THE ROLE OF TRANSFER-APPROPRIATE PROCESSING IN THE EFFECTIVENESS OF DECISION-SUPPORT GRAPHICS

A Dissertation

by

MICHAEL E. STISO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2003

Major Subject: Psychology

THE ROLE OF TRANSFER-APPROPRIATE PROCESSING IN THE EFFECTIVENESS OF DECISION-SUPPORT GRAPHICS

A Dissertation

by

MICHAEL E. STISO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Steven M. Smith
(Chair of Committee)

Colin Allen
(Member)

Winfred Arthur
(Member)

Stephanie Payne
(Member)

Steve Rholes
(Head of Department)

August 2003

Major Subject: Psychology

ABSTRACT

The Role of Transfer-Appropriate Processing in the Effectiveness
of Decision-Support Graphics. (August 2003)

Michael E. Stiso, B.A., Purdue University;
M.S., University of Oregon
Chair of Advisory Committee: Dr. Steven M. Smith

The current project is an examination of the effectiveness of decision-support graphics in a simulated real-world task, and of the role those graphics should play in training. It is also an attempt to apply a theoretical account of memory performance—transfer-appropriate processing—to naturalistic decision making. The task in question is a low-fidelity air traffic control simulation. In some conditions, that task includes decision-support graphics designed to explicitly represent elements of the task that normally must be mentally represented—namely, trajectory and relative altitude. The assumption is that those graphics will encourage a type of processing different from that used in their absence. If so, then according to the theory of transfer-appropriate processing (TAP), the best performance should occur in conditions in which the graphics are present either during both training and testing, or else not at all. For other conditions, the inconsistent presence or absence of the graphics should lead to mismatches in the type of processing used during training and testing, thus hurting performance. A sample of 205 undergraduate students were randomly assigned to four experimental and two control groups. The results showed that the support graphics provided immediate performance benefits, regardless of their presence during training. However, presenting them during training had an apparent overshadowing effect, in that removing them during testing significantly hurt performance. Finally, although no support was found for TAP, some support was found for the similar but more general theory of identical elements.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Steven M. Smith, whose guidance was both flexible enough to let me explore my own research program yet firm enough to keep me from heading into serpent-infested seas. I would also like to thank the Federal Aviation Administration for the use of its Air Traffic Scenario Test simulation, upon which I based the design of the ATC task described in this document. Finally, I could not have completed my research in such a timely manner without the Graduate Student Research Program fellowship awarded to me by NASA and the Marshall Space Flight Center.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The concept of “decision making aids” has been around at least since people began using graphs and diagrams to summarize and display complex information. The phrase can refer to a variety of tools. In the past couple of decades, though, as computers have become sophisticated enough to take an active role in decision making, “decision aid” has become associated with the idea of tools intended to automate and take over parts of that process.

For example, in the realm of air traffic control (ATC), Manning and Broach (1992) describe automated decision aids that identify potential problems (e.g., conflicts between two or more aircraft) with the current trajectories of multiple aircraft. A more advanced aid would suggest solutions to the problem, based on certain criteria. Gronlund, Canning, Moertl, Johansson, Dougherty, and Mills (2002) describe a similar ATC tool that provides controllers with an initial route-sequencing plan for aircraft, which the controllers can then fine-tune. Such aids thus automate parts of decision making for the users, relieving them of some of the burden of the process and perhaps allowing more aircraft to be handled and better decisions to be made.

However, automation, at least in its current state, is not without cost. Users may become overreliant on it, for example, which can lead to complacency and impaired performance (Yeh & Wickens, 2001). Similarly, operators may demonstrate reduced

This thesis follows the style of the Journal of Experimental Psychology: Applied.

situation awareness in the presence of decision aids, particularly if trained with such aids (Endsley, 1996; Endsley & Kiris, 1995). Alternatively, users may simply mistrust automation, especially if the tool is perceived as unreliable (Moray, Inagaki, & Itoh, 2000; see also Muir, 1994; Yeh & Wickens, 2001).

Some researchers, then, perhaps hoping to sidestep some of those issues, have adopted a focus on decision support rather than decision aiding. Morrison, Kelly, Moore, and Hutchins (1998) state that the goal of decision support is to leave as much of the decision making process as possible with the user. Their decision-support system for Naval tactical decision making augments information already available in the system in an attempt to make it more meaningful, rather than extensively filtering or processing information. That system consists of a number of modules. The "Basis for Assessment" module, for example, tabulates, categorizes, and displays the data on which the system based its recommendation regarding the threat level of a given target. It is designed to facilitate story generation by attempting to explain available and missing data. Another module is more graphical in nature, showing the position of a target relative to the user's ship, how that target has been moving over time, and whether it or the user's ship can engage the other. A particularly important function of that module is its capacity for graphically depicting a target's movement history, which the designers believe reduces the demands on short-term memory imposed by trying to interpret the significance of a target. Similar "at-a-glance" features added to later versions include velocity leaders that indicate the relative speed of all aircraft on the screen, as well as course histories that indicate where a track has been relative to landmarks, air corridors, and other aircraft.

One function that the various modules in that system have in common is that they all essentially attempt to capitalize on human perceptual capabilities. The Basis for Assessment module, for example, organizes data in ways that allow humans to employ their pattern-recognition abilities. The graphical components also translate at least part of a generally conceptual process into a perceptual process (e.g., trying to determine the significance of a target by understanding where it has been and where it is going).

However, whether such a translation is actually helpful in decision making is not well-tested. For example, although Morrison et al. (1998) found that their decision support system generally led to better performance, they were unable to test which aspects of which modules were responsible for the improvement. The interface they designed was intended to serve many functions, making it difficult to determine whether it was specifically its perceptual characteristics that provided the advantage or whether it was some other aspect of the interface. That is often the case with studies involving computer interfaces, possibly because the field is still relatively new, but also because the graphics capabilities of computers have only recently become such that researchers other than computer scientists can use them. As a result, many studies examining the use of graphics in decision support tend to be usability tests or product comparisons. In other words, rather than using specific theories as guides in the investigation, the studies work to improve decision making by determining which interfaces result in the best performance.

Reasons for the Current Study

The current project is primarily an examination of the effectiveness of a type of decision-support graphics (DSGs) in a real-world task and of the role those graphics should play in training. It is also an attempt to apply a theoretical account of memory performance—transfer-appropriate processing—to naturalistic decision making (i.e., experienced decision making in a field setting; Zsambok, 1997). As such, the results should be of both applied and theoretical interest.

Specifically, using a low-fidelity air traffic control (ATC) simulation as a testbed, I planned to investigate three main issues:

1) Can the transfer-appropriate processing framework can be applied to naturalistic decision making?
2) Can computer graphics effectively support decision making on a real-world task?
3) Is it helpful or harmful to present those graphics during training?

From a theoretical perspective, this study places the idea of decision-support graphics inside a theoretical framework, and then tests that framework. More to the point, it adds to the literature by attempting to apply a theoretical framework of memory performance to the area of decision making.

From an applied perspective, the decision support graphics are intended to serve as external representations of elements that are normally mentally represented, and so this research should help to determine whether that kind of graphical enhancement is helpful or not. Put another way, as described earlier regarding Morrison et al.'s (1998) decision support system, it is believed that the graphics will improve decision making by translating some of the conceptual processing normally required by air traffic control into perceptual processing, thus capitalizing on human perceptual capabilities. In addition, this work should indicate whether those graphics should or need to be present during training.

The following sections will outline the transfer-appropriate processing account of memory and how it applies to this project, provide counterarguments and alternate predictions to that account, and describe the rationale behind the graphics involved.

Transfer-Appropriate Processing

Many types of mental operations are available for processing a stimulus at any given time—perceptually, conceptually, via different modalities, and so on. The theory of transfer-appropriate processing (TAP) states that memory performance depends on the overlap between the types of processing used during study of an item and those used during a later memory test of that item. Essentially, the greater the overlap, the better the memory.

TAP resembles other theories that are based on the similarity between study and testing conditions, such as contextual reinstatement, encoding specificity, identical-elements theory, and the like; however, it is different in that it focuses on similarity of the processing used in the different situations, rather than on the stimuli involved. In other words, the more that the mental operations used during memory testing are the same as those used during study, the greater the memory performance.

TAP and Memory

Much of the evidence directly investigating TAP to date has come from studies focusing on its role in explaining dissociations in explicit vs. implicit memory performance (e.g., Blaxton, 1989; Graf & Ryan, 1999; Horton & Nash, 1999; Leshner & Coyle, 2000; Rajaram, Srinivas, & Roediger, 1998). Blaxton (1989), for example, used TAP to predict dissociations in memory performance on five different types of memory tests—free recall, semantic cued recall, general knowledge, word-fragment completion, and graphemic cued recall. Those tests were labeled as either conceptually driven or data-driven, depending on the type of processing thought to be primarily involved in their completion. Using methods based on Jacoby (1983, as cited in Blaxton, 1989), Blaxton placed the free recall, semantic cued recall, and general knowledge tests in the conceptually driven group; the word-fragment completion and graphemic cued recall tests, having been shown to be largely dependent on physical features of stimuli, were placed in the data-driven category.

As would be predicted by the TAP account, memory performance was enhanced on tests considered to be conceptually driven when, during study, participants were instructed to process the target items in a meaningful way—for example, by generating rather than reading the items or by forming mental images of them. Those manipulations had little effect on the data-driven memory tests. In contrast, focusing participants on the physical features of the target items (including modality and typography) enhanced performance on the data-driven but not the conceptually driven tasks.

