Article Type: Original Article

Visual Search without Selective Attention: A Cognitive Architecture Account

David Edward Kieras (kieras@umich.edu)

Electrical Engineering & Computer Science Department, University of Michigan

2260 Hayward Street, Ann Arbor MI 48109-2121, USA

Keywords: cognitive architecture, visual search; cognitive modeling; eye movements

Abstract

A key phenomenon in visual search experiments is the linear relation of RT to the number of objects to be searched (set size). The dominant theory of visual search claims that this is a result of covert selective attention operating sequentially to "bind" visual features into objects, and this mechanism operates differently depending on the nature of the search task and the visual features involved, causing the slope of the RT as a function of set size to range from zero to large values. However, a cognitive architectural model presented here shows these effects on RT in three different search task conditions can be easily obtained from basic visual mechanisms, eye movements, and simple task strategies. No selective attention mechanism is needed. In addition, there are little-explored effects of visual crowding which is typically confounded with set size in visual search experiments. Including a simple mechanism for crowding in the model also allows it to account for significant effects on error rate (ER). The resulting model shows the interaction between visual mechanisms and task strategy, and thus represents a more comprehensive and fruitful approach to visual search than the dominant theory.

Abstract word count = 190 in Word

Body word count, references and figure captions section not included = 6749 in Word

1. Introduction

Visual search, the process of finding a desired object in a visual scene, is a common real-life task, and understanding it better is important for improved design of systems such as computer displays. For the decades, an especially simple visual search task has been the focus of considerable empirical and theoretical work, starting with the seminal work of Triesman & Gelade (1980), which was extended by Wolfe and his coworkers, starting with Wolfe, Cave, & Franzel (1989). In this task, subjects view a display containing several objects, and decide whether a specified *target* object is present or not, and make a corresponding keystroke response. The main independent variable is the number of objects on the display (*set size*), and the main dependent variable is the reaction time (*RT*), the time to make the response. Normally the target is present half the time (*positive* trials), and absent the other half (*negative* trials). Additional independent variables are the visual properties specified for the target and distractors, and the logical form of the target specification. For example, the target might be a single red bar among green bars, or the target might be a *conjunctive* combination of two features, such as a blue X shape among red X and blue O shapes.

The key result in these experiments (see reviews by Wolfe, 2014; Hulleman & Olivers, 2017) is a roughly linear increase in RT with set size, with negative trials producing a slope about twice as steep as positive trials. This pattern suggests a classical serial self-terminating process in which each object is examined sequentially, and the search terminated as soon as the target is found. Depending on the task conditions, positive trial slopes range from essentially zero (e.g. the target is a single red bar among green bars) to about 50 ms/item or more (e.g. a specific detailed shape among similar detailed shapes). Error rate (*ER*) is generally fairly low, and so is often ignored, but usually increases with set size and apparent task difficulty.

1.1. Covert attention theory of visual search

An obvious explanation for the linear RT effects is that subjects move the eyes to each item sequentially to perform the search. However, the typical slopes observed are much faster than

eye movements would allow. This discrepancy underlies the basic theoretical claim originally made, and still dominant in this literature, that the sequential search is done not by overtly moving the eyes, but instead by covertly moving selective attention from one object representation to another. This *covert selective attention* theory of visual search has its roots in Neisser's (1967) assertion, based on extremely early computer vision concepts, that "focal attention" is necessary to bind together primitive features into a visual object; this attention-based "binding" operation was advanced in Triesman & Gelade (1980) as an explanation for why conjunctive searches had much steeper slopes than single-feature searches. Wolfe, Cave, and Franzel (1989) tried many different visual features and search specifications and discovered that conjunctive searches could have small slopes similar to some single-feature searches. They proposed the first version of the Guided Search theory which still involved covert attention allocation as its fundamental mechanism.

However, the covert attention theory is seriously flawed, as eloquently pointed out by Findlay and Gilchrist (2003). Visual search theorists and experimenters have generally ignored the role of powerful purely visual factors, such as how visual resolution decreases from the fovea towards the periphery, but objects can still be recognized in peripheral vision if they are large enough (e.g. Anstis, 1974; see review in Rosenholtz, 2016). Another visual factor is crowding effects, in which objects in peripheral vision become harder to perceive if other objects are nearby (for reviews, see Levi, 2008; Pelli & Tillman, 2008). This effect could be important in visual search tasks because usually the objects are displayed in a fixed area, so as the set size is increased, the objects tend to be closer together; but this confounding has usually been ignored in visual search experiments. Finally, both of these factors are the basic reason why eye movements are necessary in visual tasks — moving the eyes to the object of interest improves the resolution and eliminates crowding effects, yielding accurate perception of the object. But in fact there is little or no mention of either visual factors nor eye movements in Neisser's (1967) original treatment of focal attention, nor in the subsequent mainstream of visual search work pioneered by Treisman and Wolfe, even though several studies demonstrated their relevance (e.g., Zelinsky & Sheinberg, 1995; Carrasco & Frieder, 1996; Wertheim, Hooge, Krikke, & Johnson, 2006). It has even been claimed that the RT effects are the same regardless of whether or not eye movements are made, but this claim is problematic (c.f., Carrasco, McLean, Katz, & Frieder, 1998). Thus the dominant theory of visual search ignores known visual factors and eye

movements, and instead insists that the key mechanism in visual search is the allocation of covert attention.

1.2. Active vision alternative

Findlay & Gilchrist (2003) proposed an *active vision* approach to visual search in which information from peripheral vision is used to guide eye movements that bring the high-resolution portion of the retina to bear on relevant parts of the scene. Furthermore, for many visual properties and displays, more than one object can be perceived in a single fixation, which is the long-standing concept of the *area of conspicuity* (Engel, 1977) or *functional viewing field* (FVF, see review in Hulleman & Olivers, 2017). The claim that the RT ms/item slopes are too fast for eye movements clearly fails if it is possible for more than one object to be processed at a time; the notion that individual objects would have to be foveated is simply incorrect. Accordingly, Hulleman and Olivers (2017) proposed that the ms/item characterization of visual search was a fundamental mistake, because the number of fixations, not the number of display items, accounts for visual search RT, and presented a simple process model based on the FVF that accounted for RT effects. This paper goes further and presents an active vision model using the EPIC cognitive architecture (Meyer & Kieras, 1997; Kieras, 2016), which has no conventional selective attention mechanism and is especially suitable for modeling perceptual-motor tasks that are controlled by cognitive strategies. This model demonstrates that visual factors and eye movements, together with simple cognitive task strategies, are sufficient to account for both RT and error effects in visual search tasks without any mechanism of covert selective attention.

