An Operator-Based Attentional Model of Rapid Visual Counting*

Mark Wiesmeyer
Artificial Intelligence Laboratory
The University of Michigan
1101 Beal Ave.
Ann Arbor, MI 48109-2110
wiesmeye@caen.engin.umich.edu

Abstract

In this paper we report on the use of our operator-based model of human covert visual attention [Wiesmeyer and Laird, 1990] to account for reaction times in counting tasks in which a stimulus is presented and left undisturbed until a response is made. Previous explanations have not employed an attentionally-driven model. Our model, which is based on the Model Human Processor [Card *et al.*, 1983], is an early selection model in which an attentional "zoom lens" [Eriksen and Yeh, 1985] operates under the control of cognition in order to both locate features in visual space and improve the quality of featural information delivered to short-term memory by perception. We have implemented our model and the control structures to simulate rapid counting tasks in the Soar cognitive architecture [Laird *et al.*, 1987], which has been suggested as the basis for a unified theory of cognition [Newell, 1990]. Reaction times in the counting task are explained using operator traces that correspond to sequences of deliberate acts having durations in the 50 msec range.

Background

Rapid visual counting has been a recurring focus of interest in psychology for many years, and still seems fertile ground for research—new phenomena continue to be forthcoming and no theory adequate to explain all phenomena has yet been found. Visual counting is generally agreed to be of three varieties: immediate apprehension, item-by-item counting, and estimation. Immediate apprehension, most often labeled *subitizing* [Kaufman *et al.*, 1949], is rapid, confident, error-free counting of small numbers of items where "small" is defined to be from 1-3 items [Klahr and Wallace, 1976] to 1-6 items [Kaufman *et al.*, 1949]. Item-by-item counting, which is slower and less accurate, must be employed for displays exceeding the subitizing limit, while estimation, which is fast yet quite inaccurate, is employed in time-limited situations.

Of the three modes of counting, subitizing has generated the most interest and, hence, the most controversy. Besides the question of the maximum number of items that can be subitized, opinions differ as to whether it is a serial [Klahr and Wallace, 1976, Folk *et al.*, 1988] or a parallel [Mandler and Shebo, 1982, Sagi and Julesz, 1984] process. More evidence is needed before a clear determination can be made, however with the notable exception of the Mandler and Shebo data, data supporting the parallel position were derived using blocked, limited choice, or forced choice paradigms, or were dependent on subjects' intuition of stimulus countability. These conditions may allow strategy to play an increased role and, thus, lessen the role of a counting component, which makes the subitizing process appear more parallel than serial.

We have combined a serial subitizing component and an attentionally controlled, item-by-item counting component to account for reaction times in counting tasks in which a stimulus is presented and left undisturbed until a response is made. We use data gathered by Chi [Chi and Klahr, 1975] that has been previously modeled [Klahr and Wallace, 1976]. Klahr and Wallace's model is similar in many respects to ours: it provides reaction time estimates; has a similar set of memories; has a notion of operators; and is implemented in a production system. Their model differs from ours most strongly in that it employs a different set of primitive actions; the level of description and analysis is the individual production; and it is not driven by the constraints of attention as we know them today. This last difference is a significant advantage of our model given the central importance that attention has been shown to have in visual tasks. Further, since our analysis is at the level of operators, we avoid giving too much credence to implementational details, such as the number of productions that are used to represent a process. Operators correspond to deliberate acts which are posited to be independent of implementation.

*This research was sponsored by grant NCC2-517 from NASA Ames.

Methodology

We would like our model of covert visual attention to be capable of predicting *average reaction time behavior* in a wide range of visual tasks. Currently, we limit application of our model to tasks requiring only covert visual attention in two dimensions (hereafter simply "visual attention"); however, in future work we may expand our model to include eye and head movements. Our model is task-independent and implemented on a computer. Task independence, as we think of it, means that—with the exception of task-specific control—operators, memories and the information flow assumed by our model are unchanging across the range of tasks it seeks to cover. Implementation is a check of sufficiency, allows more complex models, and may facilitate integration with other models of behavior, since demonstrably sufficient models can better serve as foundations for more complicated cognitive models.