Blaxton's explanation was that the data-driven and conceptually driven memory tasks require different types of processing or mental operations. The former tend to rely on the analysis of physical features, whereas processing of the latter tends to be more elaborate and based on meaning. When the type of processing required by the test overlapped with the type of processing participants were instructed to perform during study, memory was enhanced.

Although much of the TAP research focuses on dissociations in retrospective memory, it does seem to be gaining ground in other areas, as well. For example, Meier and Graf (2000) extended the literature from retrospective to prospective memory. They showed that TAP can account for performance dissociations due to concurrent processing between an ongoing task and a prospective memory test.

Leshner and Coyle (2000) applied the TAP theory to research on memory for television news. Based on the TAP account of memory performance, they argue that findings in the literature suggesting poor memory for such news may actually reflect an inappropriate match between the mental requirements of the memory tests and the mental requirements of watching television. To test that idea, they borrowed methods used by Blaxton (1989) and by Roediger, Weldon, Stadler, and Riegler (1992). Specifically, they presented participants with televised news stories, but encouraged those participants to process the stories either conceptually or perceptually (i.e., data-driven processing). Conceptually driven processing was encouraged by instructing participants to rate a given story's meaningfulness, personal relevance, importance, informativeness, and seriousness. Data-driven processing was encouraged by having participants rate the story's pace, audio quality, picture clarity, camera work, and the reporter's voice. Afterward, participants received one of four memory tests: graphemic cued recall, semantic cued recall, word fragment completion, or general knowledge. Drawing on Blaxton (1989), they classified the graphemic cued recall and word fragment completion tests as tasks that require primarily data-driven processing; the semantic cued recall and general knowledge tests were classified as conceptually driven.

According to the TAP account, participants who conceptually processed the news stories should demonstrate better memory on the conceptually driven memory tests; in turn, participants who perceptually processed the stories should do better on the data-driven memory tests. The results largely supported those predictions, although the performance difference on the graphemic cued recall test in the data-driven condition failed to achieve significance. (The authors offer differences in modality between study and test as a possible factor underlying that result.)

TAP and Physical Performance

Transfer-appropriate processing has even found its way into the area of sports psychology and physical performance. Peynircioglu, Thompson, and Tanielian (2000) proposed that TAP could explain a number of empirical findings regarding the relationship between mental imagery and physical performance. In general, they suggested that, according to TAP, encouraging a set of cognitive activities during practice that is similar to that used during performance would increase the effects of practice on performance. Specifically, they predicted that the effects of a particular practice strategy on subsequent performance depends on the match between the actions and thoughts used during each.

To investigate that idea, they designed a study involving two physical performance tasks: free-throw shooting and grip strength. The authors described free-throw shooting as a fine motor skill with high cognitive demand: Among other factors, participants have to concentrate on body position; distance, height, and size of the basket; wrist action; and required strength. In comparison, grip strength is a predominantly gross motor task with much less cognitive demand: Participants primarily have to concentrate only on gathering their strength.

Before they performed each task, participants used one of three preparation strategies: nonspecific arousal (i.e., "psyching-up"); mental rehearsal (i.e., imagery); or nothing (i.e., control condition). The nonspecific arousal condition involved having participants engage in physical activity (e.g., running around or pumping fists) and verbal self-encouragement. The mental rehearsal condition involved guided visualization of the actions required for the task.

Based on TAP, the authors hypothesized that the free-throw condition, given its higher cognitive demand, would benefit from the similar types of mental operations required by the imagery preparation, but not from the actions used during the nonspecific arousal preparation. The grip strength task, on the other hand, being a simple application of strength, should benefit less from the imagery than from the nonspecific arousal preparation. The results largely supported the predictions. Free-throw shooting improved after the imagery strategy but not after the nonspecific arousal strategy.

However, for grip strength, the less cognitively demanding task, improvements resulted from the nonspecific arousal preparation, but not from mental rehearsal.

The authors explain the results by suggesting that free-throw shooting, compared to grip strength, included more elements that could be successfully imaged, providing a greater match between the actual actions it required and those imagined during mental rehearsal. The actions performed during the nonspecific arousal preparation, however, did not match those performed during the actual task. In contrast, grip strength is either not as easily imaged as free-throw shooting, or the actual actions it required were different from those imagined during mental rehearsal. As a result, there was no benefit from the imagery preparation. However, it did benefit from the nonspecific arousal condition, presumably because the heightened arousal and physiological priming better matched the actual actions used in the grip strength task.

Shanks and Cameron (2000) also brought TAP into the physical performance arena. They used the theory to explain the unexpected results of their study, in which mental practice had no effect on performance of a dot-location reaction time task. Physical practice, as might be expected, did enhance performance on the task. The authors suggest that certain fine details (such as precise timing) involved in performing the task were dissimilar to the underlying operations involved in mentally practicing the task. The operations underlying actual physical practice, on the other hand, provided a better match to those used during the test, and thus enhanced performance. In comparing this study to Peynircioglu et al. (2000), in which mental practice *did* have an enhancing effect for a cognitively demanding motor task, it may be that the dot-location reaction time task was not so cognitively demanding as to benefit from practice involving cognitive operations.

A New Area: Naturalistic Decision Making

Despite the growing variety of fields in which TAP has been applied, little if any research has explored whether the theory can successfully predict cognitive performance outside of memory tests. For example, even taking the mental practice conditions used in Shanks and Cameron (2000) and Peynircioglu et al. (2000) into account, it has never been applied to long-term study situations such as skill acquisition or to testing situations involving high-level cognitive activities such as naturalistic decision making. Both of those situations presumably involve the use of a number of types of mental operations. If it can be shown that such situations follow the TAP account, that could have a number of implications for training programs. For example, if a particular decision-support tool is found to be unsuccessful in the workplace, it could be because its presence encourages the use of a type of processing different from that used during training, when the tool was not present. If the tool is then incorporated into the training program, it may actually degrade decision making performance in workplace situations that *do not* incorporate the support tool—again, because its absence would encourage a different type of mental operation than that used during training, when it was present.

The focus here is on naturalistic rather than analytical decision making. Naturalistic decision making is essentially experienced decision making in natural or real-world settings, or simulations of such settings (Zsambok, 1997). Naturalistic settings have a number of key characteristics (Orasanu & Connolly, 1993, as cited in Zsambok, 1997):

ill-structured problems; uncertain, dynamic environments; shifting, ill-defined, or competing goals; action/feedback loops, rather than one-shot decisions; time constraints; high stakes; multiple players; and organizational goals and norms.

Present Study

Whether TAP can be applied to naturalistic decision making was tested in the current project. The decision-making task is this case is a low-fidelity air traffic control simulation, and the decision-support tool is a set of graphics intended to facilitate the processing of situation elements (aircraft trajectory and altitude) that must normally be mentally represented. As will be described, the ATC task, although a simulation, includes many characteristics of a naturalistic situation, such as a dynamic environment, feedback loops, competing goals, and time constraints.

The ATC Task and Decision-Support Graphics

In the ATC task, participants are responsible for guiding aircraft quickly but safely through a square-shaped airspace. However, a number of factors complicate the task. For example, the airspace contains several aircraft at any given time, each of which needs to be guided to a specific location. En route to those locations, the aircraft cannot fly over airports or get too close to the border or to other aircraft at the same altitude. Once at its destination, an aircraft must exit or land at a particular speed, altitude, and heading. Exiting is further complicated in that new planes entering the airspace tend to appear near the exits, potentially causing a conflict or crash with the exiting plane if participants are not paying attention. Landing at airports is similarly complicated in that the wind direction changes at regular intervals, requiring aircraft to change the direction in which they land on the runways; a lapse in attention in such a case can lead to a crash.

Maintaining aircraft safety under such conditions likely necessitates the use of mental simulation to anticipate aircraft trajectories and potential conflicts in flight plans for a particular plane, or for the group of planes as a whole. Good performance requires the ability to include contingencies in those flight plans, including simply keeping an eye on potential problem spots—in other words, keeping some attentional resources in reserve. Measures such as the number of separation violations with en route vs. waiting aircraft, number of runway violations, and number of changes to aircraft altitude, speed, and heading should provide an indication of participants' evaluation and contingency-planning activities, with more such errors and changes corresponding to increases in workload. If so, then reducing workload should result in fewer such errors, which is where the decision-support graphics enter the picture.

The rationale behind the DSGs used in the present experiment is based on the idea that experts perceive unseen relationships and processes that others cannot see (Klein & Hoffman, 1993). Part of mental simulation involves representing those unseen elements in a mental model and then manipulating them. Mental simulation, however, is both taxing and time-consuming, and it may be that having to imagine and account for such unseen factors makes up much of that burden. If so, then that burden could perhaps be reduced if the need for such imagining were eliminated, such as by making those elements explicit through graphical presentation on a computer screen. In the ATC task, maintaining aircraft separation requires anticipating where planes will end up and how fast, as well as understanding where they are (vertically as well as horizontally) in relation to each other. So, visually displaying an aircraft's trajectory and speed may reduce the burden of mentally simulating that information; as well, providing a visual indication of aircraft altitudes may facilitate mentally spatially organizing of them. Both are methods of visually depicting normally unseen factors in a situation, and as such are believed to reduce the burden of simulation. In turn, because mental simulation is an aspect of situation awareness (SA; see Endsley, 1995b) the graphics may alleviate the need to maintain SA, at least as it relates to aircraft separation and navigation. For example, they provide perceptual cues for impending aircraft conflicts, meaning participants need not devote much attention (relatively speaking) to such conflicts until the cues indicate such a conflict is about to happen. An alternative way to view the role of the graphics, as described earlier, is that they may transform some of the conceptually-driven or higher-level cognitive processing normally involved in the task into perceptual or data-driven processing, so freeing up resources and aiding learning of the task. However, it may also be that such a transformation simply complicates an already visually loaded situation, instead interfering with learning and decision making.