1.3. Overview

This paper next presents the methodology and data analysis of a very high-quality visual search data set on performance in three classic visual search tasks, made available by Wolfe, Palmer, & Horowitz (2010). Next comes an active vision model of these results based on the EPIC cognitive architecture. This model is then compared in detail to the Wolfe et al. (2010) data for both RT and error rate (ER).

2. The Visual Search Experiment

Rather than spend time and resources collecting new data to test the active vision model, it is more useful to test it with previous data of the type used to support the original theories. Many

variations on the simple visual search task have been studied, and some classic examples were reported in Wolfe et al. (1989) in support of their Guided Search theory. Subsequently, additional data in these tasks were reported by Wolfe, Palmer, and Horowitz (2010) to support a theoretical analysis based on the details of the RT distributions for individual subjects. They made the data publicly available for download at http://search.bwh.harvard.edu/new/data_set_files.html. This dataset was ideal for the present modeling work because it was collected by arguably the most experienced visual search laboratory, had well-specified stimuli and task conditions suitable for replication in a model, and a relatively large number of very well-practiced subjects, which means that the mean data would be reasonably reliable and individual subject strategies were likely to be stable, making the results especially suitable for modeling. For completeness and clarity, their experimental method is re-stated here, but with additional details on how the experiment was simulated in the EPIC model based on the details in Wolfe et al. (2010).

2.1. Method

Tasks. Wolfe et al. (2010) used three different present/absent search tasks; Fig. 1 shows a sample target-present display produced by the EPIC software for each task condition. In this paper, the three conditions are referred to as Color Single Feature (CSF), Color-Orientation Conjunction (COC), and Shape (SHP). The CSF target was a red vertical bar among green vertical distractors. The COC target was a red vertical bar among distractors that were red horizontal bars or green vertical bars. The SHP target was a "digital 2" shape among "digital 5" shapes.

---------- Insert Fig. 1 about here ----------

Stimuli. The Wolfe et al. (2010) download data set includes each individual trial but does not contain the actual display configuration used in each trial, so for purposes of modeling, the display had to be generated for each simulated trial using their display parameters. The search display was an area 22.5° × 22.5°, treated as containing 25 invisible cells of 5° × 5°. In the CSF task, the objects were 1° × 3.5° vertical bars; in the COC task, the objects were 1° × 3.5° bars, oriented either horizontally or vertically. In the SHP task, the objects were 1.5° × 2.7° character-like shapes. Each object appeared in a random location within one of the cells, constrained in the model to keep the horizontal or vertical edge of an object at least 0.25° away from the cell

boundary, ensuring a minimum separation of 0.5° between adjacent objects. Set sizes were 3, 6, 12, and 18. To generate the display for each trial, the set size number of distractors were first placed in randomly chosen display cells; if the trial was positive (target present), a randomly chosen distractor was replaced with a target object.

*Design***.** The Wolfe et al. (2010) experiment had 10 subjects in the COC task condition and 9 in each of the other two. One subject was in both COC and SHP, but the data set does not identify this subject, so the task condition was treated as a purely between-subject manipulation in this paper. The set size and polarity were chosen at random for each trial. There were about 500 trials per subject for each combination of set size and positive/negative trial polarity.

Procedure. Each trial began with a centered fixation cross. Subjects were instructed to "keep their eyes focused on this cross" but eye movements were not monitored. The search display was presented and remained visible until the subject pressed a key for target-present or target-absent. Subjects were instructed to respond as "quickly and accurately as possible." Correct/incorrect feedback was presented for 500 ms after each trial.

2.2. Results

The downloaded data consisted of the RT and correct/incorrect status for each subject in each trial at each set size and trial polarity. Following common practice in RT experiments, the data were reduced as follows: For each task condition, for each subject, the mean RT for correct trials and the proportion of errors for that subject was calculated for positive and negative trials at each set size, giving a total of 8 data points for each subject for their RT and error rate (ER). These subject means were then averaged to produce the observed data points plotted in Fig. 2 and 3. The 95% confidence intervals around each data point are based the standard error of that mean using the underlying 9 or 10 individual subject means, thus reflecting between-subject variability, but not within-subject variability.

Since they were concerned with the detailed RT distributions, Wolfe et al. (2010) did not report any conventional overall statistical tests of main effects and interactions. Therefore, for this paper, unequal-*N* ANOVAs were performed using the **R ez** package on the mean values provided by each subject in each cell of the design. For RT, the main effects of Task Condition, Trial Polarity, Set Size, and all two- and three-way interactions were significant ($p < .05$). For ER, whose overall average was 2.4%, the Task Condition main effect was not significant ($p > .1$)

but the Trial Polarity and Set Size main effects, and all two- and three-way interactions were significant ($p < .05$).

---------- Insert Fig. 2 about here ----------

---------- Insert Fig. 3 about here ----------

2.3. Discussion

The RT results follow the classic pattern obtained in most experiments with this visual search task, where the slope (determined by regression analysis) is the key theoretical measure. The RT functions for the CSF task are essentially flat for both positive and negative trials in CSF (positive trial regression slope is about 1 ms/item); this prominent effect with the color property in a single-feature search task is frequently described as "pop out". Otherwise, positive and negative trial RTs have a substantial slope, with the negative trial slope about twice that of the positive trials. The color-orientation conjunction task COC has a positive trial slope of about 9 ms/item and the SHP positive trial task slopes are much greater at 43 ms/item. The error rate (ER) overall is only 2.4%, which would justify the conventional approach of focusing the theoretical analysis only on the correct trial RT. However, note that negative trials have a fairly constant low False Alarm error rate averaging 1.4%, while positive trials produce more Miss errors as set size increases, especially for the more apparently difficult tasks. Overall, this pattern rules out a speed-accuracy tradeoff effect in the RT data, but because these ER effects are statistically reliable in spite of the small number of subjects and large between-subject variability, a good theory would attempt to explain them in addition to the RT effects.