The level of detail (grain size) at which cognitive models are described determines the sort of explanations or predictions that they can make. The grain size of information in our model is the individual visual feature with accompanying descriptive information, and our model offers explanations of how featural information arrives from perception and is transformed as it travels through short term memories[1]. The grain size of actions in our model is the deliberate act. We define a deliberate act to be an action whose selection may differ according to the current task, given the same stimuli. Thus, deliberate acts are active as opposed to passive, top-down as opposed to bottom-up, and can be expected to be sensitive to knowledge and experience. We model deliberate acts as the selection and application of operators. Figure 4 shows a sequence of operators (actually, the major result of this paper), which we call an "operator trace." Each operator application has an associated duration and the total amount of time that an operator requires to accomplish an action may be increased by interactions with perceptual or motor subsystems. Total reaction time (as in the Model Human Processor) is determined by summing the durations of constituent operator applications and their perceptual and motor dependencies.

We verify the utility of the operators that we use in our model through *coverage*. Coverage in terms of our work means testing our model on a large number of tasks, while only allowing small variations for task-specific control to occur. To date we have applied our model to precuing tasks [Colegate *et al.*, 1973], search tasks [Treisman and Gelade, 1980], tasks that produce illusory conjunctions [Treisman and Schmidt, 1982], as well as visual decay [Sperling, 1960] and crowding experiments [LaBerge and Brown, 1989]. The application of our model to these experiments is described in [Wiesmeyer, 1991].

A critical part of our modeling, once a sensible operator trace for a task has been found, is to determine operator application times that make the model fit observed behavior as well as possible. We do this by systematically adjusting operator application times to minimize the average error of the model's prediction with respect to the experimental data. Times for operator creation, shifting attention, perception, and motor processes are kept constant because determining the best fit for operator traces is an underconstrained problem—unless some times are held constant, there will be many possible solutions. Holding these particular times constant allows us to see if all operators and application times in the model are "in the ball park." Reasonable times are defined in terms of the nominal times that are specified in the Model Human Processor. This methodology may seem ad hoc, but in fact it is consonant with the goal of coverage, since we seek operators and application times that are suitable across the full range of tasks that the model has been applied to. If an operator or application time is not generally suitable, then it must be rejected.

The Model Human Processor and Soar

The Model Human Processor (MHP) is a cognitive model that has been used successfully in Human Computer Interaction (HCI) research for estimating reaction time performance [Card *et al.*, 1983]. Figure 1 shows a schematic of the MHP that omits (for economy of space) all sensory modalities except vision and includes some of our extensions for visual attention. The MHP splits the human system up into three subsystems: Perception, Cognition, and Motor. Each subsystem operates semi-autonomously and has its own domain of specialization and set of performance characteristics. Perception is composed of the lower-level processes of each of the sensory modalities. Each modality within Perception delivers information to Working Memory (essentially the same as "short-term memory") independently and at a particular rate (i.e., cycle time). Thus, the perceptual cycle time that we are interested in for our model is the time required for a stimulus to travel from the retina to Working Memory. Cognition is composed of Working Memory, Long-term Memory, and the Cognitive Processor. Cognition functions at the level of operators, and its cycle time is the time required for an operator to be applied. Motor executes commands that Cognition deposits in Working Memory, and motor cycle time is the time required for a Working Memory change to affect an overt motor response. Cycle times are specified in terms of average values and ranges: Perception, 100 msec (50-200 msec); Cognition, 70 msec (25-170 msec); and Motor, 70 msec (30-100 msec). Tasks modeled using the MHP are cast as sequences of operators applications. We have found the MHP to be a good conceptual and the-

[1] Features are iconic representations of both shape and color stimuli in our model; however, since the rapid counting task does not make use of color information, details about color are omitted in this paper.