In any case, choosing what factors to represent through graphics, and how to represent them, was a fairly arbitrary decision. However, given that maintaining aircraft separation would seem to necessitate the projection of aircraft trajectory and speed (i.e., where they will end up, and how fast), as well as an understanding of their relative altitudes, it seemed logical to focus on augmenting trajectory, speed, and altitude. In addition, research has shown that trajectory and altitude are among that top considerations in air traffic controller (Mogford, 1997; Niessen, Eyferth, & Bierwagen, 1999; Whitfield & Jackson, 1982).

Given the scarcity of examples on which to base the design of the support graphics, they were designed according to the experiment's sense of what would make their use most intuitive. The goal was to create graphical indicators whose purpose would be obvious and which also would provide at-a-glance information regarding trajectory and relative altitude of several aircraft. Given that, trajectory is represented by a line that indicates the projection of the aircraft's position forward in time (specifically, three moves), adjusted for speed. Altitude is indicated by a colored circle around the aircraft, with the color representing a particular altitude. The rationale behind the altitude indicators is that ATC presumably involves maintaining something of a spatial organization of the aircraft in the vertical plane; using color to represent different altitudes alleviates that burden by allowing quick scanning and comparison of different aircraft. On a side note, it might also be the case that time is a factor in need of augmentation, given that the runways switch directions at regular intervals, and that new planes are introduced also at set intervals; however, no graphics were designed for that purpose.

In summary, although the ATC task is a low fidelity simulation, it possesses many of the characteristics of a naturalistic decision making task. In particular, it is a dynamic, ever-changing task with action/feedback loops, because any move that participants make changes the situation and requires a reassessment of the environment. For example, altering the course of one aircraft may eventually place it in the path of several others.

Similarly, although there is one main goal—get the aircraft to the exits quickly and safely—there are conflicting ways in which they meet that goal. Because more points are earned the faster an aircraft is exited, for example, participants can choose to pay more attention to aircraft speed than to aircraft separation. On the other hand, points are lost for each aircraft conflict, so some participants may choose to deemphasize speed in favor of keeping them safely separated. Most will probably try to find a balance between the two. In addition, the aircraft do not stop until they reach their destination or crash, and new aircraft are appearing at regular intervals. As a result, the time that participants can spend considering moves for a given aircraft becomes increasingly limited, thus introducing the time constraints common to real-world tasks. As well, an attempt was made to introduce the stakes involved in such tasks by offering a monetary prize for highest score; a pilot study indicated that the amount offered was attractive to most students.

Design and Predictions

Because naturalistic decision making research is concerned with experienced rather than naïve decision makers, participants were trained on the ATC task. Some were trained with the benefit of the decision-support graphics, others without. After training, participants were tested on a more difficult version of the task, half with the graphics and half without. The addition of two control groups, which received no training on the ATC task, yielded the following six groups[1]:

[1] The coding for the different groups consists of a two-letter label, in which the first letter indicates the type of training and the second indicates the type of testing; a "G" indicates the support graphics were present, an "N" indicates no support graphics were present, and a "C" (training only) indicates control groups that did not participate in any training sessions.

1) graphics during training + graphics during testing (GG)
2) no graphics during training + graphics during testing (NG)
3) graphics during training + no graphics during testing (GN)
4) no graphics during training + no graphics during testing (NN)
5) no training + graphics during testing (CG)
6) no training + no graphics during testing (CN)

If the TAP account holds, then participants whose testing condition differs from their training condition (NG and GN) should use a different set of mental operations in each situation. As a result, participants trained with the graphics should do better when tested with the graphics (GG) than when tested without them (GN). Similarly, and what would be most surprising from an applications standpoint, participants trained without the graphics should actually perform better when tested without the graphics (NN) than with them (NG).

TAP and the theory of identical elements

With the ATC task, as an example of how the type of processing involved may differ between the with-graphics and without-graphics conditions, consider a distinction commonly described in TAP studies: conceptually vs. data-driven processing (e.g., Blaxton, 1989; Leshner & Coyle, 2000). Conceptually driven processing involves processing a stimulus at a conceptual level—for example, processing a word according to its meaning. Data-driven processing, on the other hand, is a bottom-up process in which stimuli are encoded at a perceptual level, such as the appearance of a word.

TAP experiments have found that tasks in which study and testing differ in whether they are conceptually or data-driven tend to lead to poor memory performance, whereas those with greater overlap between the type of processing involved produce better performance (Blaxton, 1989; Leshner & Coyle, 2000). The implication is that the mental operations involved in the two types of processing are different, and so study and testing conditions that differ in the type they require are encouraging participants to use different mental operations in either situation, thus hurting memory performance.

It may be that a similar distinction in processing requirements can be made in the ATC task. In particular, because the decision-support graphics are intended to facilitate the mental representation of certain situational elements, it may be that those graphics actually reduce the processing requirements of the task. Alternatively, the graphics can also be thought of as transforming some of the conceptually-driven processing required to perform the task into data-driven processing. In either case, then, compared to the use of the ATC simulation without graphics, the decision-support graphics may actually encourage the use of a different set of mental operations—specifically, those more suited to data-driven or perceptual processing. If so, then the processing requirements would be expected to be greater in the absence of such graphics. In particular, participants in those conditions would need expend more resources on imagining or mentally visualizing the trajectory of several aircraft in the airspace compared to participants for whom the graphics provide such information. That mental visualization is perhaps analogous to conceptually driven processing, in which case it likely requires the use of mental operations more suited to such activity.

To summarize, if the TAP theory can predict performance on the ATC task, then participants trained with the benefit of decision-support graphics should perform better when tested with those same graphics than similarly trained participants tested without them; similarly, participants trained without those graphics should perform better when tested without them than with them. For the latter users, the graphics likely allow users to capitalize on the more-sophisticated human perceptual system. Differences in performance, then, may be the result of differences in the mental operations used in conceptually vs. data-driven processing.

However, it is difficult to be certain that people will actually use a different type of processing in the presence of the graphics. One advantage of the previous studies investigating TAP is that they have been able to manipulate explicitly the type of processing that participants used to study the stimuli. For example, if studying memory for word lists, experimenters can have participants either generate the words themselves or instead read them from a list. The former method presumably involves primarily conceptual processing, whereas the latter involves primarily data-driven processing. Knowing that, experimenters can then match those study conditions to testing conditions that encourage similar types of processing. Being able to manipulate processing in such a way is essential for differentiating TAP from similar theories such as contextual reinstatement, encoding specificity, and the like. Basically, TAP focuses on the type of processing rather than the context involved, and so it is necessary to be able to say that the type of processing was actually different or the same between study and test.

In the current project, however, participants are not made to process the same stimuli in different ways. Rather, the argument is that different stimuli will encourage different types of processing, leading to differences in performance based on whether the same type of processing is involved during testing. That assumption can perhaps be supported, though, by comparing the pattern of results to what would be predicted by the similar but more general theory of identical elements.

Identical-elements (IE) theory and TAP make the same basic predictions: similar conditions lead to better performance, different conditions lead to worse performance. However, whereas TAP theory attributes such performance differences to the type of processing involved, IE theory is more general: It predicts performance differences, but it does not explain them beyond pointing out differences in the stimulus/response elements making up the training and testing conditions.

Recent empirical evidence supporting IE theory is scarce. The concept of IE has been around a long time and seems to have become a generally accepted fact in textbooks and the like. It is based on classical conditioning theory (Goldstein & Ford, 2002), and empirical studies under that and similar headings in the older literature have perhaps contributed to its acceptance.

Goldstein and Ford (2002) describe IE predictions in terms of transfer, with the performance of participants in experimental conditions being compared to that of a baseline control group which had no training. The addition of that baseline to the current study may allow the conclusion that the support graphics encouraged different types of processing, even though processing was not explicitly manipulated. The reason is that IE and TAP will make the same predictions in certain circumstances—namely, when the type of processing is different between study and test. If processing does not change, though, the theories make different predictions.

Essentially, the elements in identical elements theory involve both stimuli and responses. Based on the differences in either stimulus or response between testing and training, Goldstein and Ford (2002) say that certain directions in transfer (positive or negative) would be expected (see Table 1). If everything stays the same between conditions, one would expect high-positive transfer from training to testing, meaning that performance should be better than that of an untrained control group. If just the stimuli change, but the type of responses required to perform the task stay the same, positive transfer would still be expected.

Table 1
Direction of transfer predicted by IE theory

Stimulus	Response	Transfer
Same	Same	High Positive
Different	Same	Positive
Same	**Different**	**Negative**
Different	**Different**	**Negative**

Source: Goldstein & Ford, 2002.

The only situation in which negative transfer (performance worse than in an untrained control group) would be expected is when there is a change in the responses required to perform the task. In the current project, the physical responses that participants must make remain the same across all conditions, and so IE would predict positive transfer in all cases, but much better transfer in cases in which the stimuli stay the same. However, responses can be both physical and mental. So, although the physical responses do not differ between the graphics and no-graphics conditions, it may be that the mental responses do change. In that case, negative transfer would be expected in mismatched training-testing conditions.

