, 1981). To our knowledge, this issue has not yet been studied empirically.

Two experimental paradigms were used. The first was the standard MOT with identical objects (Experiment 1) and the other was a modified version, where each target and distracter had a unique visual identity (multiple identity tracking; MIT; Experiment 2).

## EXPERIMENT 1

The purpose of Experiment 1 was to examine the issues raised above in the context of the MOT paradigm (Pylyshyn & Storm, 1988; Yantis, 1992). We chose the MOT paradigm with identical elements because it is the first and the most influential experimental task used to study the tracking of multiple moving objects. MOT has its pros and cons. On the one hand, it forms a good testing ground for the hypothesis about the automatic nature of tracking. Using identical elements ensures that the participant must continuously track all the elements. If no performance decrement were found in this task as a function of time-on-task, then the tracking process may be judged to be in some sense automatic. On the other hand, identical elements are nonoptimal to study the processes involved in dynamic binding, when identity information has to be connected to the correct spatiotemporal objects. This motivated us to create a new version of MOT with nonidentical elements in Experiment 2.

MOT was presented to a large sample of participants to examine the general tracking capacity, possible individual capacity differences, the effortful versus

automatic nature of tracking, and the possible relationships of MOT into other measures of cognition. Tracking capacity was measured by manipulating the number of targets tracked. The automatic versus effortful nature of tracking was examined by comparing different tracking durations. Individual differences in tracking capacity were assessed by calculating individual estimates of tracking capacity. The question whether the mechanism underlying MOT performance is an independent encapsulated system separate from higher cognition, or alternatively, makes recourse to other cognitive processes, was examined by correlating individual differences in MOT to other cognitive abilities: visuospatial short-term memory, verbal working memory, and attention switching. Measure of visuospatial short-term memory and attention switching were chosen as we considered these cognitive functions potentially relevant in MOT on the basis of several alternative models to the FINST hypothesis (Pylyshyn, 1989, 1994, 2001). A measure of verbal working memory was included to rule out the possibility that a possible involvement of the visuospatial working memory could be ascribed to a general, modality nonspecific working memory capacity.

## Method

*Participants.* Participants in the experiment numbered 201, 180 males and 21 females. The age of the participants ranged from 19 to 42, the median was 23 years (95% 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). A group of 220 participants was selected to this final selection stage from the total of 1500. The first selection was based on the previous school achievement (mathematics and English); 550 individuals passed these criteria. The final selection from 550 to 220 was based on standardized reasoning tests (a Raven type of intelligence test, a verbal reasoning test, and a mathematical reasoning test) and clinical personality tests. Only those who were among the best 11% in the norm scores (stanine 8 or 9) were selected (the norms were based on a representative sample of normal Finnish young adults). This screening was done as a part of an air pilot recruitment process. Nineteen participants were excluded from the experiment due to a technical error.

*Apparatus.* The stimuli were presented on six 19-inch Eizo FlexScan F730 monitors with a resolution of  $800 \times 600$  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 separate software that generated the motion sequences was written in Visual Basic.

*Stimuli.* The stimuli consisted of 12 identical solid white squares ( $15 \times 15$  pixels) subtending  $0.63^\circ \times 0.63^\circ$  with a dark background, the background

subtending a visual angle of  $25^\circ$  horizontally and  $32^\circ$  vertically. An empty framed square ( $0.63^\circ \times 0.63^\circ$ ) was present during every trial in the centre of the display as the fixation point. A randomly chosen subset of two to six of the total of twelve objects was designated as targets. The remaining objects were distractors.

*Movement sequences.* The experimental trials consisted of 166, 300, or 433 static frames presented one after another for 30 ms each. This yielded an animation sequence of 5, 9, or 13 s in duration. In the resulting motion, items could move a minimum of 3 and a maximum of 6 pixels per frame. Because each frame had duration of 30 ms, the resulting item velocities were in the range of 4.19–8.38°/s. Animation sequences (trajectory files) were generated and stored offline.

Initial item positions were generated at random. Directions for each object were chosen randomly from among the eight compass directions. Each object was assigned a movement duration, randomly selected from 7 to 27 in 30 ms increments (210–810 ms), and speed, randomly selected from 3, 4, 5, or 6 pixels per frame. 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, between objects and the fixation square, and between objects and the edges of the display monitor during the motion phase. Several actions were taken to avoid these collisions. First, an invisible cushion surrounded the objects and the fixation square. Thus, two objects could not be closer than  $0.63^\circ$  apart. Second, before an object was moved to the new position, a possible collision to another object cushion area and to the edges of the display was checked. If a collision was going to happen with some other object, a reverse direction was chosen to the collided objects (the new direction depended on the collision event, 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 the case of potential collision. 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 (210–810 ms or until a collision was about to happen), and then changed direction and speed abruptly and began to move in a new direction.

Thirty trajectory files were generated and stored offline and divided into two sets of fifteen files for the two experimental blocks. 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 and durations were similar. However, the

chosen target and distractor objects were different in different target sets while the trajectory was the same. This technique ensured that any differences between different set-size and duration conditions were due to target set size or duration manipulation, and not to other accidental differences between the trials and trajectory patterns (see Scholl & Pylyshyn, 1999, for a similar procedure).

*Procedure.* Participants (from three to maximum of six at the same time) were seated approximately 57 cm from the display; a chinrest was used to reduce head movements and control the viewing distance. A screen was placed between the participants in order to prevent subjects from seeing or disturbing each other. The “b” and “n” buttons in the keyboard were used to collect the responses. The subjects were instructed to use their dominant hand for the response. Participants were given written instructions prior to the experiment, outlining the general procedure and explaining a trial sequence. They were to note the positions of the flashing targets at the start of each trial and to keep track of them during the movement phase. Subjects were instructed to track the targets without moving their gaze from the fixation cross (eye position was not monitored). At the end of the movement phase, the probe phase occurred. In this phase, all the objects stopped moving and the probe element was highlighted by flashing it five times. The subjects were instructed to respond as accurately and quickly as possible whether the probed element was one of the target items.

At the beginning of each trial, the items and fixation point were displayed for 1 s. After that, between two and six objects flashed on and off for five times (flash duration was 150 ms). The appearance of the remaining objects (distractors) did not change during this target designation phase. All the objects then began to move in a random and continuous fashion around the screen. The participants attempted to simultaneously track each of the target objects. They tracked the targets for 5, 9, or 13 s, after which the movement stopped and a target or a distractor object (the probe item) was flashed off and on five times against the background of the other items. After that, the screen was cleared and a response screen appeared, which asked if the probe was a target or not. After a response had been given, the response screen was cleared and a new intertrial screen was presented. In that screen it was asked if the participant was ready to start the next trial. The next trial was initiated after the participant responded by pressing the space bar. Participants were provided with 10 practice trials; feedback was provided after each response. Each participant completed two blocks of 75 trials. The order of trials was randomized separately for each participant. On half of the trials the probe was one of the targets; on the other half of trials the probe was one of the distractors. There was a short rest period between the blocks. The entire session took about 65 min.

*Method used in the individual ENOT estimation.* To estimate the number of objects the participants were able to track we used a measure called “effective

number of items tracked'' (ENOT) developed by Scholl, Pylyshyn, and Feldman (2001). This measure leans on the assumption that the observed individual performance level is based on correctly tracking some of objects and guessing the others. A derivation and justification of the formula is found in Scholl et al. (2001). The formula is:

$$m = n(2P - 1)$$

where  $m$  is effective number of items tracked,  $n$  a number of targets and  $P$  is empirically observed proportion of correct answers to probe questions.

The procedure for calculating the individual ENOT estimate was the following. First, we calculated a series of ENOT values for the different target set sizes for each participant by using this formula. For instance, one individual series of ENOT values was 1.8, 3.0, 3.7, 4.2, and 4.0 for set sizes from 2 to 6, respectively (i.e., in set size 2 this individual was able to track 1.8 objects effectively, in set size 3 it was 3.0 objects, but in set size 5 it was only 4.2 objects, and in set size 6 it was only 4.0 objects). Next, a final estimation of individual tracking capacity was deduced on the basis of how much the observed performance deviated from the perfect performance. We checked the individual ENOT values one by one for different set sizes to see if the result was within preset tolerance level. The largest tolerated deviation was set to 0.5. If the observed result for a certain set size was within this tolerance limit, that set size was interpreted as successfully tracked. The integer value of the highest set size within tolerance limits was then entered as the total tracking capacity estimate for a participant. Thus, for instance, the series of values given above would yield a total estimate of 4, because the highest value within tolerance limits was 3.7 in set size 4 and the next value in set size 5 deviated too much (0.8) from the allowable limit. A separate ENOT was estimated for the three tracking durations.<sup>1</sup>

Other types of estimation procedure are of course also possible, but we argue that this type of estimation has several advantages: It is simple; it does not make too fine-grained distinctions; it takes into account the progressive nature of set sizes; it estimates the maximum capacity; and it forgives the random lapses of attention during the smaller set sizes. If anything, our individual capacity esti-

---

<sup>1</sup> In 9 s and 13 s durations the procedure yielded a few cases (19 cases or 9.5% of the participants in the 9 s and 13 cases or 6.5% in the 13 s condition), where the estimate appeared to be the highest possible six. However, because in none of these cases the ENOT estimate for set size 5 was not within the tolerance limit but was typically very low (e.g., 2.), these cases were not classified as six—the former estimate within the tolerance limits was used. The participants in these cases probably changed their response criterion and were thus able to improve their performance, as reported below in the section on the sensitivity and response bias. On the other hand, the individual ENOT estimates of six in the 5 s condition were typically a result of a progressively increasing series of ENOT values (i.e., the ENOT value in set size 5 was also within the tolerance limit).

mate overestimates rather than underestimates the true capacity. It resembles the procedure used in the working memory literature in estimating the individual working memory capacity.

