

Simulated Task Environments: The Role of High-Fidelity Simulations, Scaled Worlds, Synthetic Environments, and Laboratory Tasks in Basic and Applied Cognitive Research¹

Wayne D. Gray²

*Human Factors and Applied Cognitive Program,
George Mason University*

Simulated task environments provide a setting that adds controlled complexity to experimental tasks performed by human subjects in laboratory research. Researchers whose problems are mostly applied may find that their problems are easier to study in a simulated task environment than in the actual task environment. Researchers whose theories have been nurtured in the simple environments of the typical laboratory study may find that adding controlled complexity will allow them to study how the theoretical constructs they have studied in isolation interact with other constructs in a more complex task environment. In this article I define a taxonomy and three dimensions of simulated task environments. The dimensions are based on viewing simulated task environments from the perspectives of the researcher, the task, and the participants. Research on complex systems is inherently complex. It is my hope that the terms and distinctions introduced in this article will further the scientific enterprise by enabling us to spend less time explaining our paradigms and more time communicating our results.

Keywords: simulations, microworlds, experimental design, data collection, flight simulators, training

¹ Thanks to Deborah Boehm-Davis, Alex Kirlik, Leonard Adelman, and Sheryl Miller for comments on earlier versions of this article. Thanks to Erik Altmann for helping to name the correspondence dimension.

The writing of this article was supported by a grant from the Air Force Office of Scientific Research (#F49620-97-1-0353). The simulated task environments from the author's lab that are discussed in this article were developed with the support of AFOSR, Office of Naval Research (#N00014-95-1-0175), and the National Science Foundation (IRI-9618833).

² Contact Information: Correspondence concerning this article should be address to Wayne D. Gray; Human Factors and Applied Cognition Program, George Mason University, MSN 3F5, Fairfax, VA 22030, USA. E-Mail: gray@gmu.edu. 1-703-993-1357

The researcher's dilemma and a proposed solution

"In field research there is too much [complexity] to allow for any more *definite* conclusions, and in laboratory research, there is usually too little complexity to allow for any *interesting* conclusions". (Brehmer & Dörner, 1993 p. 172).

Those who study complex situations as well as those who wish to generalize their results to complex situations have often faced the dilemma so succinctly framed by Brehmer and Dörner. *Simulated task environments* are one solution to this dilemma. The term, simulated task environment, is meant to be both restrictive and inclusive. There are many types of simulations; however, the term is restricted to those that are intended as simulations of task environments. At the same time, the term includes the range of task simulations from high fidelity ones that are intended as a substitute for the real thing, all the way to simple laboratory environments that enable the performance of tasks that do not exist. The common denominator in these simulated task environments is the researcher's desire to study behavior in a task environment that is appropriate to his or her *research question of interest*.

Simulated task environments provide a continuum of choices for the researcher. Those whose problems are mostly applied may decide that the natural world is too complex to provide the experimental control and data collection facilities that they require to make advances on their problem. For these researchers, the controlled complexity inherent to the continuum of simulated task environments enables them to reduce the complexity of the natural world to better focus on the research question of interest. Those whose theories have been nurtured in simple laboratory environments may decide that adding controlled complexity will enable them to understand how their phenomenon interacts with a range of other phenomena. For example, those who study working memory may wish to understand the role that working memory plays in the types of strategies that are adopted in a decision making task.

Why a new term? Why simulated task environment? Those of us who study human behavior are a motley group. We represent a variety of disciplines. We are given or find a variety of problems. We are inspired by a variety of motivations. Some of us wish only to solve a particular problem in any way possible. Others wish to solve a problem in the context of a favorite theoretical approach. Still others care less about solving particular problems and more about using complex behavior as a vehicle to challenge and develop theory.

Even this latter group is not particularly unified. Some researchers are interested in low-level cognitive, perceptual, and motor operations the duration of which can be measured in milliseconds, others are interested in changes that take days, weeks, or even years to develop. Some only care to

study what goes on inside one person's head. Others are interested in how interactive behavior emerges from the constraints and opportunities provided by the task, the particular artifact designed to accomplish the task, and embodied cognition (Ballard, Hayhoe, Pook, & Rao, 1997). Others care to study groups, teams, or organizations rather than individuals.

In this diversity it seems that one of two outcomes is typical. In public forums we tend to spend more time explaining why we did what we did than we do telling what we found. In our eagerness to explain quickly the *why* so we can get onto the *what*, we often frame our motivation in terms that are perceived as dismissive of other approaches. Public discussion can quickly degenerate into squabbles in which much heat but little light is shed. As an alternative, we gather in small groups or write for specialized journals in which our motivations can be assumed but which delimit our influence.

Research on human behavior is inherently complex. It is my hope that the term simulated task environment as modified by the other terms introduced and defined in this article will enable us to quickly and simply position our research in the space of all possible research. If we can reduce the added complexity involved in explaining why we do what we do, we can focus on communicating our findings and furthering the scientific enterprise in which we are engaged.

The next section provides a brief overview of the varieties of simulated task environments. The following section introduces and discusses the three dimensions in which I will frame my discussion of simulated task environments. The fourth section applies these dimensions to a sampling of simulated task environments. In the penultimate section I will summarize the article and conclude that there are differences among the various types of simulated task environments, and that the differences are interesting and may be important in helping us define and communicate alternative research agendas.

Varieties of simulated task environments

Simulated task environment enable laboratory research. The environment may be as complex as the most complex military flight simulator or it may be as simple as a paired-associate learning paradigm. The essence that these extremes share is that the task is not being performed for its own sake, but for the sake of a research study.

Five types of simulated task environments are discussed in this article. Any given simulated task environment may be regarded as a token of one or more of these types. However, not only are the types not mutually exclusive, but also the same token (i.e., the same simulated task environment) may reasonably be considered to be a different type depending on the research question of interest.

Hi-fidelity simulations of complex systems

Hi-fidelity simulations of complex systems attempt to mimic the complexity of the real world but in a fail-safe environment. Examples include commercial flight simulators, nuclear power plant simulators, and many simulators used by the military. Many hi-fidelity simulators are used to train teams. The teams may be as small as the flight crew in a commercial jetliner or as large as an entire U. S. Army battalion training at the National Training Center in Ft. Irwin, CA.

