Steps Towards Integrated Models of Cognitive Systems: A Levels-of-Analysis Approach to Comparing Human Performance to Model Predictions in a Complex Task Environment

Michael J. Schoelles, Hansjörg Neth, Christopher W. Myers, & Wayne D. Gray
Cognitive Science Department
Rensselaer Polytechnic Institute
Troy, NY 12180 USA
[schoem, nethh, myersc, grayw]@rpi.edu

Attempts to model complex task environments can serve as benchmarks that enable us to assess the state of cognitive theory and to identify productive topics for future research. Such models must be accompanied by a thorough examination of their fit to overall performance as well as their detailed fit to the microstructure of performance. We provide an example of this approach in our Argus Prime Model of a complex simulated radar operator task that combines real-time demands on human cognitive, perceptual, and action with a dynamic decision-making task. The generally good fit of the model to overall performance is a mark of the power of contemporary cognitive theory and architectures of cognition. The multiple failures of the model to capture fine-grained details of performance mark the limits of contemporary theory and signal productive areas for future research.

Introduction

Understanding human cognition requires knowing how control of semi-independent functional modules such as visual attention, perception, movement, and memory is integrated to accomplish complex tasks. Our understanding of this integration may be furthered by simple laboratory tasks, but as this understanding advances, it must be tested in increasingly complex task environments. In this paper we provide a progress report on our ability to predict complex behavior from our current understanding of its underlying functional components.

Our *levels-of-analysis approach* is inspired by Newell's famous timescale of human activity (Newell, 1990) that divided mental life into time-based levels where the time span of each level's processes differs from those of its neighbors by an order of magnitude. For example, Newell's *operations level* emerges at about 1/3 to 3 sec (10^0 sec) while above it is the *unit task level* (3–30 sec or about 10^1 sec) and below it is the *deliberate act level* (30–300 msec or about 10^{-1} sec). Our approach is congenial to, but distinct from, Anderson's (2002) challenge to the cognitive community to show that our understanding of low level cognitive functions can lead us to manipulations that differentially influence educational outcomes; specifically, by manipulating low-level, theory-based, functional components of cognition we can span "seven orders of magnitude" to influence educational outcomes that take weeks, months, or semesters to achieve.

In contrast to Anderson's building blocks approach, we use a wide-angle lens to characterize overall model performance as well a set of zoom lenses to magnify the differences between our model and our human subjects at increasingly fine levels-of-analyses. Our current quest starts with a multi-component complex task that takes humans 12 minutes (about 10^3 sec) to perform and requires a model that accurately predicts human performance on this task. We then proceed to zoom in on multiple components of our complex task and then to zoom in on components of those components. For each component and subcomponent we derive detailed measures of human performance and ask how well our model predicts performance on those measures.

Taking snapshots as we zoom-in leaves us with a set of conflicting images. For many of our components our measures of human and model behavior match fairly well. For other components, they do not. We use the results of these matches and mismatches to direct our attention to (a) our assumptions regarding the task analysis that underlies our model, (b) the theory-based assumptions that underlie the model's components, and (c) the mechanisms that control the sequencing and interleaving of cognitive subsystems to produce behavior that is adapted to its task environment.

In the next section we describe the complex task environment that provides the behavior for our comparisons. That section is followed by a description of the actual experiment. Data from our model and our humans are then presented and examined under increasingly higher magnifications. We conclude with a discussion of the implications of our zoom lens approach for cognitive theory as well as for cognitive research.

A Complex Task Environment

Argus Prime is a complex but tractable simulated task environment (Gray, 2002) that we have used in a variety of studies (see, e.g., Gray & Schoelles, 2003; Gray, Schoelles, & Myers, 2004; Schoelles, 2002; Schoelles & Gray, 2001b). With a small matter of programming, Argus is a flexible simulation into which we have incorporated a variety of nominally related tasks.