*Description of cognitive measures used in correlation and regression analyses.* Individual differences in the visuospatial short-term memory capacity were measured by a computer-aided Corsi-Block-Tapping-Test (Corsi, cited in Milner, 1971; Vienna Test System, 1992a, 1999), in which nine irregularly distributed dice are displayed on the screen. A hand-shaped pointer tapped sequentially on a steadily increasing number of dice (from three to nine). The dices were presented at a rate of 2000 ms, and a die was highlighted for 320 ms. The participant was instructed to tap the dice in the sequence shown. After three runs of each set size, the item length was increased by one die. The test was interrupted if the participant answered three consecutive runs incorrectly. The individual test result is the number of sequences correctly tapped.

To test the ability to continuously switch attention among multiple tasks, a computer-based synthetic work task SYNWORK1 (Elsmore, 1994) was used. During the test trial, the computer screen is divided into four quadrants each assigned to a different task. In the upper left corner of the screen is a memory task (a version of Sternberg, 1969). The initial display in this task consists of seven letters to be memorized. The memory set is then removed and followed by periodic displays of a probe letter, which is to be classified as a member of the set or a nonmember. The upper right corner of the display contains an arithmetic task. In this task, two three-digit numbers are to be added by adjusting plus and minus buttons to produce the correct sum in the row below the addends. This task was the only self-paced task. The lower left quadrant of the display contains a visual monitoring task. In this task, the participant monitors the position of a pointer moving continuously along the horizontal scale and attempts to reset it before it reaches the end of the scale. The lower right quadrant of the display contains an auditory monitoring task. High and low tones are presented periodically throughout the trial, and the task is to respond whenever a high tone occurs. A small window in the centre of the screen was used for displaying a composite performance score on all of the tasks. During training, each task was presented in isolation for 1 min followed by the four tasks presented together for 1 min. The test phase consisted of eight 5 min trials. The parameters for the three computer-paced subtasks in the first 5 trials were the following: The memory probes occurred every 10 s; auditory events occurred every 10 s; the pointer line required 10 s to move from the middle of the scale to the end of the scale. For the last three trials the following changes were made to the parameters of the subtasks: The memory probes occurred every 5 s instead of 10 s; auditory events occurred every 8 s instead of 10 s; the pointer line required 15 s instead of 10. The individual test result was the total score of the fifth 5 min trial.

Verbal working memory capacity was measured by an operation span task (Turner & Engle, 1989), in which the participant saw an equation in the computer screen (multiplication or division; e.g.,  $[6/2] + 6 = 9$ ) one at a time and responded as rapidly and accurately as possible by pressing a Yes or No key in the computer keyboard (the time limit was 15 s). Fifty milliseconds after the key press, a word to be remembered appeared for 800 ms on the screen. The words were high-frequency, two-syllable five-letter Finnish (i.e., the native language of the participants) nouns. The time limit for the presentation of the mathematical equations and the to-be-remembered words was used to reduce the time available to rehearse the words. The equation–word pairs were presented in set sizes ranging from 2 to 8 (set size 2 was used as a practice); the total number of trials was 99. At the end of each trial, the participant saw a question mark and wrote down the words that followed the equations. The operation–word span score was the total number of words correctly recalled (the presentation order had to be preserved).

*Design.* There were two manipulated factors in the experiment: The number of targets tracked (from 2 to 6) and the duration of movement (5, 9, or 13 s). Both variables were within-subject variables. There were 10 trials in each of the 15 conditions; the total of 150 trials were divided into two blocks.

## Results and discussion

The data were submitted to a two-way repeated-measures analysis of variance (ANOVA). Within-subjects variables were target set size (from 2 to 6) and trial duration (5, 9, or 13 s). Moreover, trend analyses were performed as univariate ANOVAs to estimate how well polynomial trends fit the data. Unless otherwise stated, the  $p$  values for all significant statistics reported are less than .001. A Greenhouse-Geisser correction was made to the  $p$  values whenever needed.

*Error rate in performance accuracy.* Table 1 shows the means for the error rate in performance accuracy as a function of duration and target set size. A significant main effect was found for the number of targets,  $F(4, 800) = 237.49$ . That is, performance deteriorated as the number of tracked targets increased. A trend analysis showed that in addition to a linear component (accounts for 87.6% of the variance),  $F(1, 200) = 749.59$ , there was a significant quadratic (6.6%),  $F(1, 200) = 51.42$ , and a cubic (5.3%),  $F(1, 200) = 61.68$ , component. A contrast analysis indicated that the greatest difference was observed between set size 3 and 4 and that the difference between 5 and 6 did not reach significance: 2 vs. 3 targets,  $F(1, 200) = 48.88$ ; 3 vs. 4 targets,  $F(1, 200) = 223.23$ ; 4 vs. 5 targets,  $F(1, 200) = 32.92$ ; and 5 vs. 6 targets,  $F(1, 200) = 2.98$ ,  $p > .05$ . The strong linear trend is consistent with serial models of MOT, whereas the presence of the higher order trends and participants' surprisingly good performance in set size 6

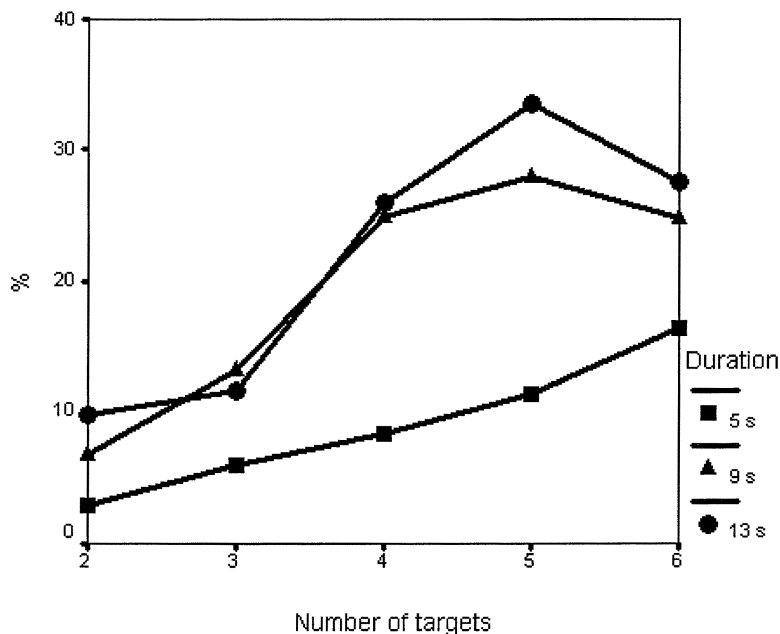
TABLE 1  
 Percentage of trials on which the participant failed to correctly respond to the probe, for the five levels of set size (2–6) and the three levels of trial duration (5, 9, and 13 s)

Duration	Target set size					Overall mean
	2	3	4	5	6	
5 s	2.84	5.92	8.36	11.29	16.42	8.97
9 s	6.67	13.09	24.98	27.91	24.73	19.48
13 s	9.70	11.64	25.82	33.48	27.56	21.64
Overall mean	6.40	10.22	19.72	24.23	22.90	16.69

suggests that a nonlinear mechanism may also need to be considered when explaining the task performance.

The main effect of trial duration was significant,  $F(2, 400) = 403.09$ . The trend analysis showed that in addition to a large linear trend (accounts for 87.4% of the variability),  $F(1, 200) = 735.10$ , there was also a significant quadratic trend (accounts for 12.6% of the variability),  $F(1, 200) = 97.56$ , in the data. The greatest decline in performance occurred between 5 and 9 s (the error percentage increased from 9.0 to 19.5), after which the performance levelled off (the error percentage in the longest condition was 21.6). These data indicate that the tracking performance seems surprisingly susceptible to the effects of time.

The interaction between the number of targets and the trajectory duration was significant,  $F(8, 1600) = 26.43$  (see Figure 1). The trend analysis showed that there were significant higher order components present, in addition to the linear one (a linear set size and a linear duration effects account only for 26.0% of the variability of the interaction; linear duration and quadratic set-size effects account for 21.4%; quadratic duration and quadratic set-size effects account for 14.8%; linear duration and cubic set-size effects account for 34.4%). Separate one-way ANOVAs (set size as the repeated factor) as well as trend and repeated contrast analyses were conducted for each different trajectory duration to decompose this interaction. A simple main effect of set size was significant for all durations: 5 s,  $F(4, 800) = 66.56$ ; 9 s,  $F(4, 800) = 113.08$ ; and 13 s,  $F(4, 800) = 137.80$ . Trend analyses showed that higher order nonlinear trends were only present in the 9 s (linear 77.7%, quadratic 17.0%, and cubic 4.0%) and in the 13 s conditions (76.1%, 8.1%, and 15.3%, respectively), whereas the data in the 5 s condition were almost completely linear (97.9%). Contrast analyses showed that in the 5 s condition there was a steady growing linear function with a significant difference in the error rates between adjacent set sizes, but in the longer durations (9 and 13 s) there was a steep decline up to five targets and a



**Figure 1.** Error rates (%) as a function of the number of targets tracked (2–6) and the tracking duration (5 s, 9 s, and 13 s).

surprising improvement in performance when six targets were tracked (causing a higher order cubic-like trend, see Figure 1). This difference in the trends between the shortest and the two longer duration conditions caused clearly the interaction. Below, we provide a plausible explanation for this admittedly rather surprising result (see the analysis of response bias).