The fidelity of even the most complex simulation is relative to the question being asked. For example, the esprit de corps developed by troops at the National Training Center might be very different from that developed by the same troops in combat. Hence, the National Training Center might not be considered a hi-fidelity simulation of the conditions that affect esprit de corps in combat.

Hi-fidelity simulations of simple systems

Hi-fidelity simulations are not necessarily complex. Sometimes one subsystem of a more complex system may be built as a stand-alone simulation; for example, Irving, Polson, and Irving (1994) used a simulation of the flight management computer used by commercial airline pilots to evaluate part-task training. In other cases the system may be relatively simple as, for example, a simulation of a VCR or global positioning system (GPS)³.

In the case of the flight management computer, the simulation sacrificed context-of-use to allow the investigators to focus on individual interactive behavior. A further loss of fidelity is that the simulation was presented on the computer screen and required a mouse and keyboard for interaction rather than being a physical device with knobs and dials. Whether this loss of physical fidelity is important depends on the exact research questions being asked.

Scaled worlds

A scaled world (Ehret, Gray, & Kirschenbaum, 2000) focuses on a subset of the functional relationships found in a complex task environment. The scaled world seeks to preserve the functional relationships in this subset while paring away others. Multiple scaled worlds of the same task environment can be constructed that differ on which functional relationships are preserved and which are pared away. This decision as to what to preserve and what to pare away must be based on the research question of interest.

³ Our perspective here is not with the system as a whole, but with that part of the system with which our population of interest interacts. Hence, a handheld GPS device is one small part of an extremely complex global system. However, if our population of interest is outdoor enthusiasts and our goal is to study the learnability of this class of GPS device for this population, then it is fair and reasonable to regard the system of interest as simple.

Performance in a scaled world typically requires prior (extensive) experience with the target task environment. Researchers who build scaled worlds are primarily interested in generalizing their findings back to the original task.

Synthetic environments and microworlds

Although the terms synthetic environment and microworld are widely used, I have been unable to locate definitional references. Judging by extent synthetic environments and microworlds, the distinctions between the two are subtle and do not seem to be practically important. In this paper, the term synthetic environment will be preferred to microworld.

Those who use scaled worlds wish to generalize to a particular task environment. For these researchers, developing theory may not be as important as generalizing back to the original task environment. In contrast, theory is foremost for those who develop synthetic environments.

One way in which synthetic environments further theory is by enabling researchers to abstract functional relationships from one or more complex task environments and to study these functional relationships in a less complex, make-believe world. Other researchers may come to a synthetic environment with a research problem that has been the focus of years of research in a simple laboratory environment. These researchers build synthetic environments because they believe that further insight into their phenomenon can only be obtained by learning how it interacts with other phenomena.

In contrast to scaled worlds, synthetic environments may be intended to be used by people with little or no experience with any of the original task environments. Similarly, the results of research with a synthetic environment are intended to generalize to many different task environments.

Laboratory tasks and simulated task environments

Although it is tempting to create a dichotomy between simulated task environments and traditional laboratory tasks, I cannot find any simple way of doing so. Rather, there seems to be a continuum of simulated task environments from hi-fidelity simulations to simple laboratory tasks that differ on where they lie on the three dimensions discussed below.

In this paper the term *simple laboratory environment* will be used to anchor the complexity dimension of simulated task environments in those tasks that are the staple of the experimental psychology laboratory. However, it must be kept in mind that what is considered a simple task depends on the research question of interest. Hence, to those who study complex decision making, the sorts of simple laboratory tasks used by Payne, Bettman, and Johnson (1993) to study the effect of cognitive effort on the choice of decision-making strategy, may be considered simple laboratory tasks. However, from the perspective of those who study visual attention (e.g., Hoffman, 1998; Logan, 1996; Mozer & Sitton, 1998; Pylyshyn, 1998; Yantis, 1998), such simple decision-making tasks may appear to be incredibly complex. Describ-

ing the role of visual attention and its many interactions and influences during simple decision-making could be a challenging research endeavor (indeed, see for example, Lohse & Johnson, 1996.).

It seems unlikely that there are any tasks that humans can perform that cannot be teased and tortured to yield some insight into human behavior. Hence, I have to part company from Brehmer and Dörner. Ultimately, the evaluation of whether a research paradigm can yield interesting conclusions cannot be answered in isolation, but must depend on the research question being asked.

Three dimensions of simulated task environments

Simulated task environments differ from each other along many dimensions. However, rather than enumerating all possible dimensions of difference, the dimensions discussed in this paper, tractability, correspondence, and engagement (see

Figure 1) were derived from looking at simulated task environments from the perspectives of the researcher, the task, and the participant. The choice of one simulated task environment rather than another can be justified by reference to where it falls on each of these three dimensions.

The most salient dimension of simulated task environments, complexity, will not be directly addressed. In this paper I take the position that the absolute complexity of a simulated task environment is less important to the researcher than its tractability, correspondence, and engagement with respect to the research question of interest.

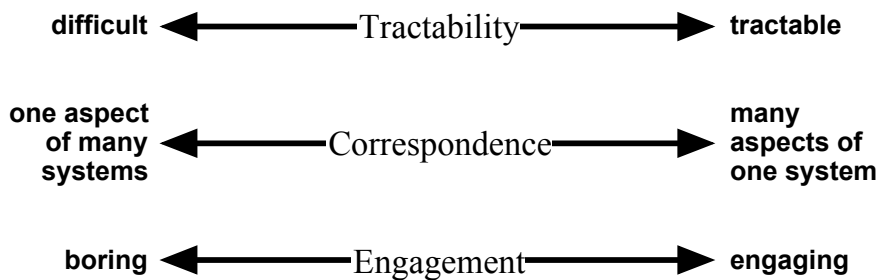


Figure 1. The three dimensions of simulated task environments represent the researcher's perspective, tractability; the task's perspective, correspondence; and the participant's perspective, engagement.

The researcher's perspective: Tractability

A thing is tractable if it is "(1) easily managed, taught, or controlled" or "(2) easily worked; malleable" (Webster's New World Dictionary of the American Language, 1960). For tractability the key issue is whether the simulated task environment allows the researcher to productively pursue the research question of interest.