Considering type of processing to be a mental response, then if the graphics actually do encourage a different type of processing, responses should differ between training and testing conditions when graphics are present during only one or the other. In that case, IE would predict negative transfer (i.e., worse performance relative to controls) for mismatched conditions, but positive transfer (better performance than controls) for the matched ones; TAP would predict the same.

Hypothesis 1—TAP & IE: If processing is different, both TAP and IE predict that mismatched groups will display negative transfer, and matched ones positive transfer, relative to controls. In terms of relative performance:

- GG > **CG** > NG
- NN > **CN** > GN

On the other hand, if type of processing does not differ in the presence of the graphics, IE and TAP make different predictions. IE would predict positive transfer (performance better than controls) across the board.

Hypothesis 2—IE: If processing remains the same, IE theory predicts positive transfer for mismatched groups and high-positive for matched ones, relative to controls.

- GG > NG > **<u>CG</u>**
- NN > GN > **<u>CN</u>**

In the same situation, TAP predicts equal performance in all conditions (barring any inherent benefit in the graphics themselves), because processing is the same in every condition.

Hypothesis 3—TAP: If processing remains the same, TAP predicts equal performance among the experimental groups.

- GG = NG = NN = GN

See Table 2 for an illustration of IE, TAP, and other competing explanations.

Table 2
Predictions of Relative Performance of Different Groups

Hypotheses	Predictions
1. TAP & IE	GG > **<u>CG</u>** > NG NN > **<u>CN</u>** > GN
2. IE	GG > NG > **<u>CG</u>** NN > GN > **<u>CN</u>**
3. TAP	GG = NG = NN = GN
4. Attention Reallocation	GG, GN > NN, NG
5. Overshadowing	GG, NG, NN > GN
6. Graphics Advantage	GG, NG > GN, NN

Note. GG: graphics during training, graphics during testing; NG: no graphics during training, graphics during testing; GN: graphics during training, no graphics during testing; NN: no graphics during training, no graphics during testing; CG: no training, graphics during testing; CN: no training, no graphics during testing.

In any case, for IE theory, the prediction is that matched conditions will lead to better performance than mismatched conditions. However, the amount and direction of transfer of those groups relative to performance of an untrained control group should shift depending on whether the decision support graphics encourage different types of processing. Hence, by looking at the pattern of results, it should be possible to determine whether or not any performance differences were due to differences in the type of mental operations used.

Attention reallocation

The line of reasoning used to describe the with-graphics and without-graphics conditions as involving data- vs. conceptually driven processing, respectively, happens to support a counterargument, as well. Capacity theories of attention posit a limited pool of attentional "resources" that an individual can spend on a task or tasks. As that pool is drained, performance worsens, as is often demonstrated in dual-task studies in which participants must divide attention between a main task and a secondary task. Divided attention may be necessary within a single task, as well. For example, in Kanfer and Ackerman's (1989a, 1989b) model of skill acquisition, attentional resources must be allocated among the main task activities as well as off-task, self-regulatory, and metacognitive activities.

If the decision-support graphics actually do facilitate or reduce the processing required to perform the ATC task, then it may be that individuals presented with those graphics have enough slack in the demand for their resources to be able to concentrate on such off-task and metacognitive activities, or on aspects of the task that are important but not very salient or immediately essential, such as patterns and timing in aircraft appearances and behavior. Without those graphics, the processing load involved may leave little attention left over for learning anything about the task beyond the basics required to perform it.

In that case, participants trained with the graphics should learn more about the task, and hence they should perform better than participants trained without them, regardless of whether the graphics are included during testing. See Table 2 for an illustration.

Hypothesis 4—Attention Reallocation: In the testing sessions, the groups trained with the support graphics should outperform those trained without them (i.e., a main effect of type of training).

- GG, GN > NN, NG

It may also be, though, that the graphics prove to be helpful only to the lower cognitive ability participants, because those with greater ability may have enough attentional slack that the advantage of the graphics becomes negligible. As well, the difference between high- and low-cognitive ability participants should be reduced in the presence of the decision-support graphics, relative to conditions not involving those graphics.

Overshadowing

Contrary to the idea that the graphics may ease processing and so facilitate learning, it may actually be that they instead overshadow important information. Overshadowing occurs when stimuli presented during training prevents the learning or processing of other stimuli. Cockrell (1979) describes that as a situation in which a very salient and distinctive feature on a target captures a trainee's attention, such that little attention is paid to remaining features. He investigated the role of overshadowing in target identification training by showing participants slides of small-scale model vehicles, which participants were later expected to identify. The vehicles were all the same size and color; the primary cue dimension for identifying the vehicles was shape. The author manipulated the salience or usefulness of that dimension by obscuring the vehicles in some of the slides, thus forcing trainees to look for other means of identification. Specifically, some participants were trained to identify the targets by way of slides in which the vehicles were partially or mostly blocked from view; the rest of the trainees were presented with normal, full-view slides. During testing, though, all participants were exposed to both normal and obstructed slides. (The testing slides had different views of the same types of vehicles used during training.)

As would be expected by an overshadowing account, participants trained with the obstructed slides more accurately identified obstructed vehicles during testing than were participants trained with normal, non-obstructed slides. There was little difference between the groups when tested on non-obstructed slides, though the participants trained on such slides were still the most accurate. Cockrell's (1979) findings suggest that participants trained in the full-visibility condition learned to rely on vehicle shape, the primary distinguishing cue, to identify the vehicles. When that cue was degraded or removed during testing, those participants had little else on which to categorize them, and so their performance dropped. For trainees in the obstructed-view conditions, however, shape was not as accessible or useful a cue, and so they learned to use other, not-so-salient cues to distinguish them.

In the present case, it could be argued that the salience of the decision support graphics hinders participants from learning other important but not-so-salient information or patterns in the task—timing and location of regular events, for example, such as the appearance of new aircraft on the radar. In that case, those participants may end up relying on the graphics and so learn less about the task than participants who did not have the aid. When faced with testing conditions in which the graphics cue was removed, participants would be expected to perform poorly compared to others or to demonstrate negative transfer compared to untrained controls. However, because the damage will have been done during training, participants for whom the graphics were presented only during testing or else not at all should fare relatively well. See Table 2 for an illustration of these predictions.

Hypothesis 5—Overshadowing: The group trained with the graphics should perform worse than the other groups when those graphics are taken away during testing.

- GG, NG, NN > GN

Graphics advantage

Finally, it may be that the support graphics have inherent advantages or disadvantages to performance, but only in an immediate or at-the-moment sense. After all, they were designed to facilitate mental simulation of aircraft position and movement by making the process more perceptual rather than conceptual. If they work, then, they should result in higher scores, fewer plane conflicts, and so on. However, contrary to what is predicted in the other hypotheses, the graphics may have absolutely no effect during training. In other words, the support graphics may make the task somewhat easier to perform when they are present, but they do not necessarily free up enough attention for participants to learn anything more about the task.

That argument is based more on supposition than on findings in the literature, because concrete research on the effects of computer graphics on learning and performance in naturalistic tasks is scant. Though graphics are often mentioned in naturalistic research involving decision support, such as in Morrison et al. (1998), the problem is that the graphics are not really the focus of the research. As a result, their influence on learning and performance is difficult to disentangle from that of the rest of the decision-support system. Even when the graphics *are* the focus, the study in question is generally theoretical and untested rather than empirical. For example, Hollan, Hutchins, and Weitzman (1984) designed a simulator (STEAMER) for teaching steam propulsion systems on Navy ships. What was new about STEAMER is that it used visual cues and signals to explicitly represent such dynamic elements as flow rates in pipes and the rising/falling state of various gauges. In previous simulators and generally in the real-world (at that time), such elements were not "seen" but rather had to be calculated, discovered, or assumed. The creators of STEAMER believed that an explicit representation of such important elements would aid learning; however, the paper was a discussion of STEAMER rather than a test of it, and so no formal results were reported.

In a similar treatment, Lewandowsky, Dunn, Kirsner, and Randell (1997) created a simulation of bushfire-spread, and introduced into it an alert that triggers in situations in which fire spread is likely to violate expectations—such as in conditions involving light wind speed and steep downhill slopes. The intention was to make explicit a normally subtle and hidden relation between wind and ground slope. Because of practical constraints, though, the alert went untested. In the present study, however, the support system is sufficiently simple and contained to be able to determine its effects on performance. See Table 2 for an illustration.

Hypothesis 6—Graphics Advantage: The groups tested with the support graphics should outperform those tested without them (i.e., a main effect of type of testing).

- GG, NG > GN, NN

Asymmetric effects

Of course, the theoretical accounts described above are not mutually exclusive. It could be, for example, that greater overlap between the type of processing used during training and testing do indeed lead to better performance on the ATC task (IE and TAP). However, it could also be that the DSGs reduce the processing burden of the task, allowing participants to allocate attention to and so learn about less salient but important aspects of the task, such as the timing of certain regular events (attention reallocation). In that case, the graphics would be expected to mitigate somewhat the negative effects of mismatched processing types, but only in those conditions in which the graphics were present during training. Or, it could be that the DSGs make the task easier to perform but not to learn, so that participants who encounter them only during testing do as well as or better than participants who do not encounter them at all (graphics advantage). Alternatively, the graphics could prove to be distracting, using up visual resources that could otherwise be applied to the task. In that case, the graphics would be expected to hurt performance, though similarities between training and testing may mitigate the negative effects. In any of those cases, the results would be expected to be asymmetrical, with the decision-support graphics either helping or hurting performance during testing, but with the difference being reduced when the training condition matches the testing condition.