*Sensitivity and response bias.* Table 2 shows the mean percentages of correct signal detection (i.e., hits) as a function of tracking time and target set size. The false-alarm rates are shown in Table 3. The parametric sensitivity index  $d'$  was then computed from the hit and false-alarm rates. The  $d'$  values are shown in Table 4. The  $d'$  values were submitted to a similar ANOVA as before. ANOVA of the  $d'$  scores yielded a significant effect for both the main effects and the interaction: set size,  $F(4, 800) = 259.87$ ; duration,  $F(2, 400) = 440.02$ ; Set size  $\times$  Duration,  $F(8, 1600) = 21.09$ .

There was a similar significant set-size effect for the sensitivity as was found in the error data reported above. The most interesting result was a significant sensitivity decrement as a function of duration. The greatest decline in performance occurred between 5 and 9 s ( $d'$  dropped from 3.54 to 2.45), and the sensitivity levelled off in the longest condition ( $d'$  was 2.28). As is evident from

TABLE 2  
Mean hit rate (target probe correctly detected), for the five levels of set size (2–6) and the three levels of trial duration (5, 9, and 13 s)

<i>Duration</i>	<i>Target set size</i>					<i>Overall mean</i>
	2	3	4	5	6	
5 s	.948	.972	.935	.891	.874	.924
9 s	.899	.838	.659	.723	.774	.779
13 s	.819	.870	.637	.685	.741	.750
Overall mean	.889	.893	.744	.766	.796	.818

TABLE 3  
Mean false-alarm rate (distractor probe not detected), for the five levels of set size (2–6) and the three levels of trial duration (5, 9, and 13 s)

<i>Duration</i>	<i>Target set size</i>					<i>Overall mean</i>
	2	3	4	5	6	
5 s	.005	.091	.102	.116	.202	.103
9 s	.032	.100	.158	.282	.269	.168
13 s	.013	.102	.153	.354	.293	.183
Overall mean	.017	.098	.138	.251	.254	.151

TABLE 4  
Mean value of sensitivity  $d'$ , for the five levels of set size (2–6) and the three levels of trial duration (5, 9, and 13 s)

<i>Duration</i>	<i>Target set size</i>					<i>Overall mean</i>
	2	3	4	5	6	
5 s	4.26	3.86	3.63	3.26	2.70	3.54
9 s	3.79	3.02	1.92	1.61	1.92	2.45
13 s	3.44	3.16	1.93	1.19	1.67	2.28
Overall mean	3.83	3.35	2.49	2.02	2.10	2.76

Table 4 and confirmed by a reliable Set size  $\times$  Duration interaction, the decrement was greater when the participants had to track four or five targets, whereas in set sizes 2 and 3 the sensitivity decline was much smaller.

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 data. The  $C$  values are shown in Table 5. In the ANOVA, both main effects and the

TABLE 5  
 Mean value of the criterion cutoff C, for the five levels of set size (2–6)  
 and the three levels of trial duration (5, 9, and 13 s)

Duration	Target set size					Overall mean
	2	3	4	5	6	
5 s	.163	-.199	-.100	-.014	-.241	-.078
9 s	.202	.197	.410	.006	-.091	.145
13 s	.515	.064	.453	-.062	-.045	.185
Overall mean	.293	.021	.254	-.023	-.125	.084

interaction all proved significant: set size,  $F(4, 800) = 60.13$ ; duration,  $F(2, 400) = 61.03$ ; Set size  $\times$  Duration,  $F(8, 1600) = 14.59$ .

The set-size effect suggests a criterion shift in the task performance. As is evident from Table 5, the C value is neutral or positive (i.e., a bias toward a “no” response) from set-size 2 to set-size 5, but in the set-size 6 the criterion shifts to negative (i.e., a bias toward a “yes” response; the C values were 0.29, 0.02, 0.25,  $-0.02$ , and  $-0.13$ , for set sizes 2–6, respectively). This pattern of results is understandable, if we assume that in set sizes that are within the capacity limits, the participants are better able to discriminate between targets and distractors and are thus more conservative in responding (i.e., if in doubt, a “no” response is given). On the other hand, in set size 6, which is above their capacity limit, they change their strategy to a more liberal one, giving relatively frequently a “yes” response in order to improve their performance. This more liberal response bias can at least partly explain the surprising improvement for set size 6, for which the participants had a poorer knowledge about the location of targets and distractors, which then probed them to change their strategy to optimize successful performance. To our surprise they succeeded in this and could in fact improve their performance (see Figure 1).

*Interim discussion.* Our results quite closely replicate the overall tracking accuracy reported by Pylyshyn and Storm (1988), although the accuracy in the present study tended to be little lower (averaged over the different tracking durations). For set sizes 2 and 3 the percentages were almost exactly the same in both studies (93% vs. 90%), whereas for set sizes 4 and 5 performance accuracy was somewhat lower here (80.3% and 75.8% vs. 90% and 85.6% in Pylyshyn & Storm). There are several possible reasons for the 10% difference: A practice effect caused by a huge number of trials in the original study (550 trials); individual capacity differences (the original sample comprised only seven participants), and/or an effect of the larger number of distractors in our experiment compared to the original study (Pylyshyn & Storm had a total of 10

moving objects, whereas we had 12). However, the observed difference cannot be ascribed to tracking duration, as the durations used in the present study and in that of Pylyshyn and Storm were almost identical.

The observed set-size effect in accuracy and sensitivity, and the lower overall performance accuracy, do not support fully Pylyshyn's claim about parallel tracking of four or five target elements. Moreover, the overall tracking capacity was found to be modulated by the duration of tracking. While tracking accuracy in the 5 s condition was quite good up to six targets (although a significant linear trend for set size was observed), in the longer durations the accuracy was only half of that, and it declined steeply when more than three targets were tracked. Thus, on the basis of these results it does not seem possible to make any generalizations about a universal tracking capacity.

The decremental effect of tracking time on task performance was surprisingly large. This result is also inconsistent with Pylyshyn's claims about the automatic nature of tracking when tracking is done within the capacity limits (i.e., up to four or five targets). Instead, we found the greatest sensitivity decline when four or five targets had to be tracked for relatively longer periods of time (9–13 s). Thus, these data suggest that the visual indexes, if they existed, may not be as sticky as Pylyshyn has assumed but are vulnerable to time decrements. These results may also be explained by assuming (1) that tracking requires effortful switching of the attentional spotlight from one target to another, and (2) that the continuous allocation of attentional resources is vulnerable to deterioration as a function of time. We return to this issue below when we predict individual capacity differences by a measure of attention switching.

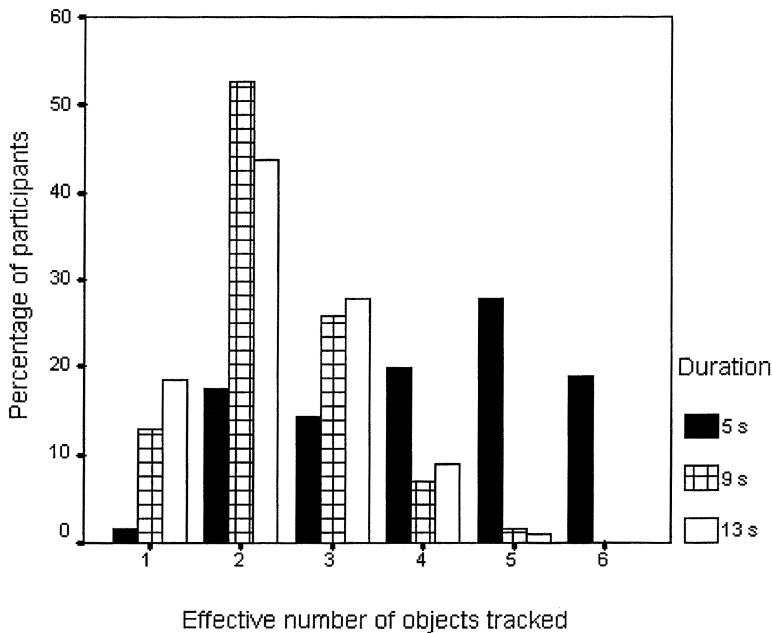
A strange improvement occurred in the tracking performance when six targets were presented. Our analyses of response bias strongly suggest that one reason for this was a shift in the response strategy. Because six targets are clearly above the capacity of every participant, they changed their response criterion to a more liberal one (i.e., to a general tendency to accept all probed items as targets). In addition, it is also possible that they adopted a strategy to track only subgroups of target items (cf. Yantis, 1992). This may explain why the performance did not continue to deteriorate when the number of targets was increased from five to six, but it cannot explain why the performance was actually slightly improved. On the other hand, the improvement may be accounted for by the shift in the response strategy mentioned above.