Easily managed, controlled, or trained

Managed or controlled. The dynamic tension in all experimental research is between control and complexity. We wish to explain phenomena that occur in the world. Important parts of this explanation include (a) the analysis of complex phenomenon into simpler phenomena, (b) a description of each of these simpler phenomena, as well as (c) a description of how one phenomenon is influenced by another. Different parts of this explanation may require that research be conducted in different task environments.

For example, Argus Prime (discussed below as well as in Schoelles & Gray, 2000; Schoelles & Gray, 2001a) is a synthetic environment that we built to study cognitive workload in a task that captures some aspects of a radar operator's task. In analyzing the behavior we observed, we noted that successful performance required that subjects switch attention many times per minute. If attention switching had a cost, then these switch costs may make an important contribution to the cognitive workload of this task.

We reviewed the existing literature on attention switching (or serial attention) (e.g., Allport, Styles, & Hsieh, 1994; Rogers & Monsell, 1995) for hints on how much it cost to switch attention and for how we should best model this process. We were unsatisfied with the answers we found. In pursuing this issue ourselves we soon realized that Argus Prime was too complex an environment in which to isolate the costs of attention switching. Hence, we built a simple laboratory environment, that is, we used a traditional experimental psychology paradigm to study attention switching (Altmann & Gray, 1998, 1999a, 1999b, 2000a, 2000b, 2002a, 2002b). The results of these studies currently inform our models of Argus Prime (e.g., Schoelles & Gray, 2000).

Training. Simulated task environments vary widely on the extent to which subjects must be trained before they can use the environment. For example, years of pilot training and experience may be prerequisite to the use of a hi-fidelity simulation of a Boeing 777. On the other hand, Argus Prime requires an hour of training, and the VCR task (discussed below as well as in Gray, 2000; Gray & Fu, 2001) requires 5 min. In general, the more training subjects require before they can use a simulated task environment, the less tractable it is.

Easily worked; malleable

Data collection. A simulated task environment is tractable only in relationship to a given set of research issues. For data collection a tractable simulated task environment must allow the researcher to collect the right data, at the right grain size, with the right timestamp. For example, for the cognitive workload questions we wished to address using Argus Prime we needed to collect every mouse click made by the subject, every system response, every mouse movement, and every point-of-gaze. The point-of-gaze and mouse movements are sampled 60 times per sec and, along with mouse clicks, are time stamped to the nearest 16.667 msec.

Usability is a tractability issue as well. Hi-fidelity simulations inherit usability problems from the real system. However, for scaled worlds and synthetic environments usability is at least as important as it is for any other software or hardware system. Care must be taken so that performance on the functional relationships of interest is not clouded (or confounded) by usability problems (unless these are the focus of the research).

Models as users. Some sets of research questions or research approaches impose their own special constraints on the tractability dimension. For example, for some of us who build computational cognitive models, it is important that the models interact with the same system with which the human subjects interact (Ritter, Baxter, Jones, & Young, 2000). Indeed, the examples in this paper of simulated task environments from my own laboratory all have the unusual property of being capable of use by simulated human users (SHU) as well as by human users. Although this situation is not rare, it is uncommon.

For those who do computational cognitive modeling, the more frequent alternative is to simulate the simulated task environment in the modeling environment. This double simulation can be a tractable alternative as it is often difficult to get two independently written computer programs to communicate. However, the double simulation runs the risk of unintentionally omitting, what are for humans, key aspects of the simulated task environment. When the simulated task environment is independent of the model then the decisions that the modeler makes (e.g., choosing to ignore the color of objects or the distance between two objects) become more obvious to the modeler and to his or her critics. (The relationship of cognitive models to human cognition is the same as that of simulated task environments to real tasks. The goal of cognitive modeling is not to build standalone artificial intelligences, but to probe and illuminate some aspect of human cognition. Hence, all models focus on some parts of human information processing while ignoring other parts. However, because of the complexity of human cognition and the complexity of task environments, it is important to make the aspects of each that are ignored or not ignored as clear as possible.)

Occasionally commercial software is written so that it is possible to query the state of key parameters and to alter the parameters by sending the simu-

lation a data stream like that normally sent by the keyboard or mouse. This is what Schoppek did (Schoppek, Holt, Diez, & Boehm-Davis, 2001) in his ACT-FLY model that flies a 747-400 simulation (Aerowinx PS 1.3) running on a personal computer. In this instance, ACT-R running on one computer communicates over the serial port with the simulation running on another computer.

Unfortunately, relying on output and input streams provided by others puts the modeler in the position of re-purposing a software feature. Hence, unless the modeler is extraordinarily fortunate, there will be aspects of the simulation with which the models cannot interact. Indeed, for those who wish to model the interaction of cognition, perception, and action it is unheard of for a model to be able to move visual attention and the mouse around a screen and to interact with commercially written software in the same manner that a human would interact.

Ritter et al. (2000) propose one solution to this dilemma; that is, to build *cognitive modeling interface management systems* that intervene between commercial software and models. In my lab we have taken a different, more limited, approach that takes advantage of features built into ACT-R/PM (Byrne, 1999; Byrne & Anderson, 1998). ACT-R/PM provides hooks between ACT-R and task environments. By incorporating a little bit of programming discipline as we write our own simulated task environments, it is simple to develop simulations that support ACT-R/PM's hands and eyes.

For example, as ACT-R/PM "knows" the location of screen objects, when it moves the mouse from one location to another it computes the degrees of visual angle between the two objects as well as the visual angle of the target object (i.e., its size) and enters these numbers into its calculation of Fitts law (see, e.g., Card, Moran, & Newell, 1983 or; MacKenzie, 1992). This feature of ACT-R/PM, combined with other features, makes it possible to collect the exact same data from SHU as from our human users. By treating the SHU and human users as two different groups we can run all of the normal statistical analyses looking for between-group differences (e.g., see Schoelles & Gray, 2001b). In Argus Prime we are comparing overall performance measures as well as reaction time and performance on more than a dozen sub-tasks.

A feature of this approach to models as users is that the SHU and simulated task environment are separate processes. Changes to one process do not necessarily require changes in the other. Hence, with a little programming discipline, we gain a tractable means of collecting both actual and simulated human data.