Summary

The theory underlying this study is that the transfer appropriate processing account of memory performance can be extended to performance on real-world tasks. The task in question is an air traffic control simulation, and there are two versions of it: a “normal” version, and a version that provided a type of decision support aid—namely, graphics that indicated plane trajectory and relative altitude. Participants saw one version during training on the task, and then either the other or the same version during testing. The assumption is that the graphics will encourage a type of processing different from that normally used when performing the task. If so, then following the TAP account, participants for whom the version used during testing matches that used during training were expected to outperform those for whom the versions mismatch.

One problem, though, is the lack of any direct check of whether the type of processing used is actually different between versions. Instead, the issue will be examined indirectly by comparing the pattern of results to what would be predicted by identical-elements theory. To that end, two control groups (one for each version) will be introduced in which participants receive no training on the ATC task.

The main interest here is the direction of transfer—positive or negative—of the performance of the experimental groups relative to the appropriate control group. For example, the groups that see the graphics version of the task during testing (GG and NG) would be compared to the control group that sees the same version (CG). Depending on whether or not the type of processing actually changes between the groups, IE theory makes two different predictions. If processing does not change, then IE predicts that all of the experimental groups will outperform the appropriate control group, but the matched groups (GG and NN) will perform the best. In other words, in terms of performance:

- GG > NG > CG
- NN > GN > CN

However, if processing actually does change, then so does the predicted performance of the experimental groups relative to the control groups. In particular, the matched groups should perform better than the corresponding control group, but the mismatched groups should perform worse, as follows:

- GG > CG > NG
- NN > CN > GN

Alternative hypotheses are posited, as well. For example, the support graphics highlight elements of the task that normally take a large amount of attention and effort. That highlighting may actually serve to reduce the amount of attention that needs to be given to those elements, thus allowing participants to reallocate their resources to less salient portions of the task—the timing of regularly repeating events, for instance. If so, then participants trained with the support graphics should perform better than participants trained without them, as would be demonstrated in a main effect of training graphics.

- GG, GN > NN, NG

On the other hand, rather than benefiting learning, it may be such highlighting actually interferes with it. In particular, the graphics may be so salient that they overshadow other important information. If so, then removing the graphics should hurt the performance of participants who have come to rely on them. Specifically, participants in the GN condition—those who were trained with the graphics but for whom the graphics were removed during testing—should perform worse than those in the other conditions.

- GG, NG, NN > GN

Finally, it may be that the support graphics have inherent advantages or disadvantages to performance, but only in an immediate or at-the-moment sense. For example, the graphics may make the task somewhat easier to play when they are present, but they may have no effect on learning the task. Such would be revealed in a main effect of testing graphics.

- GG, NG > GN, NN

Of course, not all of those hypotheses are mutually exclusive. An asymmetric pattern of results could be found that fits, for example, both the overshadowing and TAP predictions, or that shows both an IE pattern and a graphics main effect.

Operationalizing Performance

The manipulation used in this study is simple: The support graphics either are or are not present during training and during testing. Thus, the experiment's two independent variables are type of training and type of testing.

The dependent variables are somewhat more complicated, though, given the different aspects of performance that can be measured. Generally, the FAA's use of the simulation involves three main variables of interest:

- Safety: the sum of all errors made by the participant
- Efficiency: the sum of the time all aircraft spend getting to their destination; quicker times equate to better performance
- Workload: the sum of the time all newly arrived aircraft spend waiting for acceptance; quicker times equate to better performance

However, those variables will be altered slightly for the current study. In particular, Safety will be broken into three subcomponents: Plane Conflicts, Navigational Conflicts, and Timing Conflicts. Plane Conflicts is being singled out from the other types of errors because it is believed that the support graphics will have their greatest impact on this measure. Navigational Conflicts is being separated from the other types of errors because it consists primarily of procedural errors (e.g., landing at the wrong altitude, getting too close to the border, etc.). Similarly, Timing Conflicts is a combination of error types that are characterized by regular changes in the task (e.g., regular changes of runway direction, and regular appearances of new aircraft). In addition, the Efficiency measure, instead of being based on the total flight time of all aircraft, will now be based on the number of planes successfully exited; as such, it is more of an “Effectiveness” measure.[2]

The measures of interest in the current project, then, are:

- Score: the number of points for all aircraft successfully exited or landed, modified according to how quickly that is done, minus any penalties for errors; serves as a rough composite of the other five DVs
- Plane Conflicts (PlnCon): the number of times the plane-separation rule is violated (i.e., aircraft got too close to each other)
- Navigational Conflicts (NavCon): number of procedural errors (i.e., aircraft getting too close to the border of the airspace, violations to the exiting rules for gates, and airport speed and altitude landing violations)
- Timing Conflicts (TimeCon): number of errors dealing with task components that change regularly (i.e., runway violations, separation violations involving newly arrived aircraft).
- Effectiveness (NumExits): the number of aircraft successfully landed or exited.
- Workload (WaitTime): the total time newly arrived aircraft spend waiting each session; quicker times equate to lower workload

Given that the support graphics were specifically designed to facilitate mental simulation of aircraft position and movement, it is expected that they will have their greatest effect on Plane Conflicts. They may also have secondary effects on

[2] The change is being made because of a problem discovered during the course of the experiment regarding the measurement of aircraft flight time.

Effectiveness (NumExits), because the graphics may allow some greater degree of planning ahead, and possibly because fewer plane conflicts leads to fewer plane crashes—which means more planes are available to exit. Navigational Conflicts may indirectly benefit from the graphics, which make information that is important to exiting and landing (i.e., altitude and speed) more salient. The only likely way in which the graphics will benefit Workload (WaitTime) and Timing Conflicts, however, is if they free up enough attention that participants actually learn more about the task—such as the timing of regular events.

METHOD

Participants

Two hundred fifty-nine introductory psychology student volunteers participated for course credit. The data for 14 participants was dropped because it was apparent that they were not attempting to perform well on or even to play the task. For example, the computer recorded the total number of times a participant changed the direction, speed, and altitude of the aircraft during a given session. If the number of course corrections happened to be zero or even in the tens or twenties (range for the whole sample was 0-155 for the training sessions and 0-176 for the testing sessions), it was assumed that that participant either did not interact with the simulation, or else tried it at first and then quit to let the task run on its own. For 40 other participants, data was lost because of computer, network, or internet difficulties. As a result, total *N* = 205, with 118 male and 87 female young adults randomly assigned to the six conditions.

The sample sizes for the six conditions were not equal. For the training sessions, excluding the control groups (which had no training), total *n* = 145 (71 participants in the graphics training condition, and 74 in the no-graphics condition). For the testing sessions, total *n* = 205; Table 3 shows the sizes for each cell.

Table 3
Sample Sizes for Each Cell

	TEST		
TRAIN	Graphics (G)	No Graphics (N)	Grand Total
Control (C)	37	23	60
Graphics (G)	32	39	71
No Graphics (G)	41	33	74
Grand Total	110	95	205

Task

The main experimental task was a computer simulation called the Air Traffic Scenario Test (ATST). The ATST is a low-fidelity simulation of an air traffic control (ATC) radar screen that is updated every four seconds. The goal is to maintain, as efficiently as possible, separation and control of a varying number of simulated aircraft within the designated airspace.

Based on FAA usage of the ATST simulation, the four training sessions were of increasing difficulty. Session 1 started off with two planes already in flight, with new planes appearing every 40 seconds. With each subsequent session, another plane was added to those initially in flight, and new planes appeared five seconds faster. Each of the two testing sessions started out with eight planes in flight, with new planes appearing every 20 seconds. The increased number of initial planes in flight in the testing sessions, as well as the increased frequency of the appearance of new aircraft, was intended to make those sessions somewhat more difficult than the training sessions. For either type of session, the number of initial aircraft and the frequency of appearance of new aircraft were both based on examples provided by the FAA of easy vs. difficult ATST scenarios. See Table 4 for specifics.

Table 4
Specifications for Each Session of the Air Traffic Control Task

	Session				
Specs	Training 1	Training 2	Training 3	Training 4	Testing 1 & 2
difficulty	progressive	progressive	progressive	progressive	same
time limit	10 min	10 min	10 min	10 min	12 min
instructions	4 min instructions, 2 min practice				
total # planes	16	19	23	28	43
# initial planes	2	3	4	5	8
time b/w new planes	40 sec	35 sec	30 sec	25 sec	20 sec
gate destinations	2-3 each	3 each	3-4 each	4-5 each	8 each
airport destinations	3 each	3-4 each	4 each	4-5 each	8 each
initial plane speed	Varies	varies	varies	varies	varies
refresh rate	4 sec	4 sec	4 sec	4 sec	4 sec

Figure 1 shows a screenshot of the task. Each aircraft was associated with a data block indicating its present speed, altitude, and destination. Speed was represented by a letter (S = slow, M = moderate, F = fast), altitude by a number (1 = low, 2 = middle, 3 = high), and destination by a letter (A-D for gates, E-F for airports). The destination indicated either one of the four gates through which the plane must exit or one of the two airports at which it must land. For each session, the number of aircraft that participants saw was evenly split between the six possible destinations. In addition to the three possible speeds and three possible altitudes, each plane could travel in one of eight possible directions, each corresponding to one of the eight compass points.