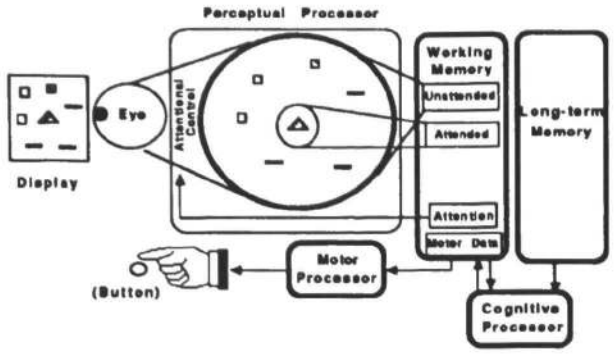

Figure 1: The Extended Model Human Processor

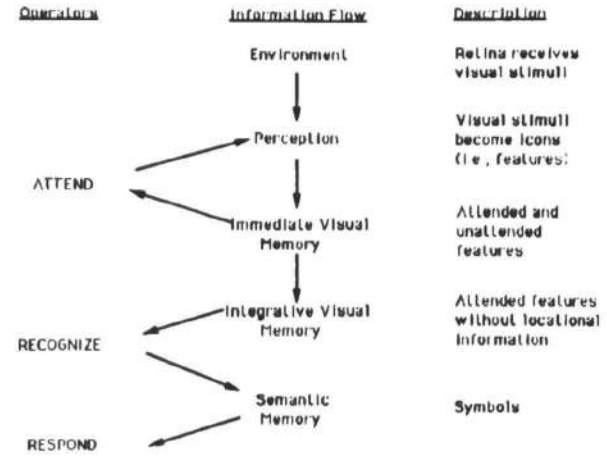

Figure 2: Information flow and operators in the Extended Model Human Processor

oretical framework to start from, since it is operator-based, already has a well-established and proven set of performance parameters, and is structurally similar to Soar (which makes implementation relatively easy).

Soar also has Perception, Cognition, and Motor subsystems[2]. Cognition is composed of a Working Memory and a Long-term Memory, which is a parallel production system. Instead of deliberation through conflict resolution at the level of productions, deliberately selected operators provide the basis of action. Thus, operators in our model map directly onto operators in Soar. Both the selection and application of operators is controlled by productions in Long-term Memory matching against Working Memory and suggesting changes to it. In Soar, specific operators (that is, instantiations of operator types) are created as soon as the data needed to instantiate them are available. In general, this means that several new operators will be created and ready for selection, during the application of the currently selected operator; however, only one operator is applied at a time. Perception is implemented as Lisp functions that transduce environmental stimuli and send input to Working Memory, while Motor is implemented as Lisp functions that receive output from Working Memory and then act on the environment. Low-level Perception occurs in parallel with the firing of productions, as does Motor.

Attentional Model

The MHP is a much more complete theory than has been presented in the previous section, however our model is only dependent upon those aspects already discussed. We capitalize on the general organization and timing of the MHP and seek to further define aspects such as timing, operators, memories, and information flow that support covert visual attention.

We use the average cycle times of both Perception (100 msec) and Motor (70 msec) in our modeling, and *usually* employ a somewhat faster value of about 50 msec for Cognition. Neither operator creation nor operator selection play any part in the MHP, its emphasis being wholly on operator application. The concepts of creation and selection in our model are derived from Soar which has been proposed as a unified theory of cognition [Newell, 1990]. Extending both the MHP and Soar, we propose that operator creation is independent of operator application and requires time. However, since next operator creation most often takes place during current operator application, these functions often overlap in a pipelined manner. Thus, creation times are often eliminated from the operator trace timing calculations. Operator selection time is assumed to be minimal and not to affect timing.