3. An EPIC model for visual search RT and ER

3.1. Summary of the EPIC cognitive architecture

The EPIC architecture for human cognition and performance provides a general computational framework for simulating a human interacting with an environment to accomplish a task. The original modeling domain was skilled performance in multitasking; the EPIC acronym reflects how *E*xecutive *P*rocesses exert *I*nteractive *C*ontrol over perceptual and motor systems to coordinate performance. Meyer & Kieras (1997) or Kieras (2016) provide detailed descriptions; the following summarizes the components of the architecture relevant to the model presented

here.

EPIC is especially suitable for computational simulation modeling in human-performance domains because it treats both perceptual and motor processes as first-class components and has a minimal set of cognitive mechanisms for executing task strategy instead of traditional mechanisms dating from pre-computational cognitive theory. Thus, EPIC has components in which the visual perceptual, ocular and manual motor, and strategy aspects of the model are explicitly represented. The visual perceptual component captures the concept of the FVF. The oculomotor component represents the mechanisms that generate saccades with realistic timing and variability. The strategy component consists of production rules applied by the cognitive processor that decide where to move the eyes and when to respond target-present or target-absent. A manual motor component represents the time for the manual response.

Of special interest in the present work, EPIC does not incorporate a covert selective attention mechanism. That is, while historically attention is clearly associated with overt behaviors such as eye movements, the concept of covert attention generally implies some kind of top-down direct internal control of perception by cognition. Rather, in EPIC, a strategy uses the available perceptual information to decide whether a response can be made or if more information is needed, and if so, what object should be fixated to collect that information. In terms of the traditional language of attention, covert attention is an *early selection* mechanism, while EPIC has a *very late selection* approach to attention. Thus, if a model built in EPIC can account for visual search phenomena, it would show that the covert attention concept that has dominated the visual search field is not in fact necessary.

In the EPIC architecture, visual objects and their properties are formed early in vision (see Scholl, 2001). The *eye processor* component contains *acuity functions* that specify whether each visual property of each object is currently *available* as a function of the size of the object and its eccentricity from the current eye position. The currently available visual properties for each object are represented in the *sensory store*; the *perceptual processor* then encodes the properties of each object, possibly in relation to other objects, and passes the encoded representation on to the *perceptual store* where they are available to the *cognitive processor* to match the conditions of production rules which represent the cognitive strategy for performing the task. The perceptual store contains the current representation of the visual world that cognition can reason and make decisions about, including decisions about where to move the eyes by commanding the

ocular motor processor.

When the eyes move away from an object, the properties of the object persist for a short time (e.g. 200 ms) in the sensory store, and a long time (e.g. 4s) in the perceptual store. But if the object disappears completely, it and all of its properties will be removed from the perceptual store fairly quickly. Thus the representation persists for a considerable time as long as the scene is present; this is supported by studies summarized by Henderson & Castelhano (2005); memory for previously fixated objects was assessed in natural visual scenes, and retention times of at least several seconds were observed. The task strategy uses this retained information to avoid re-fixating an already examined object (see Kieras, 2011).

EPIC models for other visual search tasks are presented in Kieras (2011, 2016), Kieras & Hornof (2014), and Kieras & Marshall (2006). Constructing the model for a specific search task requires a choice of perceptual mechanisms and parameters, motor parameters, and a task strategy. These are described in the following sections.

3.2. Visual resolution

The many decades of research on vision provides some useful psychophysical results on the detectability of different perceptual properties of an object as a function of the *eccentricity* (the distance in degrees of visual angle from the center of gaze) of the object, and the *size* of the object (also measured in degrees of visual angle); if the eccentricity is increased, the size of the object must be increased to be equally discriminable; the effect is known as cortical magnification (e,g,, Virsu & Rovamo, 1979). Different properties differ in detectability in peripheral vision; for example, in peripheral vision, color is very detectable (Gordon & Abramov, 1977), but letters can be recognized only if they are very large (Anstis, 1974). Findlay and Gilchrist (2003) provide a useful overview of these results. However, the psychophysical literature does not contain a comprehensive and fully parametric set of measurements that could just be "plugged into" a model, so the relevant parameters must be estimated to fit the modeled data.

In the present model, the visual processor contains a separate *acuity function* for each property of color, orientation, and shape in which a Gaussian detection function gives the probability that the property will be detected (be available) for an object with size *s* at eccentricity *e*:

$$P(\text{detection}) = P(s > N(\mu, \sigma)); \mu = a + be, \sigma = a\ constant$$

The value μ can be interpreted as the 50% threshold for object size; its value increases linearly with eccentricity, providing a simple form of the cortical magnification effect. The value of σ governs the steepness of the ogival detection function.

The color property is used in both the CSF and COC tasks and was constrained to have the same parameter values in these tasks; orientation was used only in COC, and shape only in SHP. The *a* term was held at 0.0, *b* was estimated as 0.11 for color, 0.20 for orientation, and 0.425 for shape. σ was held at 0.5. This corresponds to observations that color is widely available, orientation less so, and detailed shape even less so. Note that the shape property is treated as a unitary property like color or orientation, but it is much less available in peripheral vision. The availability of each property is independently resampled for all objects whenever the eyes are moved. The total time for a property to appear in the perceptual store was set at 50 ms.

3.3. Perceptual storage duration

As the eyes move around, the available properties of the same object can fluctuate, and so will not be reliably available from one fixation to the next. However, as described above, the information once acquired will remain for some time in the perceptual store, forming a stable visual representation. The retention time parameter was set at 4s, the value used in Kieras (2011) for modeling a search task that required individual object fixations.

3.4. Crowding effects

Crowding refers to the phenomenon in which the perception of an object in peripheral vision is impaired if there are surrounding (*flanking*) objects that are spaced closely enough (for reviews, see Levi, 2008; Pelli & Tillman, 2008; Rosenholtz, 2016). The *critical spacing* between objects at which crowding effects appear depends on the eccentricity; in fact, the critical spacing is roughly constant at about *0.5·eccentricity* (first reported by Bouma, 1970), but the magnitude of the disruption varies with the specific features involved and how similar they are. For example, letter shapes are greatly disrupted by crowding, whereas object colors much less so.