*Individual differences in the tracking capacity.* Next, we turn to examine individual differences in MOT. In their seminal paper, Pylyshyn and Storm (1988) reported some individual differences in the error scores: They ranged from 0% to 8% when tracking one target and from 6% to 28% when five targets were tracked. However, they did not discuss the implications of this variation. We argue that if individual variation is substantial and systematic, it has to be taken into account in theorizing about MOT (substantial individual variation

would go against the idea that, across individuals, there exists four or five ‘‘sticky fingers’’, or FINSTs, which are responsible for a successful MOT). The individual differences we observed in our large sample were indeed substantial. The error percentages varied from 0% to 27% in set size 2, from 0% to 43% in set size 3, from 0% to 53% in set sizes 4 and 5, and from 0% to 50% in set size 6. This means that in the larger target set sizes (>3) some participants could track all the targets, while some others performed at the chance level.

In the following analyses, we first examined more systematically how many objects each participant could track effectively. Individual capacity estimates were calculated and their distributions were assessed. Second, the nature of the underlying mechanism responsible for MOT performance (i.e., whether it is an independent encapsulated system separate from higher cognition, or a system that makes use of mental processes common to other cognitive tasks) was examined by predicting individual differences in MOT with other cognitive measures (visuospatial short-term memory, verbal working memory, and attention switching).

The distribution of the participants’ highest ENOT values is shown in Figure 2. Capacity estimates for different durations were calculated separately, because the different tracking durations behaved very differently (see above). For the 5 s



**Figure 2.** The distribution of the individual tracking capacity in terms of ENOT (effective number of objects tracked) in MOT (Experiment 1), as a function of tracking duration (5 s, 9 s, and 13 s).

duration, the median of the highest ENOT value was 4 ( $SD = 1.42$ ); for the 9 s and 13 s durations the median was 2 ( $SD$  was 0.84 and 0.91, respectively).

As is evident from Figure 2, the distribution of the ENOT values in the 5 s condition is far from normal and the variation is surprisingly large. The participants are distributed quite evenly across the different ENOT levels (i.e., from two to six). Thus, 33% of the participants could track fewer than four objects whilst 19% could even track six. Although the overall mean for the 5 s duration is almost exactly in line with Pylyshyn's claims (1989, 1994, 2001) about the number of available visual indexes, the observed distribution clearly suggests that individual differences are substantial. This large variation and the relatively even distribution of ENOT are not consistent with the notion of a universal tracking capacity. The distribution of the ENOT values in the 9 s and 13 s conditions is very different from that of the 5 s condition. In the longer tracking durations the variation is smaller and the mean is only about two. An average capacity of about two is also inconsistent with that suggested by the FINST model. In sum, these results are evidence against a general and universal multiple object tracking capacity.

*Correlations to other cognitive measures: To what extent do other cognitive processes contribute to MOT performance?* The data were explored for potential outliers using Mahalanobis and Cooks' distances, with no outliers detected. The data for Corsi and SYNWORK1 were missed for one participant, who was discarded. Correlations between MOT and other cognitive measures (i.e., visuospatial short-term memory, verbal working memory, and attention switching) are presented in Table 6. Most of the reported correlations are significant but relatively small. It is very likely that the small correlations were caused by the restricted ability range in the sample due to a strict preselection procedure described in the Method section. Because of that, it is likely that the correlations would have been much higher with a more heterogeneous sample, and thus these apparently small correlations should be taken seriously.<sup>2</sup> The results of Experiment 2 provide further support for this contention.

The question about the relative contribution of the other measures to MOT performance was addressed with a series of hierarchical multiple regression analyses with Corsi, SYNWORK1, and Operation span as the predictor variables. The Corsi tapping task was used as a measure of visuospatial short-term memory. SYNWORK1 was used as measure of multiple-task performance,

---

<sup>2</sup>The range restriction problem may be remedied by adopting a formula (e.g., in Ghiselli, Campbell, & Zedeck, 1981) that uses the standard deviation from a sample with an unrestricted range to correct the correlations in the restricted sample. If a population estimate were known for the standard deviation of Corsi and SYNWORK1, then we could correct the correlations reported here. Unfortunately, they were not available. On the basis of the results of Experiment 2 with a less preselected sample and with much higher correlations we expect that the true correlations between MOT, Corsi, and SYNWORK1 would be notably higher than observed here.

TABLE 6  
The correlations between MOT performance and other cognitive measures (Corsi, Operation span, SYNWORK1)

	<i>MOT</i>	<i>Corsi</i>	<i>Operation span</i>
Corsi	.219**		
Operation span	.091	.225**	
SYNWORK1	.211**	.231**	.255**

\*\* Correlation is significant at the .01 level (2-tailed).

which involves continuous and effortful attention switching and time-sharing between the to-be-performed tasks (for task-switching and attention control, see, e.g., Gopher, 1993; Pashler, 2000; for using SYNWORK1 as a measure of multiple-task performance, see, e.g., Salthouse, Hambrick, Lukas, & Dell, 1996). The operation span task was used as a measure of verbal working memory.

The order of entry of the variables was varied to identify the shared and unique variance attributable to each variable. First, Corsi and SYNWORK1 were in turn entered as the first variable. After that, Operation span was added to the model including Corsi and SYNWORK1, because the nonsignificant correlation between operation span task and MOT suggested that it was the least important factor. The results of these analyses are summarized in Table 7, which presents the regression analyses for each variable combination.

TABLE 7  
Summary of the results of hierarchical regression analyses of MOT

<i>Step</i>	<i>Variable</i>	$\Delta R^2$	<i>Variable</i>	$\Delta R^2$
Overall MOT score, predicted by Corsi and SYNWORK1				
1	Corsi	.048*	SYNWORK1	.045*
2	SYNWORK1	.027*	Corsi	.031*

The model predicts for 7.5% of the variance; the unique contribution of SYNWORK1 is 2.7%; the unique contribution of Corsi is 3.1%; their shared contribution is 1.7%.

Overall MOT score, predicted by adding Operation span to the model				
1	Corsi+SYNWORK1	.075**	Operation span	.008
2	Operation span	.00	Corsi+SYNWORK1	.067*

The model predicts for 7.5% of the variance; the unique contribution of Operation span is 0%; the unique contribution of Corsi+SYNWORK1 is 6.7%, the overall shared variance is 0.8%.

\*\*  $p < .01$ ; \*  $p < .05$ .

The first analysis with the overall MOT performance as the dependent variable shows that Corsi and SYNWORK1 together accounted for 7.5% of the variance. SYNWORK1 uniquely accounted for 2.7% of the variance when entered after Corsi,  $F(1, 197) = 5.82$ ,  $p < .05$ , and Corsi explained 3.1% of the variance when entered after SYNWORK1,  $F(1, 197) = 6.53$ ,  $p < .05$ . This indicates that the shared contribution of these variables was 1.7%. The second analysis shows that the operation span task does not have any unique contribution to the model with Corsi and SYNWORK1,  $F(1, 196) = 0.01$ .

The significant contributions of Corsi and SYNWORK1 to MOT performance are generally in line with the object file theory and with an attention switching model (or a mixed model), but inconsistent with the FINST theory. The amount of variance explained by these variables is low (about only 7.5%), but we strongly suspect that these figures would be much higher without the range restriction problem discussed above.

## EXPERIMENT 2

In the traditional MOT task used in Experiment 1 targets and nontargets were all visually identical. Thus, the objects were tracked only on the basis of their spatiotemporal properties. However, in a more realistic dynamic visual environment, the tracked targets are typically visually distinct from each other and have different identities. In this kind of situation the observer has to continuously bind the identities and spatiotemporal trajectories of the objects in order to know, where each object is located at any given time. To study this, we developed a new version of the task, coined the multiple identity tracking (MIT) task. Similarly to MOT, participants are required to track from two to six targets. However, the tracked objects are now visually different from each other. We used two types of stimuli, drawings of familiar objects (e.g., a picture of a coat) and pseudo-objects (i.e., object-looking items without a known identity). At the end of the movement phase all items were masked and one of them was probed. Then a list of objects was presented, and the participant was asked to identify the probed target (the response is given by clicking with the mouse on the chosen item).

Although the task is different, the predictions of the theories described in the Introduction are exactly same here with nonidentical objects. The feature-blind indexes of FINST theory (e.g., Pylyshyn, 1989), the spatiotemporally addressed object files (Kahneman et al., 1992), or a serial algorithm based on the spatial coordinates of objects are not assumed to make use of any featural or semantic information from the objects—only their spatiotemporal features. Notice, however, that the hypotheses about the automatic versus effortful nature of tracking are not possible to test in this context, because all the objects are visually distinct. When an observer does not track a certain target, she/he does not lose that target completely in this new paradigm since she/he can always find

that target again by using its visual identity. In contrast, when the observer misses a target in the traditional MOT task, she/he loses the targets completely, because the objects are identical except for their spatiotemporal histories.

The primary purpose of Experiment 2 was to study further the issue about the different cognitive functions present in multiple object tracking. The results of Experiment 1 suggest that visuospatial short-term memory and attention switching may play a significant role in MOT performance. In Experiment 2, further corroborating evidence was sought for the assumed relationships using a new tracking task and a less restricted participant sample (see the Method section for more details). In addition, two new hypotheses were tested about possible relations of MOT to other cognitive functions. First, Yantis (1992) hypothesized that MOT shares features with mental rotation. To test the possible contribution of mental rotation to MOT performance, we added a measure of the mental rotation ability to the list of predictor variables. Second, we wanted to rule out the possibility that the observed relationships of the predictors to MOT may be explained by general intelligence in terms of Spearman (1927). We did not think it was very likely that general intelligence would be a relevant contributor; nevertheless, a measure of fluid intelligence (Raven's Advanced Progressive Matrices = APM) was added to the list of predictors, because fluid intelligence has often proved to be good or even the best predictor of a range of complex cognitive activities (see Ree & Carretta, 1996; Ree & Earles, 1996).