Comparisons on the tractability dimension

How do the various types of simulated task environments compare on the tractability dimension (see Figure 2)? Many hi-fidelity simulations end up reinventing the complexity of the real world. This is fine if you want a safe

environment in which to measure performance under extreme conditions, but hi-fidelity simulations can be almost as difficult as the real world to study. Many hi-fidelity simulations do not enable the collection of performance (as opposed to “outcome”) data. For example, a researcher interested in pilot error might want to know exactly what information was available in an airline pilot’s environment, what information the pilot acquired, and exactly what changes the pilot made in the 60 sec before and after the critical event. Such information is no more accessible in the flight simulators of today than it is in the real world. A further problem with hi-fidelity simulations is the cost involved in running one. Besides the difficulties inherent in obtaining expert subjects, many hi-fidelity simulations require a small team of technicians to run. These expense considerations make hi-fidelity simulations of complex systems intractable for many researchers.

Synthetic environments are one solution to the problem of tractability. Indeed, the point of building a synthetic environment is to study complex phenomena in a controlled setting. In the example of Argus Prime, we are able to control the cognitive and perceptual-motor workload imposed by the environment and to collect and save all relevant behavioral data. However, Argus Prime is an intractable environment in which to study some of the simpler phenomena that contribute to cognitive workload. Pursuing the issue of serial attention required us to build a simple laboratory environment.

Tractability is more of an open issue for scaled worlds. A scaled world should be more tractable than a hi-fidelity simulation, else it is not worth the bother. Nevertheless, exactly how tractable a scaled world can be depends on that part of the real environment that has been preserved to study the functional relationships of interest.

The task’s perspective: Correspondence

Correspondence is “1. The act, fact, or state of agreeing or conforming. 2. Similarity or analogy.” (The American Heritage Dictionary of the English Language, 2000) Simulated task environments are not built as an end unto themselves, but to study phenomena that occur in the outside world. High correspondence simulated task environments simulate many aspects of one task environment. Low correspondence simulated task environments simulate one aspect of many task environments (see Figure 3).

For example, most commercial and military flight simulators are built to simulate many aspects of one task environment. An attempt is made to simulate all aspects of an actual flight deck that can be simulated in a ground-based system to the limits of current technology and some financial considerations.

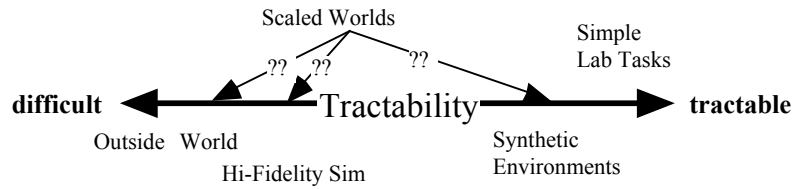


Figure 2. The tractability dimension

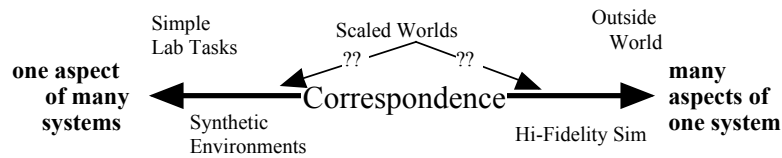


Figure 3. The correspondence dimension

For hi-fidelity simulations of complex systems the correspondence may extend not only to the device (e.g., a F-16), but the environment in which the device is embedded (e.g., communication with air traffic control, weather conditions, etc.) In contrast, hi-fidelity simulations of simple subcomponents of complex systems tend to exclude the larger context of use. For example, research on the flight management computer was conducted in the laboratory not in the airplane or in the flight simulator (Irving et al., 1994).

Scaled worlds are built to correspond to a few aspects of one task environment (see Figure 3). Scaled worlds seek to preserve certain functional relationships while paring away most everything else. In deciding what to leave and what to pare away, care must be taken to ensure that the absence of certain correspondences does not destroy the functional relationship(s) of interest.

Synthetic environments correspond to a few aspects of many tasks. For example, Moray’s pasteurization plant (Moray, Hiskes, Lee, & Muir, 1995) provides an environment in which trust in automation can be studied. The notion is that the functional relationships extracted from Moray’s studies can be applied to many diverse automated systems.

Simple laboratory environments are built to isolate phenomenon that never appear in isolation in the outside world. Compared to synthetic environments, simple laboratory environments tend to study lower level phenomena that are a ubiquitous aspect of most human tasks. For example,

compare memory retrieval with the phenomenon of trust in automation, or the phenomenon of visual attention with cognitive workload.

By combining subtle tradeoffs in generalization, scaled worlds represent an interesting intermediate point on the correspondence dimension. For example, the Ned scaled world (see below and Ehret et al., 2000) provided submarine commanders with a simulated task environment in which they could query and receive sonar information and maneuver ownship.

Ned was built to correspond to the information-processing aspects of the commander's task environment. Other aspects of the commander's task environment, most notably the mission and scenarios, were also included. However, Ned pared away aspects involving team interactions, onboard procedures, and specifics of the information displays. (As discussed below, paring away the non-cognitive factors increased tractability in several important ways.) The results of this research program are targeted to the next generation of submarines where the teams, procedures, and displays will be very different from those used today. Moreover, we expect that generalizations arising from understanding the human information-processing required for this task will be appropriate to this new type of submarine. A higher correspondence to the non-cognitive factors would have made the task of uncovering generalizable cognitive processes that much harder.

As a rule, the higher the correspondence of a simulated task environment to one system, the less the research on that simulated task environment can be generalized to other systems. (For a related discussion see DiFonzo, Hantula, & Bordia, 1998.) This rule seems validated by common practice. For example, researchers who use simulated task environments at the high end of the correspondence scale, typically wish to generalize to one system. Hence, F-16 simulators are used to train or to study pilots who fly F-16's. They would never be used to either train or study pilots who fly 777's. In contrast, researchers who, like Moray, build synthetic environments believe that the functional relationships they study can be generalized to many systems.