As mentioned above, a commonly overlooked issue in typical visual search experiments is that the objects are randomly distributed in a fixed area, so set size is confounded with average object spacing. While rarely tested directly, when spacing is manipulated independently of set size, crowding appears to be the most important factor in determining RT (e.g., Wertheim, Hooge, Krikke, & Johnson, 2006). In Monte-Carlo simulations using the Wolfe et al. (2010) displays,

assuming that the eye fixates each object, the probability that a given non-fixated object is crowded by at least one flanking object increases with set size from 0.16 to 0.74. Thus, crowding effects could well play a role in this data set.

The literature on crowding effects is extensive, but the effects and mechanisms remain unclear. There is a consensus that the visual system attempts to form visual objects by integrating information over a retinal area the size of which increases with eccentricity. If more than one physical object occupies a single such integration field, the integration process will be disrupted in some way. But if the point of fixation is closer, the smaller size of the integration fields will allow the same visual objects to be correctly formed. The problem is that the empirical work has not clarified, even in simple situations, the basic rules for the integration process and the nature of crowding disruption. Results using a common psychophysical procedure suggest that the crowding disrupts the detection or discrimination of properties of the crowded object.

But a popular hypothesis is that the existence of the crowded object is still detected, and its basic perceptual features also are still detected, but the disrupted integration process associates those features with the wrong object, such as a flanking object, and vice-versa — the features are essentially *scrambled* between the objects that crowd each other. Strong evidence for this hypothesis is sparse (e.g., Põder & Wagemans, 2007). However, more than other possible mechanisms, the feature scrambling concept has very interesting implications for errors in visual search and the role played by the strategy in mitigating these errors, and so was chosen to explore in this work.

Accordingly, a simple architectural mechanism for crowding was added to the visual perceptual processor to randomly scramble the properties between objects that are within the critical spacing of each other; an unavailable property is represented as a "blank" property and participates in this scrambling. As noted above, Shape is treated as a unitary property. The scrambling process is applied when the display appears and after every eye movement. If an object has no crowders, and all of its properties are available, these properties then become "sticky" in the visual perceptual store and are not scrambled in the future. To parameterize the magnitude of the crowding effect, scrambling for each property type and each object is performed with a certain *scrambling probability*. The estimated values for the scrambling probability parameter are 0.025 for Color and Orientation, consistent with the dissimilarity of their two values, and 0.1 for Shape, which has two highly similar values.

As the scrambling mechanism is applied repeatedly when the eyes move during a trial, an unavailable property might get replaced by some other object's property, meaning that a target object might get a non-target property, becoming an *illusory distractor*, or a non-target object might get a target property (if it was available) and thus become an *illusory target*. The likelihood of these events depends on what features are on the display in positive and negative trials, and whether more than one property has to be co-located to comprise a target. This means that crowding effects play a different role in the different search task conditions and strategies, as discussed below.

3.5. Saccade timing and accuracy

The time in ms to execute a saccade of length *e* in degrees is provided by Carpenter's (1988) estimate as:

$$saccade\ duration = 21 + 2.2e$$

A variety of studies (e.g., Abrams, Meyer, & Kornblum, 1989) have shown that saccades tend to fall short of the actual fixation target, and the standard deviation of the saccade distance tends to be proportional to the distance. In the architecture, the oculomotor processor samples the length of a saccade to an object at eccentricity *e* from a Gaussian distribution:

$$saccade\ length = N(\mu, \sigma)); \mu = g \cdot e, \sigma = s \cdot \mu$$

Typical empirical values for *g* (gain) range from 0.85-0.95, and *s* (spread) is typically around 10%. In the current model, the parameters were held constant at the values suggested by Harris (1995) as optimal, namely g=0.95, s=10%. In addition, the angular direction of the saccade is also noisy, but due to the very few available studies (e.g., van Opstal & van Gisbergen, 1989) a rough estimate was used: the angle of the saccade is perturbed by a sample from $N(0, \sigma_A)$, where σ_A = 1°. Thus large eye movements often miss the object to be fixated, reducing the chances that its properties will be accurately detected.

3.6. Task strategies

EPIC's cognitive processor applies production rules in parallel in a 50 ms cycle. The production rules in the model are a variation of a basic strategy used in previous EPIC visual search models; this *Basic* search strategy is shown as pseudocode in Fig. 4. Once the display objects appear on the screen, after a delay time held constant at 100 ms, the strategy production rules alternate between a *nomination* phase, in which rules nominate objects (possibly in

peripheral vision) that are either the target or are *possible targets* because a relevant property either matches or is unknown, and a *choice* phase, in which an action is chosen. If a target object has been nominated, a target-present response is made via a manual motor processor keystroke command. If there are no nominations, then a target-absent response is made. But if there are only possible-target nominations, the eyes are moved to the closest such object. Once the eye movement is complete, the nomination phase starts again. Thus, over time, information about the objects accumulates until either the target object becomes known, or the known properties of all objects show that none of them could be the target. The main determinant of RT is how many eye movements are made in this process.

---------- Insert Fig. 4 about here ----------

In general, the choice of strategy has a large effect on whether the model can fit the data, and a satisfactory fit can only be obtained by choosing a combination of parameter values and a strategy. These data required a different strategy for each task condition, which is plausible since the subjects were extremely well practiced in a single task, and thus had an opportunity to optimize their performance. In this section, the different strategies necessary to fit each condition are described, and then the overall goodness of fit is presented in Fig. 1 and 2 and the Model Results section below.

Using the parameter values listed above, the Basic strategy provided a good fit to the SHP condition data, but not the other two conditions - there were no parameter values that allowed this strategy to fit these RT and ER data satisfactorily. Iterative testing of competing strategies revealed that two additional strategies were needed to fit the other task conditions.