Visual attention is controlled by Cognition, as shown in Figure 1. A single ovoid region of attentional focus separates the visual field into attended and unattended stimuli. Features deposited by Perception in Working Memory may have both identity and locational information [Mishkin and Appenzeller, 1987]. Features derived from attended stimuli are guaranteed to have good identity and locational information, while those from unattended stimuli are likely to have poor identity and locational information. Differences in the quality of attended and unattended featural information create the need for attention in both recognition tasks (as in Treisman's "Feature Integration Theory" [Treisman and Gelade, 1980]) and in the counting task described later in this paper.

We posit that application of an attention shift operator, which we call `ATTEND`, causes activity at a site in the visual cortex called V4, which has been shown to be a sort of attentional gate that splits the visual field into attended and unattended regions [Moran and Desimone, 1985]. Exact timing for activity related to V4 has not been determined experimentally, but an estimate of 50 msec for deliberately changing the gate (applying the operator) and another 50 msec for receiving new visual features in Working Memory has

[2]In addition, Soar has capabilities for planning and learning that are not needed for the tasks described in this paper.

seemed to work well in our simulations. An example of using these times appears in the first five steps of Figure 4, which shows the sequence of actions that occurs when Cognition shifts visual attention to a new stimulus. It must get features from Perception that signal that a new stimulus is present (100 msec); react to the new stimulus by creating an operator (50 msec); shift attention to that new stimulus by selecting and applying the operator (50 msec); and receive new features from Perception in Working Memory that reflect a change in visual attention (50 msec), requiring a total of 250 msec. An example of shifting attention to features already in Working Memory is shown in the same figure using times at 500 and 550 msec as markers. The total time required to shift is 100 msec, since the new information appears in Working Memory at 600 msec. Another such example occurs between 700 and 800 msec. Note that the **ATTEND** operator that is applied in these latter episodes is created at some point earlier in the trace and is not shown in order to keep the trace simple.

Figure 2 shows our model from the perspective of information flow and operators. The column marked "Description" is intended to elaborate either the process that is occurring or the type of information that is available in each stage of the "Information Flow" column. To start out the information flow, light from the environment stimulates the retina and causes Perception to deposit features in *immediate visual memory*. Both immediate visual memory and *integrative visual memory*, the next memory stage, are part of iconic (i.e., pre-symbolic) Working Memory. Immediate visual memory is composed of all currently attended and unattended features. Each unattended feature in immediate visual memory causes a **ATTEND** operator to be created, which represents the possibility of deliberately shifting attention to that feature. Attended features from immediate visual memory are automatically copied into integrative visual memory. Thus, integrative visual memory is composed of features that are currently or have previously been attended. Integrative visual memory allows iconic visual memories to linger after a shift of attention or change of stimulus without deliberate recognition as discussed below [Intraub, 1985]. Although out of the scope of this paper, the fact that locational information is missing in integrative visual memory is posited to be the cause of "illusory conjunctions" [Treisman and Schmidt, 1982].

Arrows pointing to operators indicate the memories that they are dependent upon, while arrows pointing from operators indicate where their effects are felt. Operators fall into two classes: *visual operators*, such as **ATTEND**, **RECOGNIZE**, and **COUNT**, and *semantic operators*, such as **RESPOND**. It is assumed that all operators (except **RESPOND**) function identically in the tasks in which they apply, except for minimal timing variations to account for individual differences. Visual operators key off of iconic information to either affect future iconic input to Working Memory, though the **ATTEND** operator, or transform iconic memories into symbolic memories through, for instance, the **RECOGNIZE** operator. Once in semantic memory, symbolic stimuli may influence future operator selection, thus affecting behavior, or be reported through operators like **RESPOND**.

Chi's Experiment and Simulation

Twelve adults observed random dot patterns of from one to ten dots displayed on a standard video monitor that was controlled by a computer. They responded by saying the number of dots that they saw. A voice-activated relay allowed the computer to measure latencies to the nearest msec. Latencies were measured as the amount of time from when the dot pattern first appeared on the display until the relay was activated.