The participant's perspective: Engagement

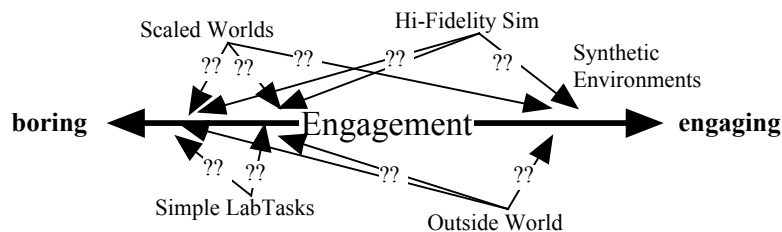


Figure 4: The engagement dimension

To engage is “INTRANSITIVE VERB: 1. To involve oneself or become occupied; participate: ‘engage in conversation’ 2. To assume an obligation; agree.” (The American Heritage Dictionary of the English Language, 2000) A simulated task environment is engaging to the degree to which it involves and occupies the participants; that is, the degree to which they agree to take it seriously. Engagement describes something about the participant’s motivation. Participants may be engaged because we are paying them money to do well. They may be engaged because they view the simulated task environment as an interesting game that they like to play. Or they may be engaged because they have deep knowledge of the real-world task, believe that it is interesting and important, and are able to fill in the blanks that are missing from the simulation.

For synthetic environments it is often, but not always, possible to build in engagement by making the task game-like. Unless a simple laboratory environment is inherently engaging, it may be impossible for researchers to increase the engagement of the task without obscuring the phenomenon they wish to study. Not all real-world tasks are especially engaging. Hence, any hi-fidelity simulation of those tasks may not be engaging either. Likewise, a scaled world that is based on a non-engaging task should probably not be more engaging than the task itself (see Figure 4).

Putting it together: How simulated task environments vary on tractability, correspondence, and engagement

To illustrate ways in which the three dimensions interact with the type of simulated task environment, I will discuss four systems, a generic flight simulator and three systems built by my laboratory for research purposes.

Example: Flight simulator

Flight simulators for commercial and military pilots correspond as highly to the real task environment as money and current technology can provide. However, as currently built most are as intractable to research as a real flight deck. It is possible to imagine changes in flight simulators that would enable the logging and time stamping of every action and every change in system state for later playback and analysis. Although such changes would improve tractability, the complexity of the flight deck with its 2-3 person crew and hundreds of instruments accessible via an eye or head movement would still make flight simulators intractable for large numbers of important research questions.

The tasks performed in these simulators can vary widely on engagement. This variation is not a characteristic of the simulation but reflects the engagement of the real-world task itself.

Example: From high fidelity simulation of a simple task to synthetic environments

We developed a simulation of a commercial VCR to pursue issues regarding performance and errors in routine interactive behavior (Gray, 2000; Gray & Fu, 2001). The simulated VCR corresponds highly to the real VCR except it does not record shows. Minor differences are that the simulated one appears on a monitor and is programmed using a mouse, whereas the real VCR is a stand-alone device that is programmed by toggling and sliding various physical switches and buttons.⁴

Not only is correspondence high, but programming the VCR is a tractable task to study. In VCR 1.0 (Gray, 2000) the current state of the system was saved to a log file at each mouse click with a 1 sec resolution. For VCR 2.0 the time stamp has a 16.667 msec resolution and point-of-gaze information can be sampled and saved to the log file 60 times a sec (Gray & Fu, 2001). Likewise, the VCR was easily learned by our subjects and was as usable as the real system. In addition, because of our concerns with computational cognitive modeling, models written in ACT-R 2.0 (Anderson, 1993) and ACT-R/PM 1.0 (Byrne, 1999) directly interact with VCR 1.0 and 2.0.

However engaging programming a real VCR might be; programming a simulated one 10 times in a row is less engaging. But, although the task itself is not inherently engaging, the demand characteristics of the experimental situation were such that our college student subjects seemed moderately engaged if not enthusiastic.

The simulated VCRs were created to study performance and errors in routine interactive behavior (Gray, 2000). To establish our basic taxonomy and approach, it was important that the simulation be based on a widely used, commercially designed product. We did not wish to be accused of building-in the patterns of performance and error behavior that we were trying to uncover.

Since the original study, the VCR has been used to study trade offs between perceptual-motor versus cognitive effort (Gray & Fu, 2001). These questions did not mandate the choice of a VCR. Rather, the questions could have been pursued using any simulated task environment that met three criteria. First, we needed a clear separation between using the task interface versus accessing information for the task. Second, we wanted a task that would not force users to keep or manipulate information in-the-head; that is, storage in memory for more than a few seconds should be an optional, not a necessary requirement of task performance. Third, the task environment had to enable us to manipulate the perceptual-motor effort involved in accessing information.

⁴ Differences along this physical dimension present minor differences for the simulated versus actual VCR in terms of the research question of interest. However, if, for example, the task being simulated were learning to ride a bicycle, then the differences between the actual bicycle and a bicycle simulation that appeared on a CRT would be considered a major difference.

Indeed, since we have begun to pursue questions regarding the trade off between perceptual-motor versus cognitive effort the simulated task environments we use have changed from hi-fidelity simulations of simple systems to synthetic environments. For one line of research we borrowed from Ballard and associates (e.g. Ballard, Hayhoe, & Pelz, 1995) to build a Blocks World synthetic environment (Fu & Gray, 2000). To pursue this work in a more complex context, we have built an interface construction kit (ICK 1.0) that allows us to directly manipulate the three cognitive engineering principles described by Gray (2000); least-effort in operating the device, least-effort in mapping prior knowledge to device knowledge, and least-effort in place-keeping. (This work is currently underway and research reports on it have not been written.)