The version of Argus Prime discussed in this paper combines our basic simulated radar-operator classification task (Schoelles & Gray, 2001a) with a preferential choice decision-making task. During the 12-min scenarios used for this study, subjects altered between performing the

Classification Task and Decision-Making Task. The Decision-Making Task presented subjects with a list of four or six targets that they had already classified and asked them to decide which of the target set had the highest threat value. When the Decision-Making Task was on the screen the targets on the radar side of the screen (see Figure 1) kept moving, but subjects were unable to access the information required to perform the Classification Task. Hence, obtaining a high score on both tasks placed some time pressure on the subject to do the Decision-Making Task quickly as well as accurately.

Classification Task For the Classification Task the subject must assess the threat value of each target in each sector of a radar screen (depicted in Figure 1). The screen represents an airborne radar console with ownship at the bottom. Arcs divide the screen into four sectors; each sector is fifty miles wide. The task is dynamic since the targets have a speed and course. A session is scenario driven; that is, the initial time of appearance, range, bearing, course, speed, and altitude of each target are read from an experimenter-generated file. The scenario can contain events that change a target's speed, course, or altitude. Current targets can fly off the screen and new targets can appear so that 18-22 targets are on the radar screen at any one time.

The subject selects (i.e., hooks) a target by moving the cursor to its icon (i.e., track number) and clicking on it. When a target has been hooked, an information window appears (this is not shown in Figure 1, but would appear at the upper-right of the display) that contains the track number of the target hooked and the current value of target attributes such as speed, bearing, altitude, and course. The subject's task is to combine these values into a total score, using an algorithm that we have taught them, and to map the total score onto a 7–point threat value scale. (This scale appears at the bottom of the information window).

Targets must be classified once for each sector that they enter. If a target leaves a sector before the subject can classify it, it is considered incorrectly classified and a score of zero is assigned. A running score that indicates percentage of targets correctly classified is shown in the upper-left of the display. For this study, each Argus Prime scenario lasted 12-min. During this period a subject had the opportunity to calculate the threat value of targets between 70 and 90 times.

Decision–Making Task (DMT) Each scenario proceeded until the subject had classified 8 targets. At this point, a Decision-Making Task presented the subject with 4 or 6 targets for which he or she had already calculated the threat value. The subject's task was to choose the target with the highest threat value.

All groups were given the track number for each of the Decision-Making Task alternatives in a *target-column* that appeared in the lower right of the display (see Figure 1). The subject's task was to determine which target had the highest threat value and select that target by clicking on its number in the target-column. The Decision-Making Task

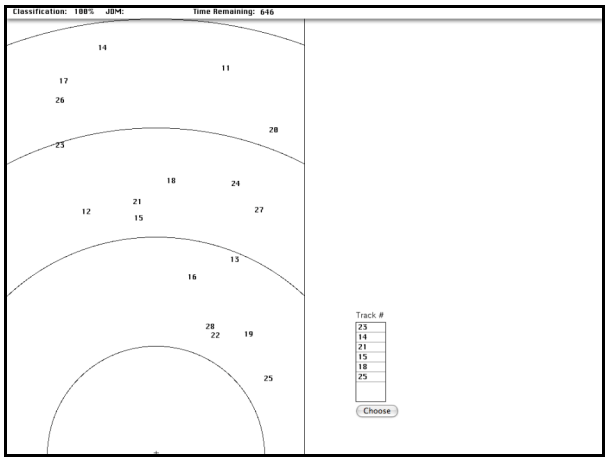


Figure 1: Argus Prime Radar Screen (left) and DMT target column (featuring 6 alternatives, to the right).

ended and the classification task resumed when the subject clicked the CHOOSE button located below the target-column.

On making a correct choice, feedback was given via a simulated explosion, the chosen aircraft was removed from the radar screen, and the overall percent score for decision-making on that scenario was increased. If the subject chose the incorrect target, the subject's overall percent score for that scenario was reduced. A running average of Decision-Making Task performance was presented to the right of the classification score. After classifying or re-classifying 8 more aircraft, another Decision-Making Task was presented. This sequence continued until the end of each scenario.