Subjects were told to determine how may dots were present in each trial as quickly and accurately as possible. At the start of each trial the word "**READY**" was displayed in the center of the monitor. Subjects fixated on the central "A" and pressed a button when they felt ready for the test stimulus to appear. After 1.5 sec, the stimulus appeared and the subject responded vocally and response time was recorded by the computer. Immediately afterwards (this part of the trial was not timed), the word "**ENTER #**" appeared on the monitor and the subject typed the number of dots seen. Stimuli were centered and at all times less than 1.8 degrees of visual angle, so they always fell entirely within fovea. Thus, eye movements during counting, although not tested for, were not likely and explanation of reaction times using a model of covert visual attention is appropriate.

Figure 3 shows a plot of the best straight line fits for Chi's results: there is a shallow slope of 46 msec with an intercept of 495 msec for the first three items, and a steeper slope of 307 msec with an intercept of -442 msec for subsequent items. Since there are two major slopes, there must be two major types of processes employed. We assume that the first process, as with most theories of rapid counting, is subitizing. In our model of subitizing, a counting operator iterates over attended items (at about 50 msec per iteration), thus generating a shallow slope. We further assume that counting a large number of items is accomplished by an initial subitizing stage followed by an item-by-item counting stage. In our model of item-by-item counting, attention shifts to each individual, uncounted item which is then counted using a counting operator. Although there is no definitive evidence from the literature, we assume that at least part of the need for individual shifts of attention is due to the inaccuracy of the shifting process. Inaccurate shifts of attention might result in items being counted more than once or not at all.

We present two attempts at fitting Chi's data that have been implemented in Soar. The initial attempt employs a single family of operator traces with default operator application times and shows that these traces

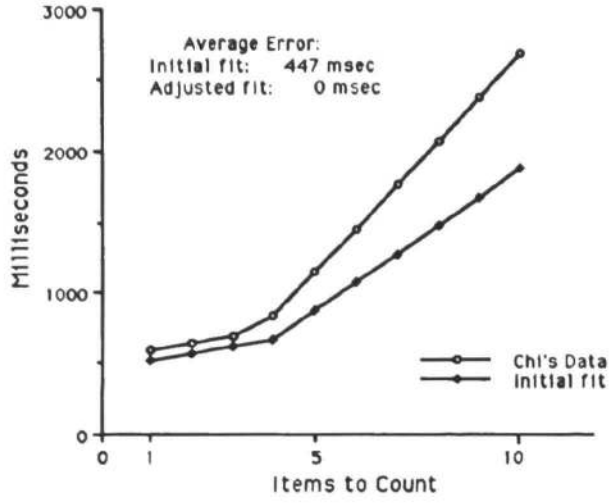

Figure 3: Comparison of Chi's data to modeling attempts

Initial fit	Operator Event	Event/Comment
0 msec	None	Stimulus at retina
0+ msec	None	Stimulus in P
100 msec	ATTEND(s) created	Stimulus in WM
150 msec	ATTEND applies	Shift to stimulus group
200 msec	ATTEND completed	Attention message at V4
250 msec	COUNT (4 created)	Stimulus attended in WM
300 msec	COUNT applies	Current count is one
350 msec	COUNT applies	Current count is two
400 msec	COUNT applies	Current count is three
450 msec	COUNT applies	Current count is four
500 msec	ATTEND applies	Shift to singleton
550 msec	ATTEND completed	Attention message at V4
600 msec	COUNT* created	Stimulus attended in WM
650 msec	COUNT* applies	Count is five
700 msec	ATTEND applies	Shift to singleton
750 msec	ATTEND completed	Attention message at V4
800 msec	COUNT* created	Stimulus attended in WM
850 msec	COUNT* applies	Count is six
900 msec	RESPOND applies	
950 msec	RESPOND completed	Motor sequence begins
1020 msec		Announce "six"

Figure 4: Operator trace for rapid counting of six items

can generate a reaction time profile that is similar to the experimental data. (A family of operator traces is required, because counting different numbers of items requires operator traces with different numbers of operator applications.) The second attempt combines the weighted predictions of two *different* families of operator traces, which are similar to those used in the initial attempt, but employ systematically adjusted operator application times.