Example: Ned – A scaled world

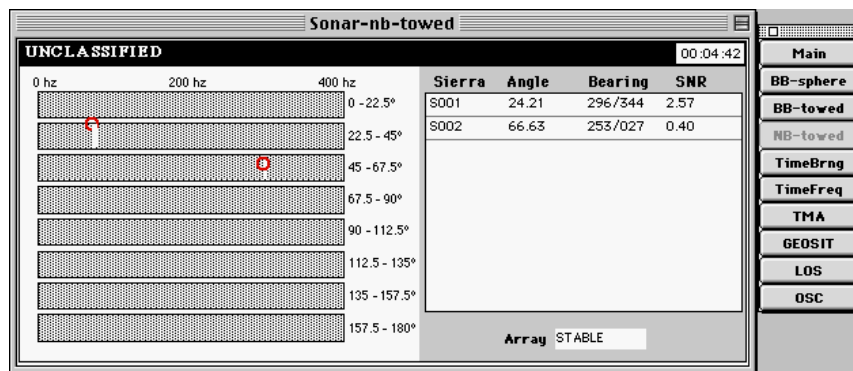


Figure 5. The sonar, narrow-banded towed display used in Ned (with the display menu shown on the right). NB-towed is one of Ned's 10 displays and one of three that include sonar sensors. The figure shows two targets (indicated by the white streaks, or waterfall, in the left display) being tracked (indicated by the circles attached to each streak). The NB-towed waterfall display indicates targets by the frequency of the sound they emit (from 0hz to 400hz at the top of the display). Target bearing from ownship is indicated by the band in which the target occurs beginning with 0-22.5° and ending with 157.5-180°. Thus, NB-towed yields an ambiguous bearing as it does not indicate which side of the ship the target is on (i.e., 0-180° or 180-360°). The table on the right indicates, for each target, its angle from ownship, its ambiguous bearing (for target S001 the bearing is either 296° or 344°), and the signal-to-noise ratio. On the bottom-right, the display indicates that the array of sonar sensors used for NB-towed is stable. To the right of NB-towed is the menu of the other displays available to the AO. Ned is generated dynamically and all information is updated every second.

Ned corresponds to many aspects of a few functional relationships of one complex, real-world task. The target audience for Ned is submarine Approach Officers. These are typically either the Executive Officer or Commanding Officer of the submarine (Kirschenbaum & Gray, 2000) (though we have used Ned with junior officers as well). The functional relationships incorporated into Ned were defined by the information-processing aspects of the Approach Officer's task environment and most other aspects of that environment have been pared away. The more we built Ned to correspond to the non-information aspects of the Approach Officers' environment, the better our analyses would have been for re-engineering the current environment, but the less able we would be to generalize our results to new submarine environments. Hence, we decreased correspondence by paring away the functional relationships provided by the Approach Officer's interactions with his crew and boat, but we maintained an information environment in which the functional relationships essential to our goals were preserved.

In deciding which correspondences to keep and which to pare away we were guided by three types of analyses. First and primarily, our decisions were based on a deep and intensive cognitive task analysis of 10 Approach Officers using a much higher fidelity simulation (Ehret et al., 2000; Gray & Kirschenbaum, 2000). Less formally, all decisions made in designing the scaled world were informed by the years of experience working with and studying submarine officers that one member of our team had. Finally, early prototypes of the scaled world were tested with former and current Approach Officers. Suggestions made by these testers were incorporated into the design. Each of the three types of analyses helped to ensure that although much of the submariners' environment was pared away, the functional relationships of interest were preserved.

For this study, being able to take the simulation to our subjects was an important tractability issue. Ned met this objective as it is written in Macintosh Common Lisp and runs well on a laptop computer. Data collection is always a tractability issue. For Ned the state of the simulation is saved to a log file along with each object that the Approach Officer clicks. All information is time stamped to the nearest tick (16.667 msec). In addition, the laptop computer feeds directly into a VCR so that everything that happens on the screen is recorded along with all verbalizations. Reducing the redundancy of the information displays was a tractability goal for data interpretation. Each of the 10 displays has been carefully designed to contain a minimum of overlapping information. Hence, if an Approach Officer went to a given display we infer that he was seeking the unique information that it contains.

Figure 5 shows one of Ned's 10 displays that Approach Officers use for situation assessment. As will be apparent to most readers, Ned requires specialized knowledge to understand its displays, and even more knowledge to localize an enemy submarine. Hence, Ned would be intractable for

any population other than experienced submarine officers. However, given their extensive knowledge, the officers we studied required approximately 15 min of training. For our purposes, a final tractability goal was that computational cognitive models written in ACT-R 4.0 (Anderson & Lebiere, 1998) be able to use Ned in the same manner that our Approach Officers do.

Ned maintains a reasonable level of engagement, but only for those with much prior submarine experience – that is, Ned is not suitable for college sophomores (or for most of our readers). The mission and scenarios used for Ned entail searching for an enemy submarine hiding in deep water. These missions and scenarios correspond highly with those on which Approach Officers train onboard submarines. We believe that this correspondence contributes to the engagement of Ned.

Example: Argus – A synthetic environment

The Argus simulated task environment (Schoelles & Gray, 2001a) was designed after studying the Advanced Cockpit task (Ballas, Heitmeyer, & Perez, 1992), Space Fortress (Donchin, 1995), the Team Interactive Decision Exercise for Teams Incorporating Distributed Expertise (TIDE2 Hollenbeck et al., 1995; Hollenbeck et al., 1997), and Tandem (Dwyer, Hall, Volpe, & Cannon-Bowers, 1992). Like the Advanced Cockpit and Space Fortress, Argus places a premium on embodied cognition (Kieras & Meyer, 1997) and rapid shifts in serial attention (Altmann & Gray, 2000b). It can be used in either single-subject (Argus Prime) or team (Team Argus) mode. Like TIDE² and TANDEM, Argus emphasizes judgment and decision-making in a multiple-cue probability task (see also, Gilliland & Landis, 1992). Argus was designed to facilitate the investigation of a broad category of research questions centered on how interface design affects cognitive workload in both team and individual performance.

Although the design of Argus was informed by our knowledge of real-world systems, specifically the U. S. Army's Patriot Air Defense system and the U. S. Air Force's AWACS (airborne warning and control system), Argus is not a hi-fidelity simulation nor is it a scaled world. Rather Argus was designed to incorporate a few aspects of several systems; that is, Argus is low to medium on the correspondence dimension.

Tractability does not vary monotonically with simplicity. Among the synthetic environments that we studied carefully, Argus and Space Fortress are the most complex, but the most tractable for data collection. For our needs, Argus is more tractable than Space Fortress. For Argus, interface elements can be quickly changed, scenarios written at varying levels of complexity, and the entire state of the simulation can be saved and played back later along with all of the subject's mouse movements, mouse clicks, eye movements, and decisions. Furthermore, cognitive models written in ACT-R/PM can interact with Argus using the same interfaces and scenarios that human subjects use (Schoelles & Gray, 2001a, 2001b).