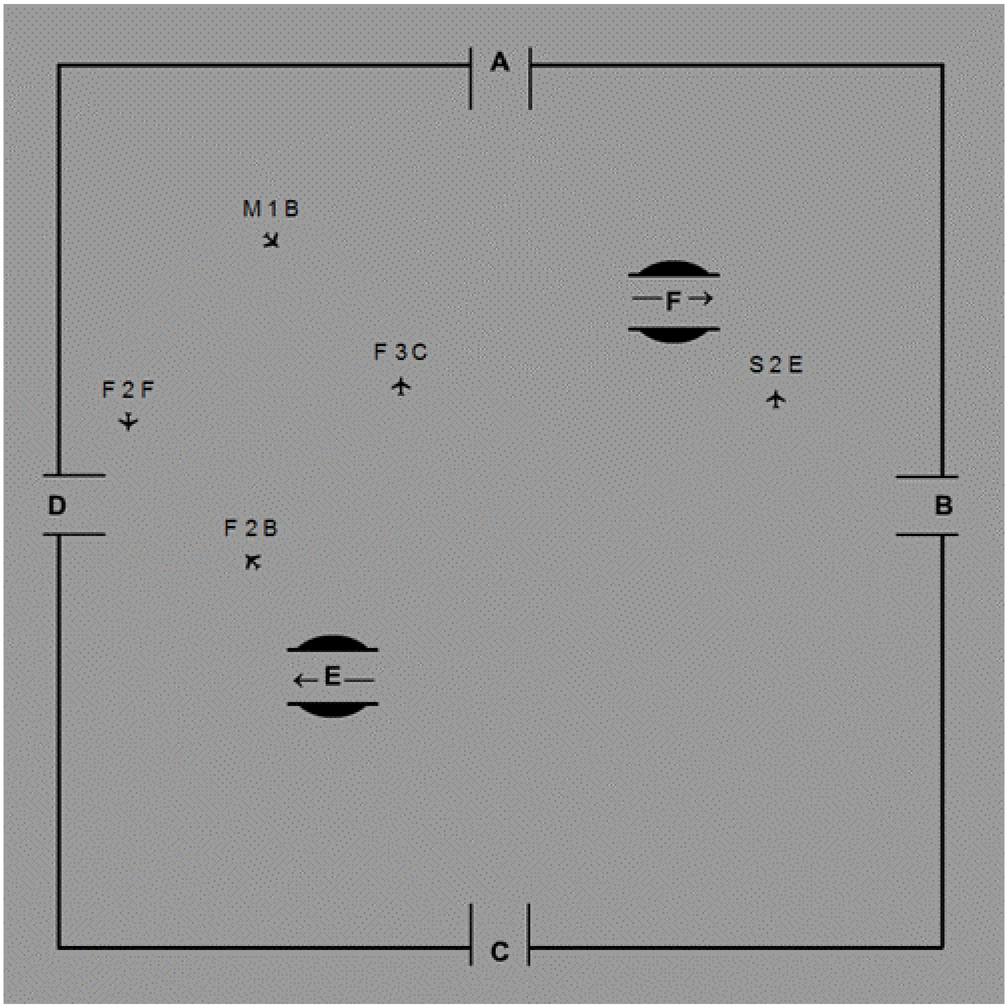

Figure 1
Screenshot of the airspace section of the ATC task.

The participants' job was to keep all planes a certain distance from each other and from the border of the airspace, to guide each one to a specific exit or airport, and to land or exit them at specified altitudes and speeds. Participants maintained separation and control over aircraft in flight by using the computer mouse to click either on the data block to change speed and altitude, or on the plane itself to change heading.

Participants were awarded 15-40 points for each plane successfully landed, depending on how quickly they landed it. For each error (conflict, collision, or broken rule), they were penalized 10 points for every plane involved. The size of the awards and penalties was fairly arbitrary, with the primary criterion being that participants should be able to compensate for a plane collision penalty by landing another plane very quickly. However, the score was mainly for the participants' benefit: It was intended to give them something to work toward. The primary measure, as in the FAA version of the task, was the number of conflicts, collisions, and errors each subject made. The score was also intended to encourage participants to land planes quickly by sending them along a relatively direct route to their destinations, rather than letting them linger along the sides of the borders; the latter would have been an easier strategy, but it would also have ruined the task. Hence, the score was presented throughout each session, so that participants could see that landing planes quickly resulted in a higher score and could compensate for earlier mistakes. Scores and other data were collected by the computer program, which then emailed the information to the experimenter.

The decision-support graphics associated with the task included a trajectory indicator and an altitude indicator for each aircraft on the radar; see Figure 2. The altitude indicator was a filled circle surrounding the aircraft; its color indicated the altitude of the aircraft, allowing quick scanning and comparison of aircraft. The trajectory indicator was a white line pointing out from the nose of the aircraft icon in the direction the aircraft was traveling. Its length changed as a function of aircraft speed, so that the end farthest from the aircraft indicates where that plane would be in four moves, whatever the craft's speed. That length was set according to the following, admittedly subjective, criteria: It had to be long enough to provide useful information about the craft's trajectory; it had to be short enough so that it was not a distraction; and it had to be long enough so that users had time to respond to the crossed indicators of two aircraft before those aircraft came into conflict.

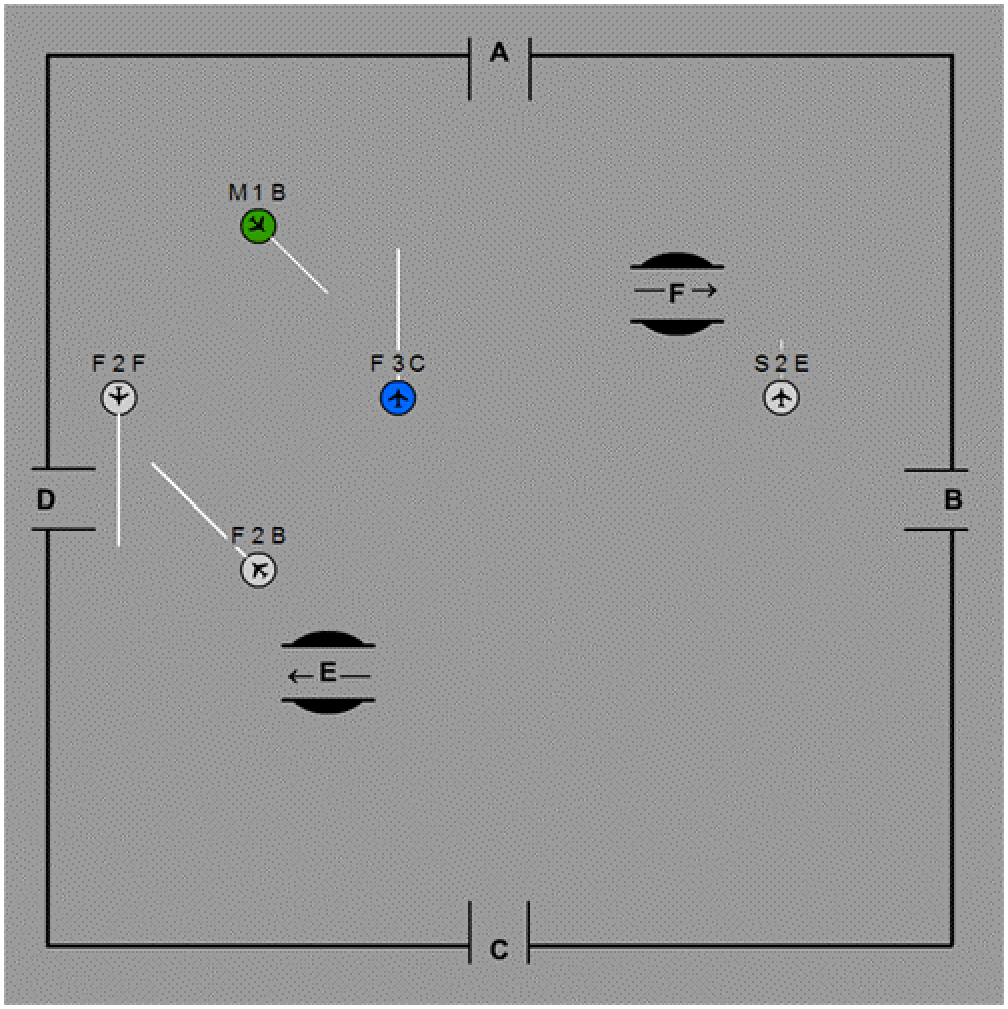

Figure 2
Screenshot of the airspace section of the ATC task showing the support graphics.

Individual-Difference Measures

The following variables were included as covariates to add power to the statistical analyses.

1) The Wonderlic Personnel Test (1992), a measure of cognitive ability. The test contains 50 items, and participants are given 12 minutes to complete it. The Wonderlic Test Manual (1992) states that the instrument's internal consistency reliabilities range from .88 to .94.
2) Nine items from the Computer Usage Survey (CUS; see Young, Broach, & Farmer, 1997) that measure video game experience. These items were summed to produce a single score.

Procedure

Participants were randomly assigned to one of the six conditions: GG, NG, NN, GN, CG, CN. They were initially run in groups in the presence of the experimenter so that they could complete the Wonderlic Personnel Test, which is a timed paper-and-pencil test. The rest of the experiment was accessible online, and so after participants were finished with the Wonderlic, they were allowed to complete the CUS and the ATC task at their own convenience either at home or in a computer lab at the university.[3]

Participants logged onto the online portion of the experiment using individual codes provided by the experimenter. After entering that code, the program asked them to

[3] Note that ecological validity might be a concern here, given that participants may be performing the task in an uncontrolled, possibly distracting environment. However, air traffic control environments are often busy and distracting themselves, and so performing the task in a computer lab may actually be more similar to the real-world environment than would a controlled environment. Also, the focus of the present study is on the support graphics; air traffic control is simply medium to investigate them, and as such is not a primary concern.

indicate where they performing the task (home or computer lab), as well as the size of the computer monitor they were using; that information was then used to introduce two covariates into the data analyses in an attempt to partial out some of the variance caused by the different environments in which the task was performed.