The CSF condition RTs can be fit pretty well by the Basic strategy since the high availability of the color property means that extremely few eye movements are required even at the largest set size, but this did not account for why Miss errors were more frequent than False Alarms. Further testing showed that a good fit for both RT and ER was provided by the extremely simple *Fixed-Eye* strategy shown in Fig. 5. No eye movements are done; instead the target-present or target-absent response is chosen after a single nomination phase.

---------- Insert Fig. 5 about here ----------

Exploration of different strategies showed that the COC conjunction condition requires the somewhat complex *Time-Out & Confirm-Present* strategy shown in Fig. 6. The first option in the

choice phase is to immediately respond absent if more than a certain number of fixations, estimated at 3, have already been made. Also, if a target has been nominated, rather than immediately responding, the eye is moved to that object, and if it indeed has the target properties, then the response if made; if not, the strategy goes to the nomination phase again. What is noteworthy about this strategy is that it deals with possible errors due to illusory targets, explained more below.

The nomination and choice rules in the CSF and SHP tasks simply test for a single object property. For example, in the CSF condition, an object is nominated as the target if it has a red color, or as a possible target if it has an unknown color. In contrast, for COC, there are three possible-target nominations, and the strategy chooses one to fixate in the following descending priority order: Red color and unknown orientation, unknown color and vertical orientation, unknown color and unknown orientation.

---------- Insert Fig. 6 about here ----------

3.7. How the model makes errors

Errors have two sources under the strategies used in the model. First is a conventional idea in human performance research, that a certain number of errors stem from simple slip or "oops" errors at response execution; for example, the subject intends to respond target-absent, but at random happens to hit the target-present button instead. In the model, when the strategy calls for a response, the *opposite* response is made with an "oops" error probability. Since the False Alarm rate in the Fig. 3 ER data is very low and fairly constant across tasks and set sizes, the "oops" error probability was set at the average False Alarm rate of 1.4% for all conditions.

The second source of errors are illusory targets and illusory distractors produced by crowding scrambling. Note how the Miss error rate in Fig. 3 increases with set size and apparent task difficulty. Clearly if a Time-Out strategy terminates the trial before all the perceptual information is available, a Miss error could result. However, another reason for a Miss error is that the strategy rule that detects the absence of possible targets fires when the target is in fact present on the display. This would happen if all of the relevant perceptual information appears to be available and all of the objects appear to be distractors. This will be exactly the situation if crowding scrambling turned the target into an illusory distractor and at the same time, all of the other objects appear to be distractors.

Thus, the consequences of crowding scrambling depend on the search task and the strategy for that task. In CSF, a target-present response should be made if the target color is visible, regardless of which object it is associated with, and the wide availability of color means that it will rarely go undetected. In this case, crowding scrambling will be essentially irrelevant, and the Eyes-Fixed strategy should suffice for both low ER and very fast RTs independent of set size.

The SHP task is similar in that if the target *2* shape is detected, it doesn't matter whether it is the correct object or not. However, because the shape property is not very available, the Basic strategy is required to move the eyes possibly many times until a shape has apparently been detected for all of the objects, leading to a long RT. Also, the similarity of the *2* and *5* patterns means that scrambling will happen fairly frequently. The result is that relatively often the target will become an illusory distractor, and a Miss error will be made before all of the objects have been fixated.

The CSF and SHP tasks and their strategies have an important property in common. In a negative trial, the target perceptual property will not be available on the display, so crowding scrambling will never produce an illusory target, and the strategy will never conclude that the target is present when it is not. So, the False Alarm error rate in these conditions is just the "oops" error rate.

In contrast, for the COC task, a target is both red and vertical, but some other objects have the red target color, and some other objects have the vertical target orientation, so crowding scrambling has many opportunities to create illusory targets even on a negative trial, causing potentially many False Alarm errors. To prevent this, the strategy has to confirm that an apparent target is an actual target by fixating it before responding — this is a fundamental strategic property of the COC task compared to the CSF and SHP tasks. Subjects can learn from practice in COC that acceptable ER and reasonably fast RTs can be achieved with only a few eye movements. The result is that the Time-Out & Confirm-Present strategy provides a good fit.

3.8. Model results

The parameter values and choice of strategies described above were determined by informal iterative fitting, with the model being run a total of 100,000 trials in each task × polarity × set size condition. Fig. 2 and 3 show the predicted and observed RT and ER values, with the observed shown as solid lines and points, and predicted as dashed lines and open points. Over all

three search task conditions, the fit for RT is very good, with r^2=0.98, average absolute relative error of 6%. The fit for ER is good in terms of r^2=0.95, but the average absolute relative error is 21% even though almost all of the predicted points fall within the confidence intervals. In addition to the fit not being quite as good as for RT, the combination of the ER data being relatively noisy and having small absolute magnitudes could have inflated the relative error metric. But on the whole, the model provides a good account of the effects of set size and search task in both the RT and ER data.

Individual differences. Some of the discrepancies in the fits can be explained by individual differences in strategy selection and parameter values, which due to space limitations, can only be summarized here. In each task condition, there are about three subgroups of individual subjects, which can be identified with a simple cluster analysis based on RT and ER metrics. The mean performance of some of these subgroups has markedly different patterns of RT and ER effects compared to the overall means for that condition shown in Fig. 2 and 3. For example, in the SHP condition, a third of the subjects have RT curves that are strongly negatively accelerated with set size and very high ERs (as much as 23%) at the largest set sizes. As another example, in the COC task, one subgroup has steeper RT slopes than the mean data, and another subgroup has much flatter RT slopes than the mean data. These performance differences imply both parametric and strategy differences between individual subjects, meaning that fitting a model to the overall average data is problematic both in the quality of the fit and whether that model represents what those subjects actually do. The model described in this paper can be fit to the mean RT and ER for each subgroup in a straightforward way by choosing one of the strategies described above and making modest modifications of parameter values. So the model itself can provide a satisfactory account of the overall mean performance, and appears to works well to explain performance at the individual level.