The key to performing the Decision-Making Task well is to obtain an accurate threat value for each target in the decision-making table. The threat value for a target could be accessed in one of two ways. First, as the subject had already classified the target, its threat value might be accessed by a memory retrieval. Alternatively, if the mouse cursor were moved to the target's icon in the radar window, its threat value would appear next to the target in a popup window. In considering these two alternatives, it is important to point out that although the Decision-Making Task appeared after every 8 classifications, the targets in the Decision-Making Task were not necessarily from the set that had been classified most recently. Rather, the 8 were chosen at random from the set of all previously classified targets with the constraint that the highest threat value in each Decision-Making Task set be unique to a single target (more than one target could share all but the highest threat value).

In summary, for this complex task environment there are two major subtasks: the Classification Task and Decision-Making Task. Both tasks heavily rely on interactive behavior and incorporate subtasks of visual search, memory encoding and retrieval, and decision making. Information for the Decision-Making subtask may be obtained by either memory retrieval or by moving the mouse cursor to over the target's icon in the radar window. Hence, a key feature of this version of Argus Prime is that knowledge obtained in

the course of performing one task component (the Classification Task) is directly relevant to performing the other task component (the Decision-Making Task).

Experimental Procedure[1] Subjects were randomly assigned to either the 0-Second Lockout (0-Lock) or 2-Second Lockout (2-Lock) condition. These two between-subjects conditions differed in their cost of information access. To obtain a threat value, the 0-Lock and 2-Lock groups had to locate the target on the radar screen and move the cursor to it. Similar to a *tool-tip*, the threat value then appeared next to the target. For 0-Lock, the threat value appeared as soon as the cursor moved to the target. For 2-Lock, the threat value appeared after a 2-sec delay.

Model Description

The Argus Prime Model is written in ACT-R 6.0 and is fully compliant with all changes in the ACT-R architecture. To perform the task, the model uses the same task environment software that the subjects use. The model consists of 276 productions of which 56 are specific to the Decision-Making Task and 3 are required to recognize and switch attention to the Decision-Making Task when the Classification Task is interrupted. A run of the model is the length of a scenario, which is 12 minutes. The model runs in real time in order to maintain synchronization with the dynamic Argus task environment.

The Argus Prime Model uses the standard ACT-R parameters for the activation and decay of declarative memory elements. It does not, however, learn the utilities of productions in ACT-R's procedural memory system. The parameters of the vision and motor system are also the standard ACT-R values.

The Decision-Making Task portion of the Argus Prime Model is enabled when a production detects the appearance of the Decision-Making Task table on the screen. The model moves visual attention to the table (bottom-right of Figure 1) and finds and reads the first track number (from the top-down) that it has not previously read. With equal probability it tries to find the target on the screen or remember how it classified this target. The search for the target on the screen is a random search. Search is a three-step process. First, the location of a track on the radar screen is determined. Second, attention is moved to the track to read it. Third, the number read is compared to the track number that is the target of the search. (Remember that at all times the tracks of from 18–22 targets appear on the radar side of the display, and only 4 or 6 of these are potential matches to those listed in the decision-making table.) If the number on the track matches the target of the search then the cursor is moved to the track and kept over it until the threat value appears.

Both the model and humans know that each decision-making table has one target that has the highest threat value. They also know that the highest possible threat value is 7. Hence, the model has the heuristic of choosing a target as the highest threat value in a decision-making table if it has a threat value of 7.

If the threat value is not 7 then a comparison is made as to whether the current threat value is the highest seen in this particular Decision-Making Task. The "highest-so-far" information is stored in a slot in the current goal. If the threat value just obtained is higher than the current highest, it and its associated track number overwrite the current highest threat value and track number. At this stage, if the current highest threat value is a 6, then it tries to remember if it classified any targets higher than 6 (i.e., 7). If it fails on this retrieval the model will gamble that 6 is currently the highest and choose this target even if not all the targets in the table have been checked. Otherwise the model will try to find a target in the table that it has not yet checked.