In the initial attempt to fit to the data, we assume that counting operators for both subitizing and item-by-item counting are identical. Further, we place the subitizing limit at four, rather than the three items implied by the best straight line fits for Chi's data. This is to limit our error, since the first really big skip in Chi's times that needs to be accounted for by a shift of attention (part of the item-by-item stage) occurs at five items. Figure 4 shows an operator trace for counting six items that incorporates subitizing between 250 and 500 msec and item-by-item counting between 500 and 900 msec. (Please read **COUNT*** as **COUNT** for this initial attempt.) Due to space limitations, it was impractical to present the complete family of operator traces for counting different numbers of items. However, subitizing traces can be produced from Figure 4 by removing the item-by-item stage and traces for larger numbers numbers of items can be created by adding shifts of attention with counting operations. Figure 3 shows a plot of reaction times predicted by such traces for this initial attempt.

The initial attempt does not fit Chi's data well for two reasons. First, if deliberate acts employed vary between identical task instances, a single family of operator traces is usually insufficient to predict reaction times. The initial fit has a single discontinuity at four items, while Chi's best fit data has discontinuities at both three and four items. Thus, Chi's data is likely the result of sometimes subitizing three items and sometimes four when there are four or more items to be counted. To account for the intermediate slope between three and four items, the reaction time predictions of a family of operator traces employing three **COUNT** operators in the subitizing stage must be averaged, weighted by frequency of occurance, with the predictions of a family of operator traces employing four **COUNT** operators in the subitizing stage. Second, a new operator, **COUNT***, is required because the item-by-item slope is about 100 msec too shallow. Since the 50 msec operator application time of **COUNT** creates just about the right slope for subitizing we would like to keep it. In order to tailor the model to the experimental data, we derived equations for reaction times of both subitizing and item-by-item counting based on new families of operator traces that use the new **COUNT*** operator. (The traces again are not shown, but can easily be derived from Figure 4.)

$$T_{Subitizing} = T_{InitialShift} + T_{CreateCOUNT} + i * T_{ApplyCOUNT} + T_{ApplyRESPOND} + T_{Motor}$$

$$T_{Item-by-item} = T_{Subitize} + (i - NumberSubitized) * (T_{NewShift} + T_{CreateCOUNT*} + T_{ApplyCOUNT*})$$

where i is the number of items to be counted, $T_{InitialShift}$ is 250 msec, $T_{CreateCOUNT}$ and $T_{CreateCOUNT*}$ are both 50 msec, $T_{NewShift}$ is 100 msec, T_{Motor} is 70 msec, and $T_{Subitize}$ is the total amount of time required to subitize either three or four items and respond. In order to get the best fit to Chi's data, we systematically altered $T_{ApplyCOUNT}$, $T_{ApplyCOUNT*}$, and $T_{ApplyRESPOND}$ in the equations and the weights by which three item and four item subitizing versions of the equations were averaged until the minimum average error was found. Times for operator creation, shifts of attention, Perception, and Motor were kept constant.

Operator application times found for the best fit: $T_{ApplyCOUNT}$, 46 msec; $T_{ApplyCOUNT*}$, 157 msec; and

$T_{ApplyRESPOND}$, 125 msec. All of these times are within the MHP nominal ranges and significantly decrease the average error of the model (from 446 msec to 0 msec per item). This fit requires that four item subitizing occurs six out of ten times (59%) and three item subitizing occurs on the rest of the trials, when there are four or more items to be counted. The adjusted fit is not shown in Figure 3 because it is identical to the experimental data.

Since the best application time for COUNT* (157msec) is about three times the usual operator application time used in our modeling (50 msec), it is likely that COUNT* is a complex operator composed of simple operators. It would be too speculative to guess exactly what those operators might be, but it is likely that at least one of those operators is a semantic operator. Some evidence that COUNT* has a semantic component derives from the fact that 157 msec is very close to the silent counting rate, which has been found to be 167 msec [Landauer, 1962]. On the other hand, COUNT is clearly a simple visual operator since its application time (46 msec) is too short for anything other than a single iconic to semantic transformation to take place.