Unlike the VCR, Argus Prime requires one hour and Team Argus requires two hours of instruction for subjects to become proficient. Hence, for some researchers the time required for training may make Argus intractable. The usability aspect of Argus is a research question; by design, some variations of Argus are easier for subjects to use than are other variations.

Argus has enough elements of a video game that some subjects find it engaging, whereas it is clear that other subjects do not. Changes in the Argus interface do not change the overall task, but do affect the ease with which subtasks are performed. Interestingly, the easiest interface seems to be associated with the most reports of boredom.

Dimensions of simulated task environments

By capturing three different perspectives, the researcher's, the task's, and the participant's, the dimensions of tractability, correspondence, and engagement (see Figure 1) provide a vocabulary for discussing the differences and similarities among various simulated task environments.

Tractability

Tractability is an important but relative dimension. Whether a given simulated task environment is tractable or not is defined by the research question. For example, Argus Prime is a tractable environment for the study of cognitive workload but is intractable for the study of serial attention.

For most researchers the essential aspect of tractability is whether the simulated task environment enables them to collect the data they need with the frequency and accuracy that they need it. For data collection, high correspondence may be the enemy of tractability. High correspondence systems may impose requirements that are at odds with data collection. For example, even if flight simulators could be instrumented to the satisfaction of a researcher, it would remain extremely difficult to know what information a pilot had or was acquiring at any given moment. Similarly, tactical engagement simulation systems for small unit combat training (Gray, 1983) take place over an extended terrain with each soldier doing his best to be invisible to the enemy. For such simulations, knowing "who was where when" let alone knowing "who knew what when" may be an inherently intractable task.

Training, usability, and the expense of running the simulation are important considerations that affect tractability. Likewise, different research questions or constraints may impose idiosyncratic but vital requirements for tractability. Examples discussed above included portability and the ability to support cognitive models as users.

In considering tractability, it is important to bear in mind the distinction between questions and research questions. Not all good questions can profitably be addressed through experimental research. A research question is

one that is tractable given current limits on funding, time, technology, or training.

Correspondence

Correspondence can vary from one aspect of many systems, to some aspects of one system, to many aspects of one system as the goal of research shifts from basic research, to building next generation systems, to fixing problems in current systems. Unfortunately, the goal of building high correspondence simulated task environments exerts a siren-like lure to the unwary researcher and, all too often, to those who fund and sponsor research. For these individuals I can do no more than to paraphrase Brehmer and Dörner (1993) as well as DiFonzo et al. (1998) and state that too high of a correspondence may limit generality.

Engagement

The engagement dimension arises not from the needs of the task or researcher but from the needs of the participant. If the task is not engaging it simply may not be done or, at least, not performed with the attention and detail desired by the researcher. Engagement may be a two-edged sword. It is possible to imagine a simulated task environment that is so engaging that some participants ignore instructions in an attempt to do whatever is possible to “win” the game.

Summary

In this article I have introduced the term simulated task environment, discussed different types of simulated task environments, as well as dimensions on which they may differ. The distinctions between hi-fidelity simulations, scaled worlds, synthetic environments, and simple laboratory environments are not crisp and clean. Rather, such distinctions are inherently fuzzy and overlapping. For example, if the research question concerns such issues as how airline pilots perform in emergency situations when they believe that their own lives are in danger, then the correspondence between a hi-fidelity simulation of a commercial jetliner and the emergency situation may be too low to yield valid results. Similarly, the dimensions provided – tractability, correspondence, and engagement – are meant to capture three perspectives on simulated task environments, but are surely not an exhaustive set.

Despite limitations my goal in making these distinctions is to facilitate communication among researchers and between the research community, our sponsors, and the public. Although the distinctions are imprecise, if widely adopted they will facilitate communication across these diverse groups.

The researcher's dilemma revisited

This article began with a quotation that I expected would get many who use simulated task environments nodding in agreement. Unfortunately, the quotation is half-right and half-wrong. The spirit of the quotation is right. Simulated task environments allow us to address a range of research questions that cannot be addressed by either field research or by simple laboratory environments. However, the implications of the quotation are demonstrably wrong. Field research can yield definite conclusions and even the simplest laboratory environment can yield interesting conclusions.

Research paradigms, whether field research, simulated task environments, or simple laboratory tasks are inherently neither good nor bad. The key issue for any researcher is not paradigm but productivity: are the questions addressed of interest and does the methodology employed support the inferences that the researchers wish to draw?

In a diverse research community we must listen carefully as individual researchers define their research questions and research approaches. If we find the questions interesting, then we must insist that all researchers – despite paradigm – convince us that their research is both reliable and valid and can be generalized to the situations defined by the research question (see also, Gray & Salzman, 1998).

References

- Allport, A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance IV* (pp. 421-452). Cambridge, MA: MIT Press.
- Altmann, E. M., & Gray, W. D. (1998). Pervasive episodic memory: Evidence from a control-of-attention paradigm. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 42-47). Hillsdale, NJ: Erlbaum.
- Altmann, E. M., & Gray, W. D. (1999a). Functional decay in serial attention, *Proceedings of the Sixth ACT-R Workshop*. Fairfax, VA: ARCH Lab.
- Altmann, E. M., & Gray, W. D. (1999b). Serial attention as strategic memory. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society* (pp. 25-30). Hillsdale, NJ: Erlbaum.
- Altmann, E. M., & Gray, W. D. (2000a). An integrated model of set shifting and maintenance. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modeling* (pp. 17-24). Veenendal, NL: Universal Press.
- Altmann, E. M., & Gray, W. D. (2000b). Managing attention by preparing to forget, *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.
- Altmann, E. M., & Gray, W. D. (2002a). *Forgetting to remember: The functional relationship of decay and interference*. *Psychological Science*, 13(1), 27-33.
- Altmann, E. M., & Gray, W. D. (2002b). *The anatomy of serial attention: An integrated model of task switching and maintenance*. Manuscript submitted for publication.