In summary, the model will process all targets in the decision-making table unless it encounters a target with a threat value of 6 or 7. If the threat value is 7, the model will immediately know that this is the highest threat value target. If the threat value is 6 the model will try to remember if it has classified any 7s. If it cannot remember classifying a target as a 7 it will conclude that the 6 is the highest possible threat value and stop processing table items.

The model has two potential ways of accessing a target's threat value. It randomly decides to either try retrieving the threat value from memory or obtaining the threat value from a search of the radar screen. The search of the radar screen is always successful, whereas memory retrieval might fail, in which case a search of the radar screen is initiated.

Human & Model In Harmony

Eleven human subjects were run in each of our two lockout conditions. Each subject completed 4 scenarios with just the classification task followed by 8 scenarios that mixed the Classification Task with the Decision-Making Task. In contrast, the model was run 11 times in each condition. Across conditions, no scenario was used by the model more than twice.

Classification Task

Although we have much to say concerning how the model compared to human performance in the Classification Task, in this short report we focus on the Decision-Making Task and limit our discussion of the Classification Task to one overall measure of performance. Hence, we can report that the model compares very well with human performance with a mean score of 58% compared to the human mean of 61%. The 95% confidence interval (CI) for the human data is 58% to 65%; hence, the model's performance falls just inside this interval.

[1] The full study used three between-subjects conditions, only two of which will be discussed here. (For more procedural details see, Gray, Schoelles, & Myers, 2004).

Decision-Making Task

Overall Accuracy At the highest level of analysis, we can compare mean score per scenario of the model and humans on the percentage of correct decisions. Both groups were overwhelmingly accurate in their decisions. Humans showed a small and non-significant difference of 94% accuracy in 0-lock and 92% for 2-lock [$F(1, 20) = 2.12$, $p = .161$] with a 95% CI of 92% to 97% for 0-Lock and 89% to 94% for 2-Lock. The Argus Prime Model was also highly accurate with a mean of 92% for 0-Lock and 93% for 2-Lock. Not only does the model do very well, but its performance falls within the confidence interval of the human data.

Another measure of overall performance is the percentage of Decision-Making Tasks for which a decision was made. Humans and models had a maximum of 60-sec for each Decision-Making Task. After 60-sec, the Decision-Making Task was scored as an error and the Classification Task resumed. This 60-sec timeout imposed some time pressure on both humans and model as neither could deliberate without limit.

Our human subjects in the two lockout conditions made a selection in 99.9% of all Decision-Making Tasks. The model chose a target candidate 100% of the time.

Number of Decision-Making Tasks Another basic measure of performance is the number of Decision-Making Tasks the subject or model received. This number depends on the speed with which the Classification Task is performed as well as the speed with which the highest threat valued target is chosen in the Decision-Making Task.

Humans show a small, but not significant difference between lockout conditions of 5.75 (0-Lock) versus 5.48 (2-Lock) DMTs per scenario [$F(1, 20) = 0.351$, $p = .56$]. There was, however, a small but significant difference of number of DMTs as a function of the number of targets, 2.86 for DMT-4 versus 2.75 for DMT-6, [$F(1, 20) = 7.87$, $p = 0.011$].

The model matched the humans well on these measures showing an average number of 6.09 DMTs in the 0-lock and 5.63 in the 2-lock conditions. These figures are within the 95% CI for this measure (5.2 to 6.3 for 0-Lock and 4.9 to 6.0 for 2-Lock). Likewise the model matches the humans on this measure when we compare across Decision-Making Task size with 2.86 for DMT-4 and 2.75 for DMT-6. Again this falls within the 95% CI for this measure of 2.6 to 3.1 for 0-Lock and 2.5 to 3.0 for 2-Lock.

We conclude from these comparisons that humans and models did not differ in the number of Decision-Making Tasks performed.