4. Conclusions

The model built in the EPIC computational cognitive architecture provides an accurate account of the RT and ER data using a surprisingly simple combination of architectural components and task strategy, which together implement an active vision approach to visual search. Notably, there is no need for a covert selective attention mechanism, as proposed in the dominant theory, to successfully account for the effects; basic perceptual mechanisms, eye movements, and

strategies that meet the task demands are all that is required. The addition of a crowding mechanism to the visual processor not only provides a novel approach to accounting for errors in visual search, but also motivates a more thorough account of how the strategy is affected by the task characteristics. In particular, as originally proposed by Triesman and Gelade (1980) conjunctive search is indeed different from single-feature search, but rather than attentional binding and a special "pop out" mechanism, the difference is due to how visual crowding produces ambiguities in the perceived objects that require a different strategy for eye movements. These results demonstrate that the covert attention theory can be replaced by quantitative computational models whose architectural structure can take advantage of vision science much more thoroughly, and whose explicit representation of task strategy provides much more articulated and rigorous explanations of how fundamental mechanisms are deployed in these tasks.

The presence of individual differences in task strategies, shown by substantial differences in the patterns of RT and ER, was briefly summarized above. Most researchers assume that ample amounts of practice will result in stable performance because subjects seek to optimize their time and accuracy. In this case, they were given accuracy feedback, and surely the tedious nature of the experiment encouraged them to respond quickly. But they were not incentivized to trade time and accuracy in any particular way, so each subject picked their own tradeoff and a strategy to meet it. This can produce large variability in what should theoretically be a fundamental process, as can be seen in the individual search time slopes reported in Wolfe et al. (1989), and the clusters of subjects discussed above. In contrast, imposing an explicit payoff function can induce more efficient strategies which show more consistent patterns of performance even in seemingly simple perceptual tasks, as powerfully demonstrated by Thompson, Iyer, Simpson, Wakefield, Kieras, & Brungart (2015). Certainly, future visual search studies should use this technique to remove extraneous variability in performance and thereby facilitate modeling and theory-building.

But in the meantime, identifying subgroups of subjects in existing data is a promising approach to dealing with individual differences. Normally, modeling individual differences is impractical because the data will often be too noisy, and fitting a model to each one of many datasets is very labor-intensive. The clustering approach summarized above combines individual subjects into a small number of subgroups with statistically reliable properties, and so makes it practical to both

describe and model the individual differences more effectively. Future work could explore this approach further.

Finally, the proposed simple feature-scrambling mechanism of crowding is especially interesting because of its strong implications for the sources of errors and effective task strategies; as mentioned above, other mechanisms are possible, and could be explored by constructing other simple models, which in turn would suggest fruitful empirical manipulations. In this way, more cognitive modeling work on visual tasks could help guide future empirical work to more quickly arrive at a comprehensive understanding of the visual system.

Acknowledgements

This work was supported by the Office of Naval Research Cognitive Science Program under grant N00014-16-1-2560. Thanks are due to David Meyer, Gregory Wakefield, and Anthony Hornof for useful discussions and comments.

References

Abrams, R.A., Meyer, D.E. & Kornblum, S. (1989). Speed and accuracy of saccadic eye movements: Characteristics of impulse variability in the oculomotor systems. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3), 529-543.

Anstis, S.M. (1974). A chart demonstrating variations in acuity with retinal position. *Vision research*, 14, 589-592.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, 226, 177–78.

Carpenter, R.H.S. (1988). *Movements of the eye*s (2nd ed). London: Pion.

Carrasco, M., & Frieder, K.S. (1996). Cortical magnification neutralizes the eccentricity effect in visual search. *Vision Research,* 37, 63-82.

Carrasco, M., McLean, T. L., Katz, S. M., & Frieder, K. S. (1998). Feature asymmetries in visual search: Effects of display duration, target eccentricity, orientation and spatial frequency. *Vision Research*, 38, 347-374.

Engel, F. L. (1977). Visual conspicuity, visual search and fixation tendencies of the eye. *Vision Research* 17:95–108. doi: 10.1016/0042-6989(77)90207-3.

Findlay, J.M., & Gilchrist, I.D. (2003). *Active Vision.* Oxford: Oxford University Press.

Gordon, J., & Abramov, I. (1977). Color vision in the peripheral retina. II. Hue and saturation. *Journal of the Opthalmological Society of America*, 67(2), 202-207.

Harris, C.M. (1995). Does saccadic undershoot minimize saccadic flight-time? A Monte-Carlo study. *Vision Research*, **35**, 691-701.

Henderson, J.M. & Castelhano, M.S. (2005). Eye movements and visual memory for scenes. In G. Underwood (Ed.), *Cognitive processes in eye guidance*. New York: Oxford University Press. 213-235.

Hulleman, J. & Olivers, C.N.L. (2017). The impending demise of the item in visual search. *Behavioral & Brain Sciences*, 40(1), 1-20. doi:10.1017/S0140525X16000121, e142

Kieras, D. (2011). The persistent visual store as the locus of fixation memory in visual search tasks. *Cognitive Systems Research*, 12, 102-112.

Kieras, D.E. (2016). A summary of the EPIC Cognitive Architecture. In S. Chipman (Ed.), *The Oxford Handbook of Cognitive Science, Volume 1*. Oxford University Press. 24 pages. DOI: 10.1093/oxfordhb/9780199842193.013.003

Kieras, D.E & Hornof, A.J. (2014). Towards accurate and practical predictive models for active-vision-based visual search. In *Proceedings of CHI 2014: Human Factors in Computing Systems*. New York: ACM, Inc.

Kieras, D.E, & Marshall, S.P. (2006). Visual Availability and Fixation Memory in Modeling Visual Search using the EPIC Architecture. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 423-428.

Levi, D.M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, 48, 635-654. doi:10.1016/j.visres.2007.12.009

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review*, **104**, 3-65.

Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.

van Opstal, A.J, & van Gisbergen, J.A.M. (1989). Scatter in the metrics of saccades and properties of the collicular motor map. *Vision Research*, 29(9), 1183-1196.

Pelli, D.G., & Tillman, K.A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, 11(10), 1129-1135. doi:10.1038/nn.2187.

Põder, E., & Wagemans, J. (2007). Crowding with conjunctions of simple features. *Journal of Vision*, 7(2):23, 1–12, doi:10.1167/7.2.23.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science*, 2, 437–57. doi: 10.1146/annurev-vision-082114-035733

Scholl, B.J. (2001). Objects and attention: the state of the art. *Cognition*, 80, 1-46.