- The American Heritage Dictionary of the English Language*. (Fourth ed.)(2000). New York: Houghton Mifflin Company.
- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Lebiere, C. (Eds.). (1998). *Atomic components of thought*. Hillsdale, NJ: Erlbaum.
- Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7(1), 66-80.
- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(4), 723-742.
- Ballas, J. A., Heitmeyer, C. L., & Perez, M. A. (1992). Evaluating two aspects of direct manipulation in advanced cockpits. *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems* (pp. 127-134).
- Brehmer, B., & Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior*, 9(2-3), 171-184.
- Byrne, M. D. (1999). *ACT-R Perceptual-Motor (ACT-R/PM): A users manual*. <http://chil.rice.edu/byrne/RPM/docs/index.html>.
- Byrne, M. D., & Anderson, J. R. (1998). Perception and action. In J. R. Anderson & C. Lebiere (Eds.), *The atomic components of thought* (pp. 167-200). Hillsdale, NJ: Erlbaum.
- Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- DiFonzo, N., Hantula, D. A., & Bordia, P. (1998). Microworlds for experimental research: Having your (control and collection) cake, and realism too. *Behavior Research Methods, Instruments, & Computers*, 30(2), 278-286.
- Donchin, E. (1995). Video games as research tools: The Space Fortress game. *Behavior Research Methods, Instruments, & Computers*, 27(2), 217-223.
- Dwyer, D. J., Hall, J. K., Volpe, C., & Cannon-Bowers, J. A. (1992). *A performance assessment task for examining tactical decision making under stress* (Special Report 92-002). Orlando, FL: Naval Training System Center.
- Ehret, B. D., Gray, W. D., & Kirschenbaum, S. S. (2000). Contending with complexity: Developing and using a scaled world in applied cognitive research. *Human Factors*, 42(1), 8-23.
- Fu, W.-t., & Gray, W. D. (2000). Memory versus Perceptual-Motor Tradeoffs in a Blocks World Task. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society* (pp. 154-159). Hillsdale, NJ: Erlbaum.
- Gilliland, S. W., & Landis, R. S. (1992). Quality and quantity goals in a complex decision task: strategies and outcomes. *Journal of Applied Psychology*, 77, 672-681.
- Gray, W. D. (1983). Engagement simulation: A method of tactical team training. *Training and Development Journal*, 37(7), 29-34.
- Gray, W. D. (2000). The nature and processing of errors in interactive behavior. *Cognitive Science*, 24(2), 205-248.
- Gray, W. D., & Fu, W.-t. (2001). Ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head: Implications of rational analysis for interface design. *CHI Letters*, 3(1), 112-119.
- Gray, W. D., & Kirschenbaum, S. S. (2000). Analyzing a novel expertise: An unmarked road. In J. M. C. Schraagen, S. F. Chipman & V. L. Shalin (Eds.), *Cognitive task analysis* (pp. 275-290). Mahwah, NJ: Erlbaum.

- Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*(3), 203-261.
- Hoffman, J. E. (1998). Visual attention and eye movements. In H. Pashler (Ed.), *Attention* (pp. 119-153). East Sussex, UK: Psychology Press.
- Hollenbeck, J. R., Ilgen, D. R., Sego, D. J., Hedlund, J., Major, D. M., & Phillips, J. (1995). Multilevel theory of team decision making: Decision performance in teams incorporating distributed expertise. *Journal of Applied Psychology, 80*(2), 292-316.
- Hollenbeck, J. R., Sego, D. J., Ilgen, D. R., Major, D. A., Hedlund, J., & Phillips, J. (1997). Team decision making accuracy under difficult conditions: Construct validation of potential manipulations using the TIDE 2 simulation. In M. T. Brannick, E. Salas & C. Prince (Eds.), *Team performance, assessment, and measurement: Theory, research, and applications*. Hillsdale, NJ: Erlbaum.
- Irving, S., Polson, P., & Irving, J. E. (1994). A GOMS Analysis of the Advanced Automated Cockpit. In B. Adelson, S. Dumais, & J. Olson (Eds.), *ACM CHI'94 Conference on Human Factors in Computing Systems* (Vol. 1, pp. 344-350). New York: ACM Press.
- Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction, 12*(4), 391-438.
- Kirschenbaum, S. S., & Gray, W. D. (2000). The précis of project Nemo, phase 2: Levels of expertise. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society* (pp. 753-758). Mahwah, NJ: Erlbaum.
- Logan, G. D. (1996). The CODE theory of visual attention: an integration of space-based and object-based attention. *Psychological Review, 103*(4), 603-649.
- Lohse, G. L., & Johnson, E. J. (1996). A comparison of two process tracing methods for choice tasks. *Organizational Behavior and Human Decision Processes, 68*(1), 28-43.
- MacKenzie, I. S. (1992). Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction, 7*(1), 91-139.
- Moray, N., Hiskes, D., Lee, J., & Muir, B. M. (1995). Trust and human intervention in automated systems. In J.-M. Hoc, P. C. Cacciabue & E. Hollnagel (Eds.), *Expertise and technology: Cognition & human-computer cooperation* (pp. 183-194). Hillsdale, NJ: Erlbaum.
- Mozer, M. C., & Sitton, M. (1998). Computational modeling of spatial attention. In H. Pashler (Ed.), *Attention* (pp. 341-388). East Sussex, UK: Psychology Press.
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York: Cambridge University Press.
- Pylyshyn, Z. (1998). Visual indexes in spatial vision and imagery. In R. D. Wright (Ed.), *Visual attention* (pp. 215-231). New York: Oxford University Press.
- Ritter, F. E., Baxter, G. D., Jones, G., & Young, R. M. (2000). Supporting cognitive models as users. *ACM Transactions on Computer-Human Interaction, 7*(2), 141-173.
- Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*(2), 207-231.
- Schoelles, M. J., & Gray, W. D. (2000). Argus Prime: Modeling emergent microstrategies in a complex simulated task environment. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modeling* (pp. 260-270). Veenendaal, NL: Universal Press.

- 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..
- Schoppek, W., Holt, R. W., Diez, M. S., & Boehm-Davis, D. A. (2001). Modeling behavior in complex and dynamic situations – the example of flying an automated aircraft. In E. M. Altmann, A. Cleermans, C. D. Schunn & W. D. Gray (Eds.), *Proceedings of the Fourth International Conference on Cognitive Modeling (ICCM-2001)*. Mahwah, NJ: Erlbaum.
- Webster's New World Dictionary of the American Language*. (1960). New York: The World Publishing Company.
- Yantis, S. (1998). Control of visual attention. In H. Pashler (Ed.), *Attention* (pp. 223-256). East Sussex, UK: Psychology Press.