Speed A more exacting measure of performance is the speed with which model and humans made their choices. For humans the between-subjects comparison of lockout conditions (see Figure 2a) was marginally significant with a mean time per DMT of 16.45 sec for 0-Lock and 23.56 sec for 2-Lock [$F(1, 20) = 4.13$, $p = .056$]. The model mirrored human performance showing a mean of 15.59 sec for 0-Lock and 22.34 sec for 2-Lock [$F(1, 20) = 42.55$, $p < .001$]. Both of these times fall within the 95% CI (11.29 to 21.61 sec for 0-Lock and 18.40 to 28.72 sec for 2-Lock).

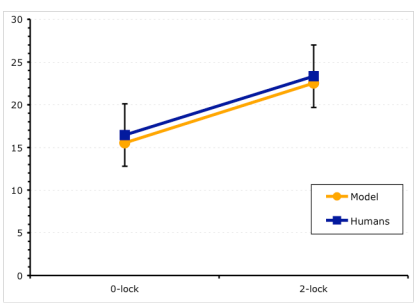
The human data also shows a significant main effect of size (see Figure 2b) with DMT-4 at 17.41 sec being (not surprisingly) faster than DMT-6 at 22.59 sec [$F(1, 20) = 15.53$, $p = .001$]. Again this difference is captured by the model [$F(1, 20) = 32.24$, $p < .001$] with mean speeds of 15.96 sec (DMT-4) and 21.97 sec (DMT-6) that fall well within the 95% CI for this measure (14.62 to 20.20 sec for DMT-4 and 17.84 to 27.34 sec for DMT-6).

Summary of Human & Model in Harmony

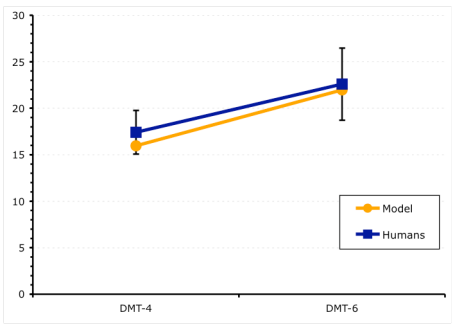
As shown by the results reported in this section, the model does a good job of replicating both overall and detailed effects found in the human data. Clearly, we have reported enough good fits to declare victory and to pat ourselves on the back for having successfully modeled human performance in a complex task environment. Although we neither dismiss nor distain our success, we feel we can learn more about human cognition by zooming in on a higher resolution of performance.

Zooming in to Reveal Differences

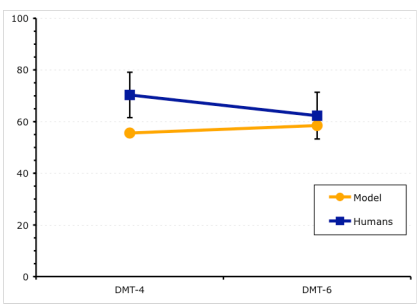
The above measures represent the standard sorts of factors on which human or model performance are typically compared. In this section we zoom in further on human and model performance in an attempt to find the points at which they begin to diverge.



(a) Choice times (sec) by lockout.



(b) Choice times (sec) by DMT-size.



(c) Percentage of the targets checked.

Figure 2: Comparisons between human and model subjects. (Error bars denote 95% confidence intervals for human data.)

Percentage of Alternatives Checked

Going deeper than the number of Decision-Making Tasks performed, we can ask of human and model what percentage of the targets in the decision-making table were checked? Humans apparently guess or rely on memory as they check only 66% of the targets by moving to and clicking on its trace on the radar screen. The model never guesses and is, apparently, able to rely on its memory as across both DMT-sizes it only checks 57% of the table targets. Interestingly, the model and target diverge in that the model checks about the same percentage of targets for DMT-4 as DMT-6 (see Figure 2c) whereas humans check a significantly larger percentage of targets in the DMT-4 condition than in the DMT-6 condition [70% vs. 62%, respectively, $F(1, 20) = 18.46$, $p = .000$; 95% CI: 60% to 82% for DMT-4 and 50% to 74% for DMT-6].