After entering that information, they began the CUS, which was then followed immediately by the ATC task. The ATC task had to be completed in one sitting. To verify that participants actually followed that restriction, the computer recorded how long it took each one to complete the task.

The ATC task contained a brief description of the simulation and instructions on how to use it. After the participants read that information, they were allowed to practice controlling a single aircraft for two minutes. Following that, all non-control participants completed four 10-minute training sessions, with a one-minute break between them. A five-minute break followed training, after which participants completed two testing sessions. Participants in the control conditions skipped the training sessions.

The testing sessions were more difficult than the training sessions in terms of the number of initial aircraft, the rate of appearance of new planes, and so on. Each testing session was also 12 minutes long, with a two-minute break between them. Participants were given two testing sessions rather than one in order to examine whether performance differences attributable to the introduction or removal of the support graphics in the middle of the game were simply because of the novelty of the presence or absence of the graphics. The increased difficulty was an attempt to make the testing task different enough from the training task to be considered a transfer test.

Design

The experiment was a 3 (training graphics) x 2 (testing graphics) between-participants design. A separate was performed on the training data, which was a 2 (training graphics) x 4 (training session) within-subjects design.

RESULTS

Each of the six DVs were repeated-measures variables, recorded for each of the six task sessions (four training sessions followed by two testing sessions). In the following analyses, a number following the variable name is a reference to a particular session. For example, PlnCon1 refers to the number of Plane Conflicts for the first training session; similarly, PlnCon5 and PlnCon6 refer to the number of such conflicts in the last two sessions, which are the testing sessions.

Separate analyses were performed on the training and testing sessions. In addition, each analysis was further broken down into separate analyses for Score and for the set of five remaining DVs (WaitTime, NumExits, PlnCon, NavCon, and TimeCon). Score was run separately because it could serve as a rough but single indicator of performance—a composite of the other variables, essentially. The other five DVs were entered into a multivariate analysis to provide more detail.

Finally, although four covariates (cognitive ability, video game experience, gender, and location of experiment) were measured, only two (gender and location) were included in the final analyses. For the univariate analyses, given that the number of covariates used in ANCOVA is best kept to a minimum (Tabachnick & Fidell, 1996), only the two most reliable covariates—gender and location—were used. Reliability here is based on Tabachnick and Fidell's (1996) definition, which refers to the degree to which covariates can be measured without error. For the multivariate analyses, cognitive ability and video game experience had no significant effect on the DVs, so they were left out.[4] See Appendix A for analyses of the remaining covariates.

Training Data

Composite measure of performance

The training data were analyzed using a 2 (training graphics) x 4 (training session) mixed-design ANCOVA on Score, a composite measure of performance. Adjustment was made for two covariates: gender and the type of location in which the experiment was performed. Type of training (graphics vs. no graphics) and training session (one through four) served as independent variables. SPSS ANOVA with Method 1 adjustment for unequal cell sizes was used to analyze the data.

Evaluation of the assumptions of normality and linearity was acceptable. The cell sizes were unequal, but only by a few participants, so the assumption of homogeneity of variances was supported, as well. Some groups had outliers on Score for the first and last training sessions, which was determined by examining box plots. Transforming the distributions generally made matters worse, so they were left unchanged, and the outliers were instead dealt with by changing the outlying case so that it was either one unit above the highest or one below the lowest non-outlying case (as suggested in Tabachnick & Fidell, 1996). If the outlier already happened to be just one unit away from the highest or lowest non-outlier, it was left unchanged. Changing the outliers in such a way did not

[4] When included with all the covariates, cognitive ability and video game experience approached but did not reach significance: for cognitive ability, $F(5, 191)=2.092$, $p = .07$, $\text{eta}_p^2 = .05$, $\beta = .69$; for video game experience, $F(5, 191)=2.092$, $p = .08$, $\text{eta}_p^2 = .05$, $\beta = .66$.

influence the overall effects found; rather, effect size and power were simply improved by a hundredth of a point or two.

The only significant effect found was that of training session. $F(3, 139) = 6.76$, $p < .001$, with a modest effect, $\text{eta}_p^2 = .14$, and high power, $\beta = .99$. The effect was quadratic, $F(1, 141) = 18.99$, $p < .001$, $\text{eta}_p^2 = .13$, $\beta > .99$, with performance increasing over sessions 1-3, but decreasing on session 4.

Sub-scores

The training data were analyzed using a 2 x 4 mixed-design MANCOVA on WaitTime, NumExits, PlnCon, NavCon, and TimeCon, as well as CourseChanges and AltitudeChanges. Adjustment was made for two covariates: gender and the type of location in which the experiment was performed. Type of training (graphics vs. no graphics) and training session (one through four) served as independent variables. SPSS MANOVA with Method 1 adjustment for unequal cell sizes was used to analyze the data.

The distributions for the five DVs within the six conditions were generally somewhat skewed. There were no multivariate outliers detected; however, within each condition, several of the DVs contained univariate outliers, as determined through an examination of box plots. Square-root transformations were applied to the distributions in an attempt to normalize them and to remove the outliers. For NavCon, the transformation removed the skew and most of the outliers. For the other DVs, though, the transformations either did not help or instead worsened the situation. Hence, because multivariate analyses, even with unequal *n*, are generally robust to violations of normality when there are at least 20 participants per cell (Tabachnick & Fidell, 1996), the distributions for those variables were unchanged. Instead, the outliers were handled as described in the previous section for Score.

The between-groups IV, type of training, was significantly related to the set of DVs, $F(5, 137) = 3.67, p = .004$, $\text{eta}_p^2 = .12$, $\beta = .92$. The same was true for the within-participants IV, training session, $F(15, 127) = 9.61, p < .001$, $\text{eta}_p^2 = .53$, $\beta > .99$.

The effects of the IVs on each covariate-adjusted DV were investigated in a series of univariate analyses. The main effect of training session was primarily on PlnCon, NumExits, and TimeCon. The curves for each were quadratic. For NumExits, performance improved over sessions 2 and 3 but leveled off for session 4. For PlnCon, performance worsened over sessions 2-4, likely reflecting the progressive difficulty of the training sessions. For TimeCon, performance dropped sharply between sessions 1 and 2, but improved just as sharply over the remaining sessions; it should be noted, though, that power was weak for TimeCon, and the effect was found only after adjusting for outliers.

- NumExits: univariate $F(3, 423) = 18.12, p < .001$, $\text{eta}_p^2 = .11$, $\beta > .99$.
- PlnCon: univariate $F(3, 423) = 11.31, p < .001$, $\text{eta}_p^2 = .07$, $\beta > .99$.
- TimeCon: univariate $F(3, 423) = 2.74, p < .043$, $\text{eta}_p^2 = .02$, $\beta = .66$.

For the sub-scores, type of training had a significant effect only on PlnCon, although it did approach significance for NumExits (p = .07). For PlnCon, the participants who had the graphics during training had fewer plane conflicts than did the other participants. The effect and power were both small, however; see Table 5 and Figure 3.

- PlnCon: univariate $F(1, 141) = 6.01$, $p = .015$, $\text{eta}_p^2 = .04$, $\beta = .68$.

Table 5
Means and Standard Deviations for Type of Training x Training Session Interaction on Plane Conflicts

	Type of Training		
Session	Graphics (G)	No Graphics (N)	*Grand Total*
1	4.45 (3.59)	5.74 (3.50)	*5.11 (3.59)*
2	4.87 (3.88)	5.73 (3.50)	*5.31 (3.70)*
3	6.37 (4.45)	7.95 (4.92)	*7.17 (4.75)*
4	8.66 (5.83)	10.32 (6.38)	*9.51 (6.16)*
Grand Total	*6.09 (3.14)*	*7.44 (2.99)*	*6.78 (3.13)*

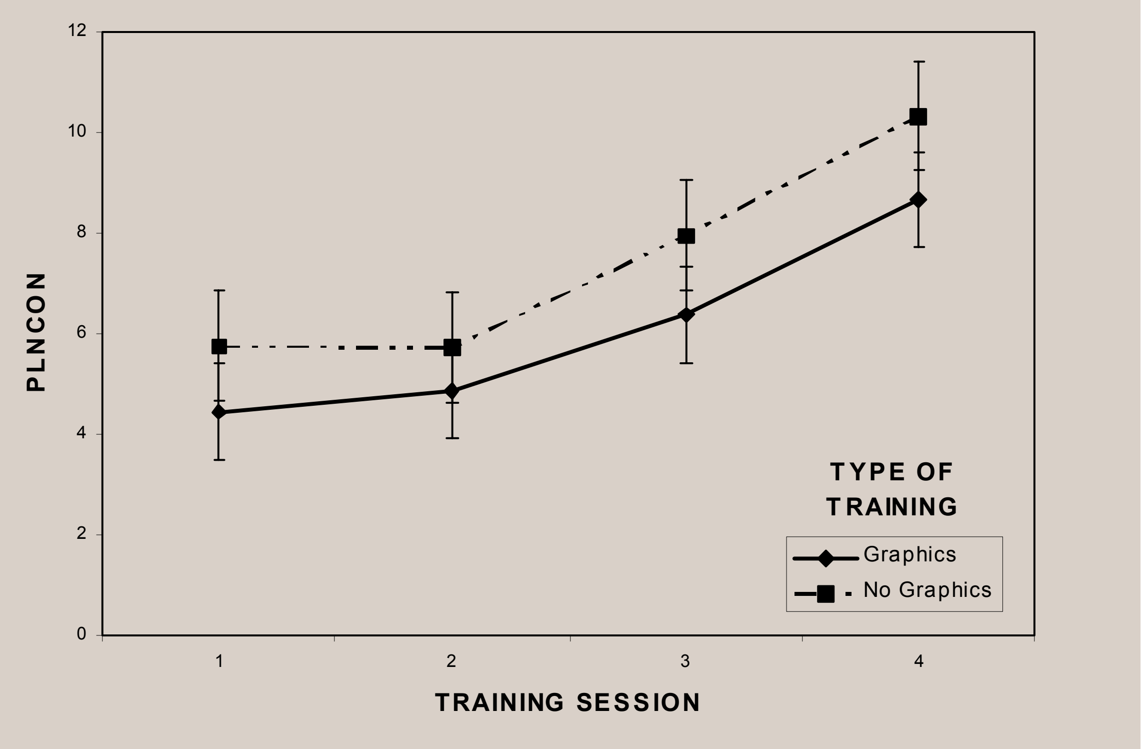

Figure 3.
Type of training main effect for plane conflicts.