We also added a new measure of visuospatial working memory capacity, coined the Spat-Span (Oinonen, 2002). This new version was a "complex span" task (in the spirit of Daneman & Carpenter, 1980) that requires participants to simultaneously process and memorize visuospatial information. Mental rotation was used as the processing task, and the memory task was similar to the Corsi test used in Experiment 1. The motivation to use this new complex span version was that this type of measure is often proved to be a better predictor than a simple short-term memory task in a range of complex cognitive activities (e.g., Daneman & Carpenter, 1980). It is suggested that the complex span task including a processing component prevents the use of mnemonic strategies and thus provide a better, noncontaminated measure of the short-term memory (Cantor, Engle, & Hamilton, 1991; la Pointe & Engle, 1990).

## Method

*Participants.* Participants in the experiment numbered 78, 76 male and 2 female subjects. The age of the each participant was about 20. The experiment was conducted as a part of a large test battery that was administered to the Finnish Air Force applicants. All participants had normal, uncorrected vision. The previous selection before that stage was based on the previous school achievement (mathematics and English). Note that in contrast to Experiment 1,

for the preselection of this sample no psychological tests were used. Thus, the range of abilities in the sample was less restricted than in Experiment 1.

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

*Stimuli.* Two sets of six stimuli were used, familiar objects and pseudo-objects. As familiar objects, six vertically oriented line drawings of real objects (flower, coat, lobster, rocking chair, rooster, and watch) were used. The objects were selected to represent different semantic categories. The pictures were selected and scanned from a standardized set of black-and-white line drawings (see Snodgrass & Vanderwart, 1980, Appendix A, Figures 91, 125, 142, 188, 191, and 250). 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 were above average. The nonobject targets consisted of six pseudo-objects with an object-like appearance. Vertically oriented, visually rather complex, black-and-white line drawings were selected from the pictures provided by Kroll and Potter (1984, see Appendix, Figure A-1, Figures 3, 4, 10, 26, 68, and 85), which were similar to each other and to the real objects. A visual complexity estimation was done subjectively by the authors, because the visual complexity of the nonobjects was not rated by Kroll and Potter.

The stimuli (75 pixels in height and from 41 to 69 pixels in width) were black outline drawings on a white background subtending a visual angle of  $2.4 \times 1.3$ – $1.7^\circ$ . The computer screen subtended a visual angle of  $25^\circ$  horizontally and  $32^\circ$  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 nontargets 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 2, twelve times in the target set 3, etc.). After the movement phase, the target probing was carried out by flashing a black frame ( $75 \times 75$  pixels, 2 pixels in width,  $2.4 \times 2.4^\circ$ ) around the target. The frames were not visible during the movement phase. Visual masks ( $75 \times 75$  pixels) of 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.

*Movement sequences.* The experimental trials consisted of 167, 233, or 300 static frames presented one after another for 30 ms each. This yielded an

animation sequence of 5, 7, or 9 s, respectively. In the resulting motion, items could move a minimum of 3 and a maximum of 9 pixels per frame. Because each frame had duration of 30 ms, the resulting item velocities were in the range of 2.6–7.7°/s. Each object was assigned a movement duration randomly selected from 7 to 37 in 30 ms increments (210–1110 ms), and speed, randomly selected from 3, 5, 7, or 9 pixels per frame. Twenty-four animation sequences (trajectory files) were generated and stored offline and divided into two sets of twelve files for two experimental blocks. The same trajectory files were used for both stimulus versions (i.e., for objects and nonobjects; for further details, see the Method section of Experiment 1).

*Procedure.* Participants (maximum of ten at the same 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 subjects from seeing or disturbing each other. The mouse was used to collect the responses. 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 1 s. After that, a black frame flashed on and off 10 times (flashing duration was 150 ms; total flashing time was 3000 ms) around the designated targets (2–6 objects). The objects then began to move in a random and continuous fashion around the screen. The participants tracked the targets for 5, 7, or 9 s, after which the movement stopped, and the objects were masked. After that a black frame flashed for five times (flashing duration was 150 ms; total flashing time was 1500 ms) around one of the target objects (i.e., the probed item). Finally the screen was cleared and a response screen appeared, where all the presented six pictures were arranged into an array with two rows and three columns. A response frame (100 × 100 pixels) surrounded each picture. The initial mouse pointer position was set into a middle of the array between the rows. The participant was asked to select (point and click within a framed picture) the probed picture as accurately and quickly as possible. He/she was asked to guess if she/he did not know the answer. After the response was given, the response screen was cleared and an intertrial screen was presented. The next trial was initiated by the participant pressing the space bar or after the maximum intertrial time (3000 ms) expired. Participants were provided with 10 practice trials; feedback was given after each response in the practice session. Each participant completed two blocks of 60 trials for each experimental set (i.e., the real objects and the pseudo-objects), 240 trials altogether. The order of trials was randomized separately for each participant. The order of the experimental sets was

counterbalanced across participants. The probed item was always one of the targets. There was a short rest period between the blocks. The entire session took about 75 minutes.

*Description of cognitive measures used in correlation and regression analyses.* The reader is asked to consult Experiment 1 for details about the computer-aided Corsi-Block-Tapping-Test, SYNWORK1, and the operation span task. The speed of mental or image rotation was measured by a variant of the task used by Dror and Kosslyn (1994). The task required the participants to determine as fast as possible whether or not two shapes were identical, regardless of their orientation. Pairs of shapes were shown simultaneously on the computer screen, one on the left and one on the right. The shape on the right was always the one rotated, and the shape on the left was always presented upright and used for reference. The rightmost shape was rotated 45°, 90°, 125°, 180°, 225°, 270°, or 315° clockwise. The shapes used in the task were composed of squares. A square corresponded approximately to 1° × 1° of visual angle. One shape always consisted of three juxtaposed rectangular elements, which included one, two, or three squares. The whole shapes were approximately 4° × 3° of visual angle and the distance between the middle points of the shapes was approximately 7° of visual angle. A black square was placed at the top of each shape, which helped the subjects to locate the corresponding parts of figures. The task included 12 practice trials and 84 test trials. In half of the trials the rightmost shape was identical to the leftmost shape and in half they were mirror reversed. The pairs were presented in a random order. During a trial the participant studied a pair of shapes and then decided whether the shapes were identical. If they were, the subject pressed K and if not (in that case they were mirror reversed), she/he pressed D in the keyboard. Immediately after the subject responded, another pair of shapes were presented. The pair of shapes was present for a maximum of 15 s. The mean response times for the correct responses were calculated and used as the measure of mental rotation ability in the subsequent correlational analyses. The task was programmed in E-prime (Schneider et al., 2002a, 2002b).

Visuospatial working memory was measured by a complex span type of task (Spat-Span), in which both the processing and the storage task components were visuospatial in nature. The processing component was the mental rotation task described above. The storage component was a variant of the Corsi block task used by Smyth and Scholey (1994). In the Corsi block task, an array of nine squares was presented to measure the individual visuospatial short-term memory span. The squares were white, outlined in black. During each trial, a black spot appeared on a subset of squares. The squares were 1.5° × 1.5° of visual angle and the diameter of spot was 0.5° of visual angle. The size of whole pattern of squares was 10° × 10° of visual angle. During a trial a black spot was presented on one of the nine squares for 1000 ms. Immediately after that a mask pattern,

which consisted of black spots on each square, was presented for 250 ms. The second spot was presented on some other square followed again by a mask. This went on as long as all the spots of each span level were presented. The participant's task was to recall in the proper order the squares where a spot was presented; she/he responded by clicking the squares by mouse. The response time was unlimited. Two-spot trials were presented first as practice. The test proper started with three-spot trials; in the subsequent trials, task difficulty was increased by adding the number of presented spots, up to eight spots. Each span level was tested with three trials. A trial was scored correct if all the spots were recalled in the correct order.

The mental rotation task and the Corsi block task were combined in a similar manner as in operation span tasks (measuring verbal working memory capacity), where a processing and a storage trial are presented after each other. At the beginning of a trial, a pair of shapes was presented, and the participant responded as fast as possible whether or not the shapes were identical (for more details, see the description of the mental rotation task above). Immediately after the response, the pair disappeared, and the array of nine squares used in the Corsi task appeared after 50 ms (see above for more details). After responding to the Corsi task, the participant initiated a new trial by pressing the space bar. The total number of sequences recalled in the correct order was calculated and used in correlation analyses (other possible measures, such as the number of rotation tasks correct or the average rotation speed were not used). The task was programmed in E-prime (Schneider et al., 2002a, 2002b).

In the standardized Raven's Advanced Progressive Matrices (APM) test (Raven, 1958), participants were presented with 36 patterns and 2 practice items. Each pattern included a missing piece. For each pattern, six alternatives to fill in the pattern were presented and the participants selected the one that fit. The patterns increased in difficulty as one progressed through the test. There was a 40 min time limit. A computerized version of APM was used (Vienna Test System, 1992b, 1999). The total number of correct items was used in the correlational analyses.