Thompson, E.R., Iyer, N., Simpson, B.D., Wakefield, G.H., Kieras, D.E., & Brungart, D.S. (2015). Enhancing listener strategies using a payoff matrix in speech-on-speech masking experiments. *Journal of the Acoustical Society of America*, 138(3), 1297-1304. doi:10.1121/1.4928395

Treisman, A.M, & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97-136.

Wertheim, A. H., Hooge, I. T. C., Krikke, K., Johnson, A. (2006). How important is lateral masking in visual search? *Experimental Brain Research*, 170, 387-402. DOI 10.1007/s00221-005-0221-9.

Wolfe, J.M. (2014). Approaches to Visual Search: Feature Integration Theory and Guided Search. In A. C. Nobre & S. Kastner (Eds.). *Oxford Handbook of Attention*. (pp.11-55). New York: Oxford University Press.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 419–433.

Wolfe, J.M., Palmer, E.M, & Horowitz, T.S. (2010). Reaction time distributions constrain models of visual search. *Vision Research*, 50, 1304-1311.

Zelinsky, G., & Sheinberg, D. (1995). Why some search tasks take longer than others: Using eye movements to redefine reaction times. In J.M. Findlay, R. Walker, & R.W. Kentridge (Eds.), *Eye movement research: Mechanisms, processes and applications*. North-Holland: Elsevier Science Publishers., 325-336.

Figure File Names and Captions

KierasFig1.eps

Fig. 1. Sample search displays produced by the model using the information in Wolfe et al. (2010). The tasks conditions, left-to-right, are color single feature (CSF), color-orientation conjunction (COC), and shape (SHP). The concentric gray circles show the simulated eye position at the initial fixation location; for scale, the inner circle has a diameter of 1°; the outer circle is 10°.

KierasFig2.eps

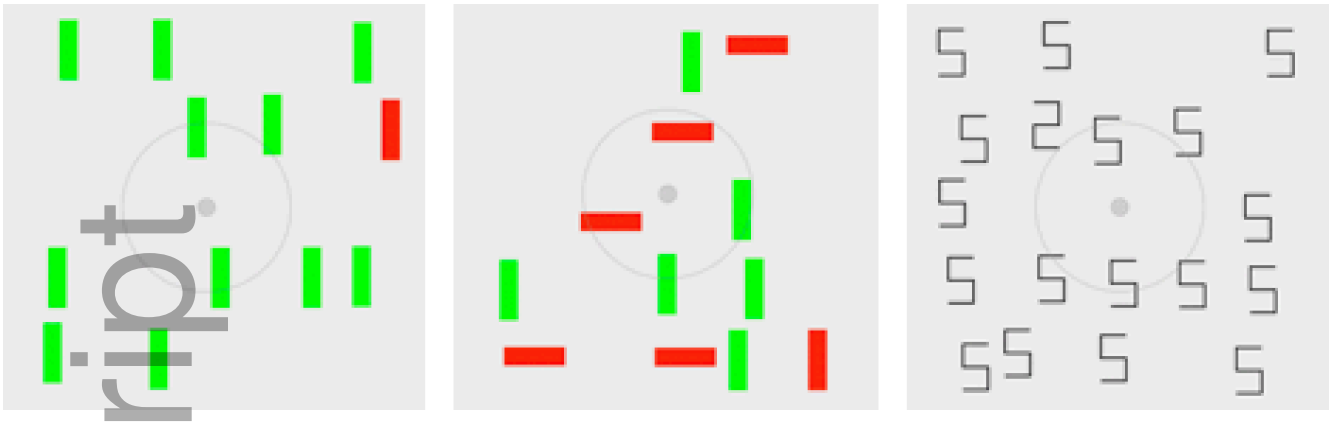
Fig. 2. Observed (solid points and lines) and predicted (open points and dotted lines) for correct trial RT in each task condition. CSF: circles, COC: triangles, SHP: squares. Positive (target-present) trials: red; negative (target-absent) trials: black.