Nonetheless, the fact that the model only checks 57% of all targets and achieves accuracy levels above 90% means that we at least partially succeeded in our objective of creating a model that not only successfully performs the task, but does so by using potentially fallible memory retrievals.

Accuracy

As mentioned above, the model captures the overall effect of level of accuracy of the humans as well as the lack of difference in accuracy between lockout conditions. However, zooming in we find that the model mismatches human behavior when we look at the effect of task size. Whereas humans perform significantly better on DMT-4s than on DMT-6s [96% vs. 90%, respectively, $F(1, 20) = 15.1$, $MSE = 21.6$, $p = .001$] the model shows a slight difference in the opposite direction 92% vs. 94%. In addition, on this measure the model's result fall outside of the 95% CIs for human data (93% to 98% for 0-Lock and 88% to 92% for 2-Lock).

A close examination of the human data revealed that for one-third of the errors humans chose a target with a lower threat value when one of the checked targets had a higher value. In the other two-thirds of errors humans did not check the target with the highest threat value; that is, they either satisficed before getting it or relied on an erroneous memory. In contrast, the model never forgets the highest-so-far threat value and never retrieves an erroneous memory. (It may fail to retrieve any memory, but in that case it performs an overt search and check of the radar display.) Clearly, a better handling of memory is needed to bring model performance into line with human performance at this detailed level of analysis.

Time Spent per Target

Earlier we looked at the overall speed with which decisions were made. In this section we examine the durations of sub-components. An important indicator of the methods used by operators in a Decision-Making Task is the average time per target check. We computed this time by dividing the total sum of choice times by the number of targets checked (including duplicate checks) and subtracting 2-sec from the target check times of the 2-lock group (due to the enforced wait before the threat value was displayed). This procedure makes the simplifying assumption that subjects did nothing but checking targets on a Decision-Making Task; i.e., it counts the time spent on visual search for targets as part of the time per check.

An intriguing finding is that, even with lockout time subtracted, humans spend twice as long on 2-Lock target checks (7.2 sec) as on 0-Lock checks (3.1 sec). Gray et al. (2004) interpreted this significant difference ($p < .01$) as a strategic effort to adjust to the longer memory retention requirements in this condition. Unfortunately, the model currently has average target check times of 5.8-sec regardless of condition and does thus not reflect the same behavioral pattern.

Check and Check Again

Although humans and model check a similar number of the table targets, the model differs from humans in its percentage of duplicate checks; i.e., the proportion of all checks that were to targets that had already been checked on the current Decision-Making Task.

For humans, duplicate checks make up 19.3% of all checks, whereas the proportion of duplicate checks for the model is a mere 2.2%. Humans also display a sensitivity to the costs of checking that the model does not. In the human 0-lock group, 32.9% of all checks were rechecks; that is, checks of a target that had been already checked at least once. In contrast, for the human 2-lock group, rechecks constitute only 5.8% of all checks. As the costs for checking in the 2-lock case are higher (due to the lockout) this is a functional adaptation on part of the humans. By contrast, the corresponding model data for duplicate checks are only 2.8% and 1.5% for the 0-lock and 2-lock conditions. Thus, the model generally tends to not check targets repeatedly, regardless of the checking costs.

Summary of Zooming In

Unlike the pattern in the previous section of the paper, the data reviewed in this section reveal intriguing shortcomings of the model. A consideration of these differences suggests a profound lack of knowledge on our part as to how to repair them without imperiling our impressive successes.

Summary and Conclusions

In addition to giving us a way to think about processes that emerge at different timescales of human activity (Newell, 1990), Newell also warned us that an unremitting focus on isolated components of cognition would never enable us to see how these components fit together (Newell, 1973). We argue that Newell was right and that the time to build integrated models of cognitive systems is now. In some sense, our position is neither bold nor novel as there are many examples of other researchers engaged in much the same enterprise (for a collection of examples, see Gray, in press).