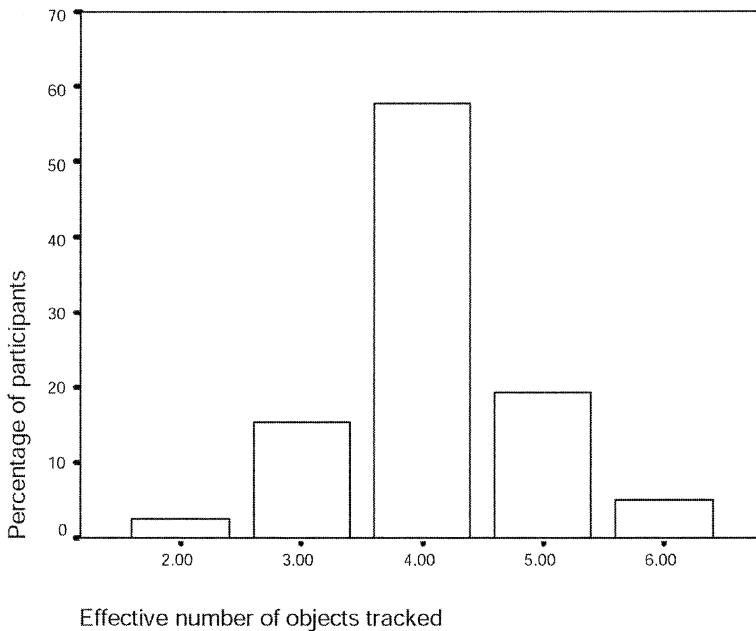
## Results and discussion

*Individual differences in tracking capacity.* Individual capacity estimates were calculated using a similar procedure described in the Method section of Experiment 1. The formula described in Scholl et al. (2001) is based on the probability of guessing at the level of .5 in the standard MOT task. However, in the present version of the task the probability of guessing is not 0.5 but 1/6, because the participant has to choose the response among six alternatives. A new formula adapted to this guessing level was derived. The new formula is:

$$m = n(6P - 1)/5$$

where  $m$  is effective number of items tracked,  $n$  is number of targets, and  $P$  is an empirically observed proportion of correct answers to probe questions. The distribution of the participants' highest ENOT values is shown in Figure 3. As is evident from the figure, the median value is four ( $SD = 0.81$ ), and most values are concentrated on the ENOT level of four. The observed median is in line with Pylyshyn's claims (1989, 1994, 2001) about the number of available visual indexes, and it is also line with the results of Kahneman and Treisman (1984; Kahneman et al., 1992) obtained using a very different methodology. However, again substantial interindividual variation existed with about the same performance range as was observed in Experiment 1 for 9 s and 13 s durations. Next we examine the extent to which this variation is accounted for by other cognitive measures.

*Correlations to other cognitive measures: To what extent do other cognitive processes contribute to the MIT performance?* The data were explored for potential outliers using Mahalanobis and Cooks' distances. One influential outlier with a very high performance level in MIT and a zero performance in Spat-Span was detected and removed from the data. The correlations between the MIT performance and the other cognitive measures (i.e., Corsi, SYN-



**Figure 3.** The distribution of the individual tracking capacity in terms of ENOT (effective number of objects tracked) in MIT (Experiment 2).

TABLE 8  
 The correlations between the MIT performance and other cognitive measures  
 (Spat-Span, Corsi, APM, SYNWORK1, Operation span)

	<i>MIT</i>	<i>Spat-Span</i>	<i>Corsi</i>	<i>APM</i>	<i>SYNWORK1</i>	<i>Operation span</i>
Spat-span	.594**	—				
Corsi	.396**	.636**	—			
APM	.426**	.462**	.436**	—		
SYNWORK1	.405**	.303**	.201	.167	—	
Operation span	.283*	.309**	.249**	.293**	.293**	—
Mental rotation	-.244*	-.270*	-.130	-.388**	-.324**	.047

\*\* Correlation is significant at the .01 level (2-tailed).

\* Correlation is significant at the .05 level (2-tailed).

WORK1, operation span, Spat-Span, and APM) are presented in Table 8.<sup>3</sup> Most of these correlations were highly significant reflecting a high degree of interrelationship among the variables. For instance, Corsi and SYNWORK1 correlations to the MIT performance are now twice the size to those observed in Experiment 1. This finding is consistent with our argument about the restricted range of abilities causing lower correlations in Experiment 1.

Experiment 1 provided evidence for the view that visuospatial short-term memory and attentional switching may be involved in multiple object tracking. In the first set of regression analyses, we sought for further corroborating evidence for this view. However, for reasons given above, Corsi was replaced by Spat-Span as a measure of visuospatial working memory. Analogously to Experiment 1, the order of entry of these variables was manipulated to identify the shared and unique variance attributable to each. The results of these analyses are summarized in Table 9.

The first analysis shows that Spat-Span and SYNWORK1 jointly accounted for 40.8% of the variance. Spat-Span uniquely accounted for 24.4% of the variance when entered after SYNWORK1,  $F(1, 74) = 30.55, p < .001$ , and SYNWORK1 explained 5.6% of the variance when entered after Spat-Span,  $F(1, 74) = 6.99, p < .05$ . This indicates that the shared contribution of these

<sup>3</sup> For the SYNWORK1 measure, we chose the sixth 5 min round as the predictor, while in Experiment 1 the fifth round was used. The reason was that the sixth round correlated somewhat more strongly with MIT than the fifth round (the difference in the correlations was only about .03). The difference between the fifth and the sixth round was that the sixth round was performed with changed parameter settings of the tasks (see the Method section of Experiment 1). The reason why the fifth and the sixth round were chosen was based on a pilot study by Marja-Leena Haavisto where participants (psychology students) performed 16 × 5 min rounds in SYNWORK1. The data of this pilot study showed that the learning curve in terms of standard deviation and mean levelled off at the fifth round. The first rounds before the fifth seem to be susceptible to random variation due to initial learning phase in a multiple task environment.

TABLE 9  
Summary of the results of hierarchical regression analyses of MIT

<i>Step</i>	<i>Variable</i>	$\Delta R^2$	<i>Variable</i>	$\Delta R^2$
MIT score, predicted by Spat-Span and SYNWORK1				
1	SYNWORK1	.164**	Spat-Span	.353**
2	Spat-Span	.244**	SYNWORK1	.056*
The model predicts for 40.8% of the variance; the unique contribution of Spat-Span is 24.4%; the unique contribution of SYNWORK1 is 5.6%; their shared contribution is 10.8%.				
MIT score, predicted by adding APM to the model				
1	Spat-Span+SYNWORK1	.408**	APM	.181**
2	APM	.027	Spat-Span+SYNWORK1	.254**
The model predicts for 43.5% of the variance; the unique contribution of Spat-Span+SYNWORK1 is 25.4%; the unique contribution of APM is 2.7%; their shared contribution is 15.4%.				
MIT score, predicted by adding Operation span to the model				
1	Spat-Span+SYNWORK1	.408**	Operation span	.080*
2	Operation span	.003	Spat-Span+SYNWORK1	.331**
The model predicts for 41.1% of the variance; the unique contribution of Spat-Span+SYNWORK1 is 33.1%; the unique contribution of Operation span is 0.3%; their shared contribution is 7.7%.				
MIT score, predicted by adding Mental rotation to the model				
1	Spat-Span+SYNWORK1	.408**	Mental rotation	.060*
2	Mental rotation	.001	Spat-Span+SYNWORK1	.349**
The model predicts for 40.9% of the variance; the unique contribution of Spat-Span+SYNWORK1 is 34.9%; the unique contribution of Mental rotation is 0.1%; their shared contribution is 5.9%.				
MIT score, predicted by SYNWORK1 and Corsi (analogously to Experiment 1)				
1	SYNWORK1	.149**	Corsi	.172**
2	Corsi	.111*	SYNWORK1	.087*
The model predicts for 26.0% of the variance; the unique contribution of Corsi is 11.1%; the unique contribution of SYNWORK1 is 8.7%; their shared contribution is 6.2%.				

\*\* $p < .01$ ; \* $p < .05$ .

variables was quite large (10.8%), which presumably reflects task demands common to them. There are two alternative explanations for this shared contribution. It may either be attributed to visuospatial memory requirements in both tasks, or it is related to attention switching requirements present in both tasks. The first alternative does not seem very likely, because there is no apparent

visuospatial memory task in SYNWORK1. On the other hand, the second option appears much more likely, because also in our visuospatial working memory task the participant must continuously switch her/his attention between processing and storage modes. By attributing the shared contribution between Spat-Span and SYNWORK1 to the allocation of attentional resources, the relative importance of attention switching would rise from a rather small unique contribution of SYNWORK1 up to over 16% (5.6% SYNWORK1-unique + 10.8% shared).

The subsequent analyses indicated that none of the other variables (APM, Operation span, or Mental rotation) made any unique contribution to the MIT performance when entered after the model involving Spat-Span and SYNWORK1 ( $p > .05$ ). However, they share some variance with the two variables of our main model; especially APM has a substantial shared contribution with the model (the shared variance of APM was 15.4%, that of Operation span was 7.7%, and that of Mental rotation was 5.9%). These shared contributions reflect common task demands that may either be attributed to visuospatial working memory constraints or to the time-sharing of attentional resources between multiple tasks. In particular, the high shared contribution between the main model and APM seems to reflect a strong visuospatial working memory involvement in APM. This became apparent from regression analyses, in which the shared contribution between APM and Spat-Span (i.e., without SYNWORK1) was 15.3%, while the shared contribution between APM and SYNWORK1 (i.e., Spat-Span excluded) was only 4.9%. For a successful completion of APM, the participant needs to be able to keep active in working memory many different visuospatial patterns (matrices) while studying different solutions or problem spaces.