KierasFig3.eps

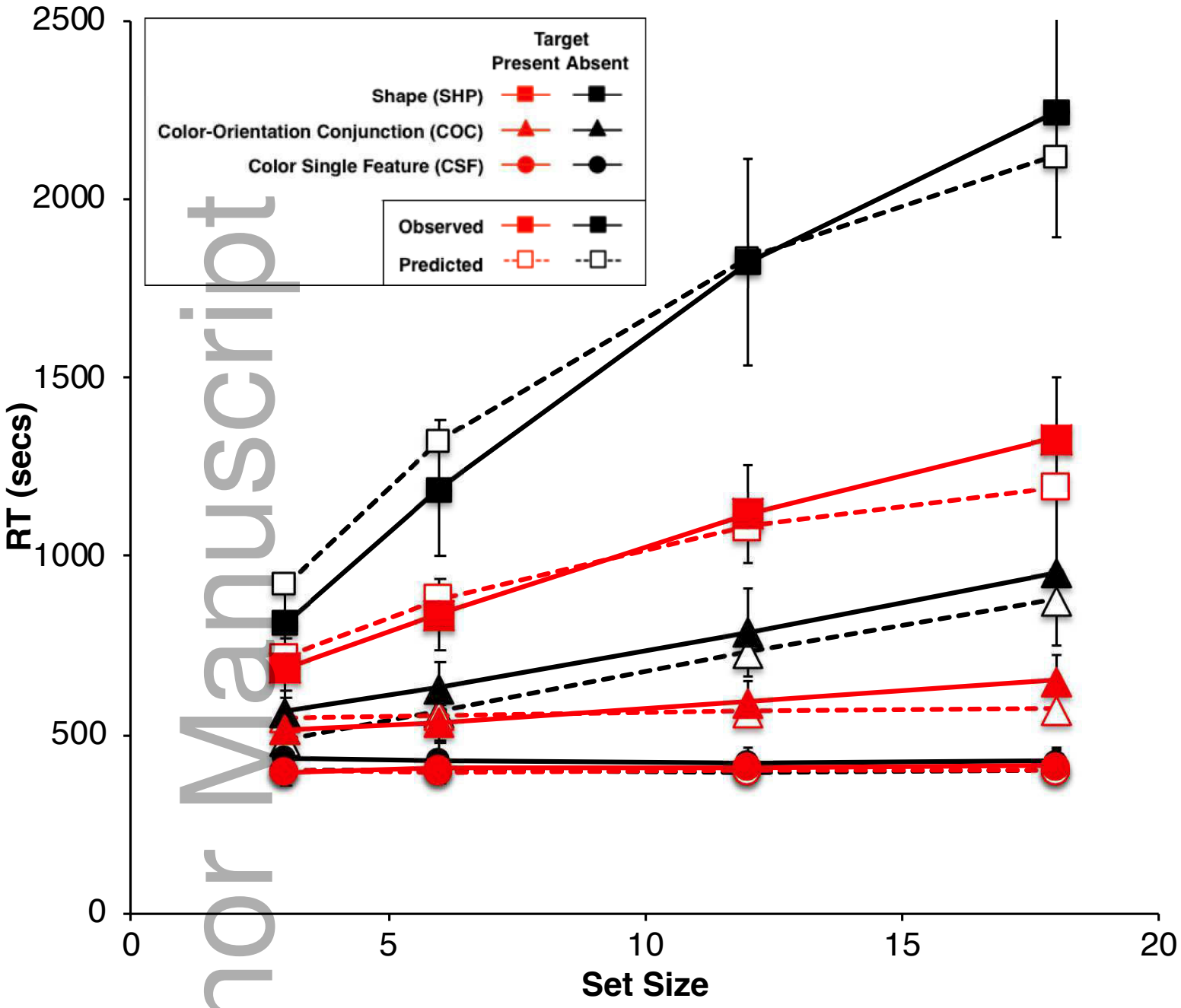
Fig. 3. Observed (solid points and lines) and predicted (open points and dotted lines) proportion of errors (error rate, ER) in each task condition. CSF: circles, COC: triangles, SHP: squares. Positive trials (Miss errors): red; negative trials: (False Alarm errors) black.

Text-only Figures and Captions

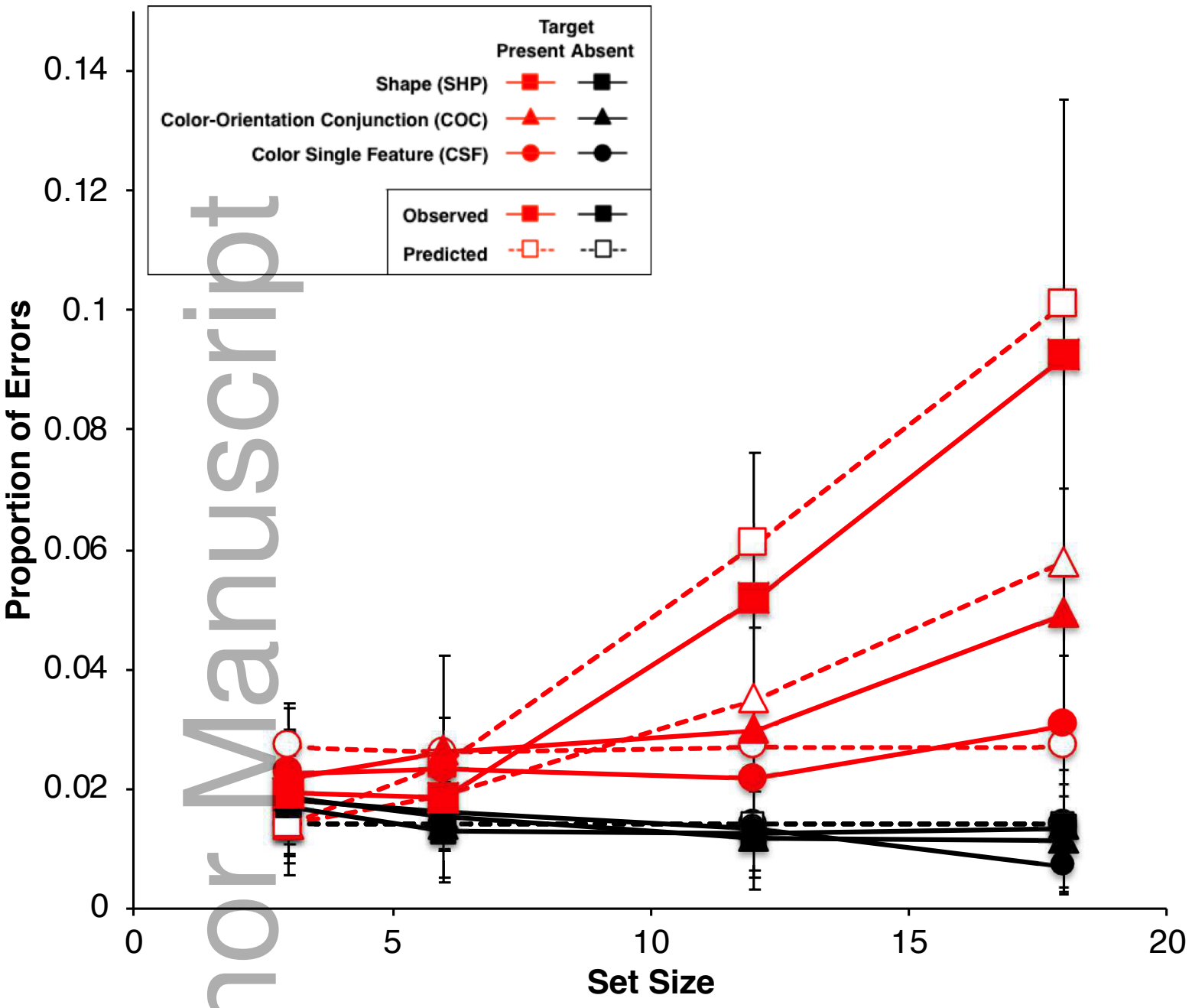
(Please maintain indentation as shown; font & box can be changed if necessary).



tops_12406_f1.eps



tops_12406_f2.eps



tops_12406_f3.eps