However, our call is a bit different than a call to build models using an architecture of cognition. Ours is a call to build models that faithfully reflect not only the cognitive, perceptual, and motor operations of embodied cognition, but that reflect detailed and accurate models of subsystems of perception (such as visual search and audition) as well as subsystems of cognition (such as memory and attention) along with more complete models of motor movements. Instead of building complex models for complexity's sake we argue for modeling increasingly complex tasks and then examining the success and failure of the model on this complex task through an array of wide-angle and zoom lenses.

Our goal is to gain a better understanding of how the human control system orchestrates and interleaves the resources at its disposal. A failure of the model to accurately capture increasingly detailed data should be regarded not as a dead end but as an opportunity to increase our understanding of one or more components and their integration as part of the cognitive system.

We believe that our proposed approach can be fruitful in the scientific sense of leading to interesting research and productive advances in theory. Early attempts to model Argus Prime have resulted in research focused on visual search (Myers & Gray, 2005; Neth, Gray, & Myers, 2005), task switching (Altmann & Gray, 2002), stable but suboptimal performance (Fu & Gray, 2004, in press), as well as to a theory of resource allocation at the under 1,000 millisecond level of analysis (Gray, Sims, Fu, & Schoelles, 2006).

Acknowledgments

The work reported was supported by a grant from the Air Force Office of Scientific Research AFOSR #F49620-03-1-0143. Thanks to Chris R. Sims and Vladislav D. Veksler for running subjects as well as many other contributions to this project.

References

Altmann, E. M., & Gray, W. D. (2002). Forgetting to remember: The functional relationship of decay and interference. *Psychological Science, 13*(1), 27–33.

Anderson, J. R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science, 26*(1), 85–112.

Fu, W.-T., & Gray, W. D. (2004). Resolving the paradox of the active user: Stable suboptimal performance in interactive tasks. *Cognitive Science, 28*(6), 901–935.

Fu, W.-T., & Gray, W. D. (in press). Suboptimal tradeoffs in information seeking. *Cognitive Psychology*.

Gray, W. D. (2002). Simulated task environments: The role of high-fidelity simulations, scaled worlds, synthetic environments, and microworlds in basic and applied cognitive research. *Cognitive Science Quarterly, 2*(2), 205–227.

Gray, W. D. (Ed.). (in press). *Integrated models of cognitive systems*. New York: Oxford University Press.

Gray, W. D., & Schoelles, M. J. (2003). The nature and timing of interruptions in a complex, cognitive task: Empirical data and computational cognitive models. In R. Alterman & D. Kirsch (Eds.), *25th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Gray, W. D., Schoelles, M. J., & Myers, C. W. (2004). Strategy constancy amidst implementation differences: Interaction-intensive versus memory-intensive adaptations to information access in decision-making. In K. D. Forbus, D. Gentner & T. Regier (Eds.), *26th Annual Meeting of the Cognitive Science Society, CogSci2004* (pp. 482–487). Hillsdale, NJ: Lawrence Erlbaum.

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review, in press*.

Myers, C. W., & Gray, W. D. (2005). Influencing saccadic selectivity: The effect of and interplay between stimulus-driven and strategic factors on initial fixations during visual search. *Manuscript submitted for publication*.

Neth, H., Gray, W. D., & Myers, C. W. (2005). Memory models of visual search: Searching in-the-head vs. in-the-world. *Journal of Vision, 5(8)*, 417.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York: Academic Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Schoelles, M. J. (2002). *Simulating Human Users in Dynamic Environments*. George Mason University, Fairfax, VA.

Schoelles, M. J., & Gray, W. D. (2001a). Argus: A suite of tools for research in complex cognition. *Behavior Research Methods, Instruments, & Computers, 33*(2), 130–140.

Schoelles, M. J., & Gray, W. D. (2001b). Decomposing interactive behavior. In J. D. Moore & K. Stenning (Eds.), *Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 898–903). Mahwah, NJ: Lawrence Erlbaum Associates.