In sum, based on the above analyses we argue that the most important predictor of the MIT performance is visuospatial working memory, which explains about 25% of the interindividual variance in MIT. Another significant predictor is the attention switching ability, which accounts for up to 16% of the variance (although its unique contribution is clearly smaller). On the other hand, while tasks measuring verbal working memory, mental rotation or fluid intelligence do show significant correlations with the MIT performance, the regression analyses suggest that their contribution is transmitted via the cognitive mechanisms related to visuospatial working memory and/or attention switching.

## GENERAL DISCUSSION

In the present study, several issues related to the mental processes underlying multiple object tracking (MOT) were examined. The studied issues included the possible contribution of higher order cognition to tracking, the general tracking capacity, the size of individual differences in tracking capacity, and the automatic versus effortful nature of tracking. Existing theories relevant to multiple object tracking provide qualitatively different predictions concerning these

issues, as discussed in the introduction to this paper. These predictions were examined in two experiments. Experiment 1 employed the traditional MOT paradigm with identical objects, and in Experiment 2 a new version of the tracking task (multiple identity tracking; MIT) was used, where nonidentical (either visually or both visually and semantically) distinct objects had to be tracked.

The following findings were obtained: (1) Both experiments demonstrated a substantial individual variation in the estimated tracking capacity of MOT and MIT. (2) Tasks measuring visuospatial short-term and working memory as well as attention switching proved to be significant predictors of multiple object tracking. (3) Tracking performance deteriorated as a function of tracking time. This is taken as evidence for the view that tracking is not automatic, but the maintenance of the targets during tracking is attention demanding. (4) The tracking performance deteriorated linearly as a function of set size. Thus, tracking is not fully parallel but also involves a serial component, or alternatively, tracking is based on capacity limited parallel processing. In the next section, we discuss the theories described in the introduction in the light of our main results.

### The compatibility of our data with theories of multiple object tracking

*Parallel processing model of Pylyshyn.* The visual indexing model of Pylyshyn (1989, 1994, 2001) assumes that MOT is based on an encapsulated, low-level mechanism, which operates preattentively, in parallel, and automatically. No general memory system is needed, because independent local processing components carry out the tracking. Clearly, our results stand in sharp contrast to these assumptions. Our results suggest that in tracking visuospatial working memory and attention is required, serial processing is involved, and the process is effortful and non-automatic. We were quite surprised about how much the general conclusions we are forced to make on the basis of our results depart from those made by the visual indexing theory that has been very influential in the field. It is quite obvious that our individual differences approach clearly provides a new perspective on the phenomenon. However, our results may not be considered so detrimental to the FINST theory, if one assumes that the observed correlations are not associated with “pure tracking” but reflect additional processes related to the tracking task as a whole (see Pylyshyn, 2001; Pylyshyn et al., 1994), such as responding to the probed item, during which one may be required to scan through the indexes provided by the early vision mechanism. To test this view, one would need an empirical measure of MOT that is “clean” from all other processes that may intervene. At present, we cannot think what such a measure would be.

*Serial processing and attention switching models.* In contrast to the parallel models, serial attention switching models assume tracking to be based entirely on an attentional spotlight and the spatial memory of the coordinates of the to-be-tracked-targets (a serial algorithm is described in Pylyshyn & Storm, 1988; Yantis, 1992). Our results are generally in line with these premises. The observation that visuospatial working memory and attention switching are significant predictors of multiple object tracking fits perfectly well with the assumptions of the serial model. So does our conclusion about the effortful nature of tracking and the relatively linear set size effect.

The serial mechanism needs memory for the target coordinates simply because it is the only way for it to know where the targets are; it does not have a direct access to anything outside of the scope of the narrow spotlight. The system obtains these coordinates in the dynamic situation by constantly shifting its spotlight from one target to another. The purpose of these discrete attention shifts is to encode and update the target coordinates. Otherwise the system would immediately lose track of the moving objects. At the end of tracking, the serial system uses the spatial memory for the target coordinates to decide whether or not the probed item was among the target set (in MOT of Experiment 1) or what identity the probed target had (in MIT of Experiment 2).

This kind of serial system needs a very accurate temporary memory system and very fast attention movements. Models of this kind have been frequently rejected in the simulations by Pylyshyn and Storm (1988) and Yantis (1992), because they require an unrealistic fast attention velocity. Yantis estimated that the attention movement velocity must be at the level of 150–200°/s in order to accomplish the empirically observed performance. However, the estimated level of human attention movement is assumed to be clearly below that (an informal meta-analysis by Pylyshyn & Storm, 1988 suggests a velocity of 50°/s). In addition, the algorithm has other problems as described in Pylyshyn and Storm and Yantis. First, if we assume that the system does not know anything but the last coordinates of the targets (perceptually a very stupid algorithm) when attention returns back from its “switching tour” to the same coordinates again, the target previously in that location is no longer there. As a consequence, the system has to start to search the target. During the search, there is a high possibility that the system chooses a wrong object that happens to be in a near location. The second additional problem of the completely serial models is that even an addition of an extra assumption about the system’s ability to extrapolate the movement vector does not improve the performance (in fact in the simulations of Pylyshyn & Storm the performance deteriorated, because the extrapolation takes time, and moreover, target movement in the MOT paradigm is so random that extrapolations are mainly incorrect).

The above arguments seem to force us to reject a purely serial model as unrealistic. Most seriously, the acceptance of a purely serial model would require an assumption of attentional velocities in the order of 150°/s that appears

to go well beyond human capacity. How should we deal with this dilemma? There seem to be at least two options available. Either, we may argue that the estimated attention movement velocity (about  $50^\circ/\text{s}$ ) is wrong, and the correct one is closer to  $150^\circ/\text{s}$ . Or alternatively, we may try to sketch a model that assimilates the results without requiring a very fast attention movement velocity.

The first alternative about the attention velocity being closer to  $150^\circ/\text{s}$  seems unlikely in the light of the present knowledge, although perhaps further research is needed on this important issue.<sup>4</sup> Very recently, a new tracking paradigm, multiobject permanence tracking (MOPT), was devised by Saiki (2002). The MOPT task consists of multiple items (e.g., four coloured disks) and an occluder pattern superimposed on them. Both the items and the occluder move regularly in a circular route. The setting creates a perception of items alternately appearing and disappearing behind the occluder. The task of an observer is to detect changes in the moving targets with different item velocities. Saiki found that the performance deteriorated monotonically as a function of the target velocity. With a velocity of  $126^\circ/\text{s}$  (the highest velocity used), the performance was at a chance level. It may be noted that the task is quite different from MOT and MIT in that the movement is nonrandom in MOPT. If anything, nonrandom movement may be assumed to be associated with a better overall performance level than random movement. Thus, it seems safe to conclude that Saiki's results demonstrate that attention movement in MOT or MIT has to be below  $126^\circ/\text{s}$ , which in turn leads to the rejection of a purely serial model. To solve the dilemma, we are left with the second alternative of constructing a model that is neither completely parallel nor completely serial. In the next section, we outline a mixed model capable of explaining and assimilating the observed pattern of results.

*Outline of a mixed model consistent with our results.* The proposed model is a mixed model containing both a parallel and a serial processing component as well as a temporary memory component. The lower level parallel component provides a coarse percept of the target objects in motion and their movement vectors. Although the component may have some resemblance to the visual indexes proposed by Pylyshyn, contrary to the FINST model, the parallel component is not assumed to be preattentive and automatic, but effortful attention is presumed to be needed for its successful operation.<sup>5</sup> A key feature in our mixed model is that the representation provided by the parallel component is vulnerable to time decay and interference caused by the existence of other

---

<sup>4</sup> In contrast to the assumption that attention moves continuously and analogously in space there exists a possibility that attention can move in a discontinuous fashion without any actual movement or without any time cost (see, e.g., Cave & Bichot, 1999, and Egeth & Yantis, 1997 for a discussion of this possibility).

<sup>5</sup> That attention is required for motion detection is convincingly demonstrated by Mack and Rock's (1998) inattentional blindness studies.

objects in the vicinity of target objects. Thus, a higher order serial attentional mechanism is needed to overcome the vulnerability of the parallel system. The main purpose of the serial component is to refresh the object representations active in temporary memory. The temporary memory for target coordinates is needed in shifting the attention to the correct locations. However, in contrast with the purely serial models, the scanning of the targets is not assumed to be exhaustive. When a tracked item is about to go below an activation threshold and/or intermix with another nearby object, presumably an alarm signal is sent from the lower level processes to the attentional system to guide the attentional spotlight to that location to raise its activation level and/or to prevent it from intermixing with a distractor object. The more targets to be tracked, the greater number of these alarm signals, and the higher the probability that the system is no longer capable of maintaining all the targets in a sufficiently active stage.

The mixed model proposed above can assimilate the pattern of observed results concerning the individual capacity differences, the relevant role of visuospatial working memory and attention, the effortful nature of tracking, and the significant set-size effect. The proposed model differs from purely parallel and serial models by assuming that the system responsible for tracking includes both a parallel and a serial component. Both components have their inherent weaknesses that can be overcome by the other subsystem. The weakness of the lower level system capable of parallel tracking is the maintenance of the multiple targets for longer motion durations. The weakness of the higher level serial subsystem, on the other hand, is its inability to constantly update all the targets (particularly when there are more than two or three targets) by discrete attention scanning due to insufficiently slow attention movement velocity. A tracking mechanism that utilizes both of the two subsystems is not vulnerable to the weaknesses of each subsystem working alone (within the limits of individual capacity).