Klahr and Wallace (1976) also used Landauer's results in the analysis of their model and had similar conclusions about the memories used in subitizing and item-by-item counting. However, their timing explanations were not as accurate as ours and did not predict a ratio of three item to four item subitizing.

Acknowledgements

I would like to thank Tony Simon for discussions that piqued my interest in modeling visual counting, and John Laird and Allen Newell for much of the guidance and inspiration that has motivated my model of visual attention.

References

[Card *et al.*, 1983] S.K. Card, T.P. Moran, and A. Newell. *The Psychology of Human-Computer Interaction.* Erlbaum, Hillsdale, NJ, 1983.

[Chi and Klahr, 1975] M.T.H. Chi and D. Klahr. Span and rate of apprehension in children and adults. *Journal of Experimental Child Psychology*, 19:434–439, 1975.

[Colegate *et al.*, 1973] R. Colegate, J.E. Hoffman, and C.W. Eriksen. Selective encoding from multielement visual displays. *Perception and Psychophysics*, 14:217–224, 1973.

[Eriksen and Yeh, 1985] C.W. Eriksen and Y.Y. Yeh. Allocation of attention in the visual field. *Journal of Experimental Psychology*, 11:583–597, 1985.

[Folk *et al.*, 1988] C.L. Folk, H. Egeth, and H. Kwak. Subitizing: Direct apprehension or serial processing? *Perception and Psychophysics*, 44:313–320, 1988.

[Intraub, 1985] H. Intraub. Visual dissociation: An illusory conjunction of pictures and forms. *Journal of Experimental Psychology: Human Perception and Performance*, 11:431–442, 1985.

[Kaufman *et al.*, 1949] E.L. Kaufman, M.W. Lord, T.W. Reese, and J. Volkmann. The discrimination of visual number. *American Journal of Psychology*, 62:498–525, 1949.

[Klahr and Wallace, 1976] D. Klahr and J.G. Wallace. *Cognitive Development: An Information-processing View.* Erlbaum, Hillsdale, NJ, 1976.

[LaBerge and Brown, 1989] D. LaBerge and V. Brown. Theory of attentional operations in shape identification. *Psychological Review*, 96:101–124, 1989.

[Laird *et al.*, 1987] J. E. Laird, A. Newell, and P. S. Rosenbloom. Soar: An architecture for general intelligence. *Artificial Intelligence*, 33, 1987.

[Landauer, 1962] T.K. Landauer. Rate of implicit speech. *Perception and Psychophysics*, 15:646–650, 1962.

[Mandler and Shebo, 1982] G. Mandler and B.J. Shebo. Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, 111:1–22, 1982.

[Mishkin and Appenzeller, 1987] M. Mishkin and T. Appenzeller. The anatomy of memory. *Scientific American*, June:80–89, 1987.

[Moran and Desimone, 1985] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784, 1985.

[Newell, 1990] A. Newell. *Unified Theories of Cognition.* Harvard University Press, Cambridge, MA, 1990.

[Sagi and Julesz, 1984] D. Sagi and B. Julesz. Detection versus descrimination of visual orientation. *Perception*, 13:619–628, 1984.

[Sperling, 1960] G. Sperling. The information available in brief visual presentations. *Psychological Monographs*, 74:1–29, 1960.

[Treisman and Gelade, 1980] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[Treisman and Schmidt, 1982] A. Treisman and H. Schmidt. Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14:107–141, 1982.

[Wiesmeyer and Laird, 1990] M.D. Wiesmeyer and J.E. Laird. A computer model of visual attention. In *Twelfth Annual Conference of the Cognitive Science Society, Boston*, 1990.

[Wiesmeyer, 1991] M.D. Wiesmeyer. *An Operator-based Model of Covert Visual Attention.* PhD thesis, The University of Michigan, Ann Arbor, 1991.