*Baddeley's multicomponent working memory model.* The observed results also fit well with the multicomponent working memory model of Baddeley (1986; Baddeley & Hitch, 1974). The model contains three components: A modality-free central executive resembling focused attention, a verbal memory subsystem (the phonological loop system), and a visuospatial memory subsystem (the sketchpad; Baddeley, 2000, has recently also added a new subsystem to his model, the episodic buffer). The overall performance of the system during a cognitively demanding task may involve both subsystems (depending on the nature of stimuli and the type of mental processes used to temporarily store information) and the central executive. The model contains many of the characteristics that would be required to accommodate the results of the present study. The required visuospatial memory system would be provided by the sketchpad, while the required modality-free attention allocation component would be provided by the central executive. Also, as suggested by our data, Baddeley's working memory

model has a limited capacity both in terms of the functions of the central executive and of the subsystems. The model is in a sense a mixed system: Memory subsystems have an ability to keep simultaneously several items active in short-term memory, while the functioning of the central executive is serial in nature. The working memory performance is assumed to deteriorate quickly as a function of time unless effortful maintenance is used—also a feature consistent with our results. The problem is, however, that the model is not very specific in details and does not provide very detailed predictions regarding MOT performance.

There is an interesting finding observed in the research of visuospatial working memory that fits nicely with our results. Phillips and Christie (1977) observed that the maintenance in the visuospatial working memory is closely related to the general processing resources (i.e., to the central executive). However, the specific mechanism that would mediate the rehearsal in the visuospatial working memory has remained somewhat unclear. Awh, Jonides, and Reuter-Lorenz (1998) and Smyth and Scholey (1994) have claimed that the online maintenance of visuospatial information involves active shifts of spatial attention. The results observed in the present study with significant correlations between tracking and a modality-free attention switching task and a modality-specific spatial temporary memory task are compatible with these claims. The above mentioned claims are also consistent with our mixed model, according to which the function of active attention shifts during multiple object tracking is to maintain the tracked objects by periodically refreshing the activation of the target representations.

*Object file theory by Kahneman and Treisman.* Our conclusion about the importance of temporary spatial memory representations in MOT is in line with the object file theory of Kahneman and Treisman (1984; Kahneman et al., 1992). Our other conclusions about the effortful nature of tracking, the role of serial attention, and the individual differences are not perhaps directly inconsistent with the object file theory, but they are not predicted by the theory either. The maintenance of several object files for longer periods of time seems to be an unspecified area in the theory. Our results suggest that the maintenance of object files is attentionally demanding and effortful, whereas the object file theory seems to assume that the maintenance (or reviewing of object files) is automatic.

*Perceptual grouping by Yantis.* The model by Yantis (1992) claims that MOT performance is based on the ability to update a “virtual” representation of the spatial coordinates of the targets.<sup>6</sup> According to Yantis, the

---

<sup>6</sup>Pylyshyn et al. (1994) argue for a post-tracking “error-recovery” stage, which is also based on polygon-like representation of the relative location of targets. However, in our view this proposal seems to be somehow different from Yantis’ (1992) proposal because it is not used during tracking.

maintenance of this “virtual polygon” is effortful—a claim consistent with our results. However, Yantis’ hypothesis about the mental rotation ability being a relevant factor in MOT was not supported by the data of Experiment 2. Unfortunately, we did not include a mental rotation task in Experiment 1 where identical moving objects were tracked. Thus, we cannot rule out the possibility that perceptual grouping could be in operation when tracking identical objects (perceptual grouping is probably much more difficult in case of visually and semantically distinct items as employed in Experiment 2). However, it may be concluded that the significance of perceptual grouping in tracking is perhaps limited to visual scenes that give rise to Gestalt-type percepts.

## CONCLUSIONS

The observed substantial individual differences in MOT and MIT seem to be largely based on individual capacity differences in effortful higher order cognitive processes, i.e., temporary spatial memory and attention switching. Our conclusion is that MOT performance is not based on an encapsulated “cognitively impenetrable” mechanism (cf. Pylyshyn, 1999) but on a much more cognitively open system utilizing both goal-driven and stimulus-driven attentive and perceptual processes as well as spatial memory representations. Our conclusions are also relevant to the discussion about space-based versus object-based attention (Duncan, 1984). Our results suggest that space-based attention plays a surprisingly significant role in an apparently object-based task.

## REFERENCES

- Awh, E., Jonides, J., & Reuter-Lorenz, P. (1998). Rehearsal in spatial working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 780–790.
- Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Oxford University Press.
- Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Science*, 4, 417–423.
- Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (pp. 47–89). London: Academic Press.
- Cantor, J., Engle, R. W., & Hamilton, G. (1991). Short-term memory, working memory and verbal abilities: How do they relate? *Intelligence*, 15, 229–246.
- Cave, K. R., & Bichot, N. P. (1999). Visuospatial attention: Beyond a spotlight model. *Psychonomic Bulletin and Review*, 6, 204–223.
- Culham, J. C., Brandt, S., Cavanagh, P., Kanwisher, N. G., Dale, A. M., & Tootell, R. B. H. (1998). Cortical fMRI activation produced by attentive tracking of moving targets. *Journal of Neurophysiology*, 80, 2657–2670.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.
- Dror, I. E., & Kosslyn, S. M. (1994). Mental imagery and aging. *Psychology and Aging*, 9, 90–102.
- Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, 113, 501–517.

- Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual Review of Psychology*, *48*, 267–97.
- Elsmore, T. F. (1994). SYNWORK1: A PC-based tool for assessment of performance in a simulated work environment. *Behavior Research Methods, Instrumentation, and Computers*, *26*, 421–426.
- Fisk, A. D., & Schneider, W. (1981). Control and automatic processing during tasks requiring sustained attention: A new approach to vigilance. *Human Factors*, *23*, 737–750.
- Ghiselli, E., Campbell, J., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W.H. Freeman.
- Gopher, D. (1993). The skill of attention control: Acquisition and execution of attention strategies. In D. E. Meyer & S. M. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 299–322). Cambridge, MA: MIT 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.
- 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.
- La Pointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1118–1133.
- Mack, A., & Rock, I. (1998). *Inattention blindness*. Cambridge, MA: MIT Press.
- Milner, B. (1971). Interhemispheric differences in the location of psychological processes in man. *British Medical Bulletin*, *27*, 272–277.
- Oinonen, K. (2002). *Ikää ja kokemusta—kadettien ja keski-ikäisten upseereiden suoriutuminen kielellisessä ja visuaalispataalisessa työmuistitehtävissä* [Age and experience—Visuospatial and verbal performance of younger army and air force cadets and middle-age officers]. Unpublished master's thesis.
- Parasuraman, R. (1985). Sustained attention: A multifactorial approach. In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance XI: Attention and neuropsychology*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Parasuraman, R., & Mouloua, M. (1987). Interaction of signal discriminability and task type in vigilance decrement. *Perception and Psychophysics*, *41*, 17–22.
- Pashler, H. (2000). Task switching and multitask performance. In S. Monsell & J. Driver (Eds.), *Attention and performance XVIII: Control of cognitive processes*. Cambridge, MA: MIT Press.
- Phillips, W. A., & Christie, D. F. M. (1977). Interference with visualization. *Quarterly Journal of Experimental Psychology*, *29*, 637–650.
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3–25.
- Pylyshyn, Z., 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. (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. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*, 341–423.
- Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, *80*, 127–158.
- Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*, 1–19.
- Raven, J. C. (1958). *Advanced Progressive Matrices*. London: Lewis.

- Ree, M. J., & Carretta, T. R. (1996). Central role of g in military pilot selection. *International Journal of Aviation Psychology*, 6, 111–123.
- Ree, M. J., & Earles, J. A. (1996). Predicting occupational criteria: Not much more than g. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 151–166). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- 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). Amsterdam: Elsevier Science.
- Salthouse, T. A., Hambrick, D. Z., Lukas, K. E., & Dell, T. C. (1996). Determinants of adult age differences on synthetic work performance. *Journal of Experimental Psychology: Applied*, 2, 305–329.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002a). *E-prime user's guide*. Pittsburgh, PA: Psychology Software Tools, Inc.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002b). *E-prime reference guide*. Pittsburgh, PA: Psychology Software Tools, Inc.
- 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.
- Smyth, M. M., & Scholey, K. A. (1994). Interference in immediate spatial memory. *Memory and Cognition*, 22, 1–13.
- 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.
- Spearman, C. (1927). *The abilities of man: Their nature and measurement*. London: Macmillan.
- Sternberg, S. (1969). The discovery of processing stages: Extension of Donders' method. *Acta Psychologica*, 30, 276–315.
- Townsend, J. T. (1990). Serial and parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science*, 1, 46–54.
- Turner, M. L., & Engle, E. W. (1989). Is working memory task dependent? *Journal of Memory and Language*, 28, 127–154.
- Vienna Test System. (1992a). *Corsi-Block-Tapping-Test manual*. Mödling, Austria: Gernot Schufried GmbH.
- Vienna Test System. (1992b). *Raven's Advanced Progressive Matrices (APM) manual*. Mödling, Austria: Gernot Schufried GmbH.
- Vienna Test System. (1999). *Vienna Test System hardware: Description*. Mödling, Austria: Gernot Schufried GmbH.
- Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology*, 24, 295–340.

*Original manuscript received July 2002*

*Revised manuscript received November 2002*

Copyright of Visual Cognition is the property of Psychology Press (T&F) and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.