Summing up, the support graphics seemed to produce a main effect of graphics, due primarily to their impact on the PlnCon measure. Participants who saw the graphics during training had fewer plane conflicts than those participants who did not see the graphics. The graphics did not appear to aid learning, though, as demonstrated by the lack of a training session x type of training interaction.

Testing Data

Composite measure of performance

The testing data were analyzed using a 3 (training session) x 2 (testing session) ANCOVA on Score, a composite measure of performance. Rather than using the two testing sessions as a within-participants variable, Score5 and Score6 were averaged to produce a single score;[5] see Appendix B for descriptive statistics. Adjustment was made for two covariates: gender and the type of location in which the experiment was performed. Type of training (control vs. graphics vs. no graphics) and type of testing (graphics vs. no graphics) served as independent variables. SPSS ANOVA with Method 1 adjustment for unequal cell sizes was used to analyze the data.

Evaluation of the assumptions of normality and linearity was acceptable. Although the cell sizes were unequal, the discrepancy between largest and smallest was less than 2:1, and the same ratio held for the variances of the different cells, so the assumption of

[5] Using the two testing sessions as a repeated-measures DV was considered; however, the primary reason for including two sessions was to factor out the element of surprise at the introduction or removal of the support graphics. In addition, it was possible that one testing session may have been more difficult than the other, potentially biasing a repeated-measures analysis. More importantly, though, two sessions do not provide enough information to indicate a trend. However, in the interest of thoroughness, an analysis was run in which testing session was included as a repeated-measures DV. Some differences between testing sessions were apparent, but the effect size was generally negligible, and the findings—because of the inability to determine trends—difficult to interpret.

homogeneity of variances held, as well. The distribution for Score was skewed for some groups, and most groups had a couple of univariate outliers, as determined through an examination of box plots. Transforming the distribution generally made matters worse in terms of outliers, so the distribution was left alone, and the outliers were instead dealt with as described in the composite measure section for the training data.

Table 6 shows the correlations among the DVs and covariates. The main effects of type of training and type of testing, as well as the interaction between the IVs, were all significant, though the effect sizes were small. For type of training, $F(2, 197) = 21.22$, $p < .001$, with a modest effect, $\text{eta}_p^2 = .18$, and high power, $\beta > .99$; see Table 7 and Figure 4. However, contrasts showed that the primary difference was between the control groups and the experimental groups, rather than between the graphics and no-graphics experimental groups.

- CG vs. NG + GG: univariate $F(1, 197) = 26.63$, $p < .001$, $\text{eta}_p^2 = .12$, $\beta > .99$.
- CN vs. NN + GN: univariate $F(1, 197) = 16.407$, $p < .001$, $\text{eta}_p^2 = .08$, $\beta = .98$.

Because the control groups, unlike the experimental ones, were untrained, the fact that they performed the worst is not particularly surprising.

Table 6
Correlations between DVs and Covariates

	Score	WaitTime	NumExits	PlnCon	NavCon	TimeCon	Gender	Location
Score	1.000	.338**	.778**	-.538**	-.594**	-.135	-.180**	.215**
	.	.000	.000	.000	.000	.054	.010	.002
WaitTime	.338**	1.000	.133	-.462**	-.418**	.370**	.019	.308**
	.000	.	.056	.000	.000	.000	.785	.000
NumExits	.778**	.133	1.000	.025	-.596**	-.201**	-.283**	.255**
	.000	.056	.	.724	.000	.004	.000	.000
PlnCon	-.538**	-.462**	.025	1.000	.212**	-.253**	-.055	-.020
	.000	.000	.724	.	.002	.000	.430	.772
NavCon	-.594**	-.418**	-.596**	.212**	1.000	-.253**	.170**	-.206**
	.000	.000	.000	.002	.	.000	.015	.003
TimeCon	-.135	.370**	-.201**	-.253**	-.253**	1.000	-.074	-.001
	.054	.000	.004	.000	.000	.	.290	.983
Gender	-.180**	.019	-.283**	-.055	.170*	-.074	1.000	-.031
	.010	.785	.000	.430	.015	.290	.	.659

(top value = correlation, bottom value = significance)
** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Table 7
Means and Standard Deviations for Type of Training and Type of Testing Main Effects and Interaction on Score

	Type of Testing		
Type of Training	Graphics (G)	No Graphics (N)	*Grand Total*
Control (C)	567.73 (134.64)	470.24 (168.46)	*530.36 (154.69)*
Graphics (G)	779.63 (164.28)	608.96 (217.03)	*685.88 (211.74)*
No Graphics (N)	750.06 (199.06)	731.95 (151.38)	*741.99 (178.44)*
Grand Total	*697.33 (192.20)*	*618.10 (208.37)*	*660.61 (203.26)*

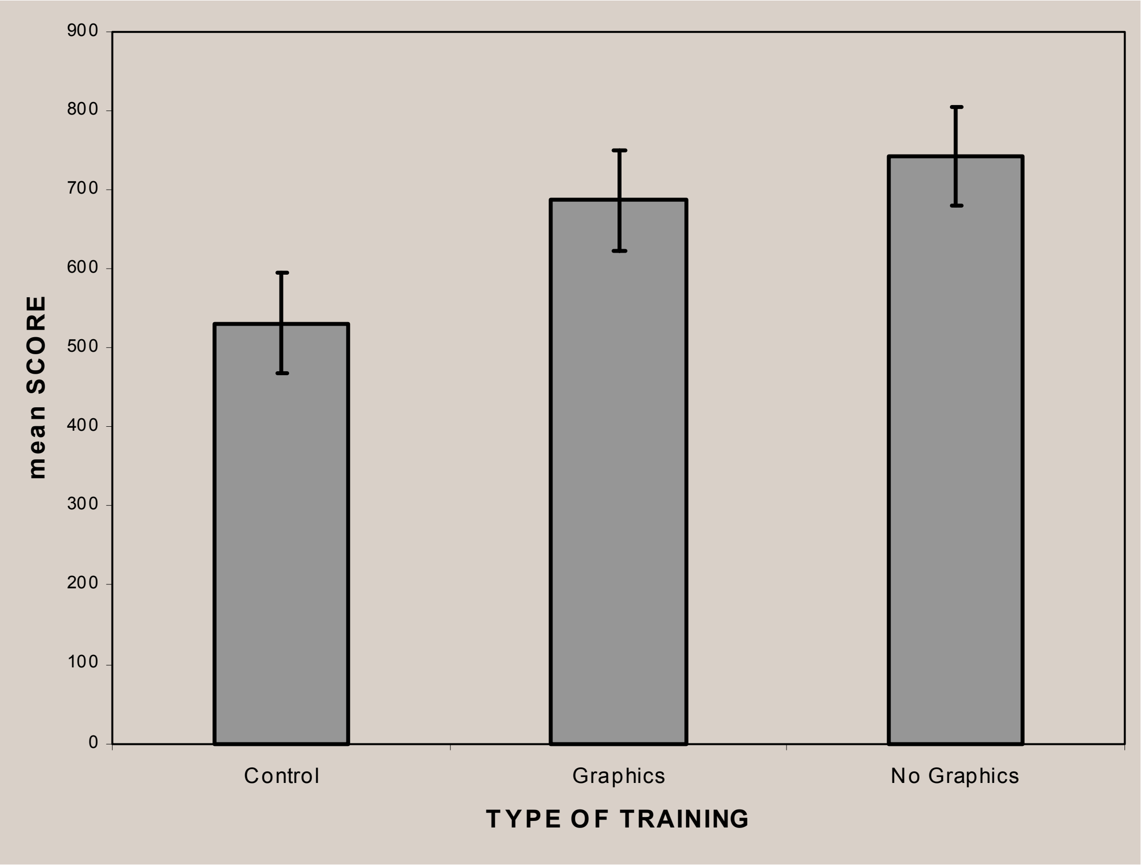

Figure 4.
Type of training main effect for Score. The trained groups outperformed the untrained groups.

Type of testing was also significant, $F(1, 197) = 17.088$, $p < .001$, with a small effect, $\text{eta}_p^2 = .08$, and high power, $\beta = .98$. The advantage was for the groups that received the support graphics during testing; see Figure 5.

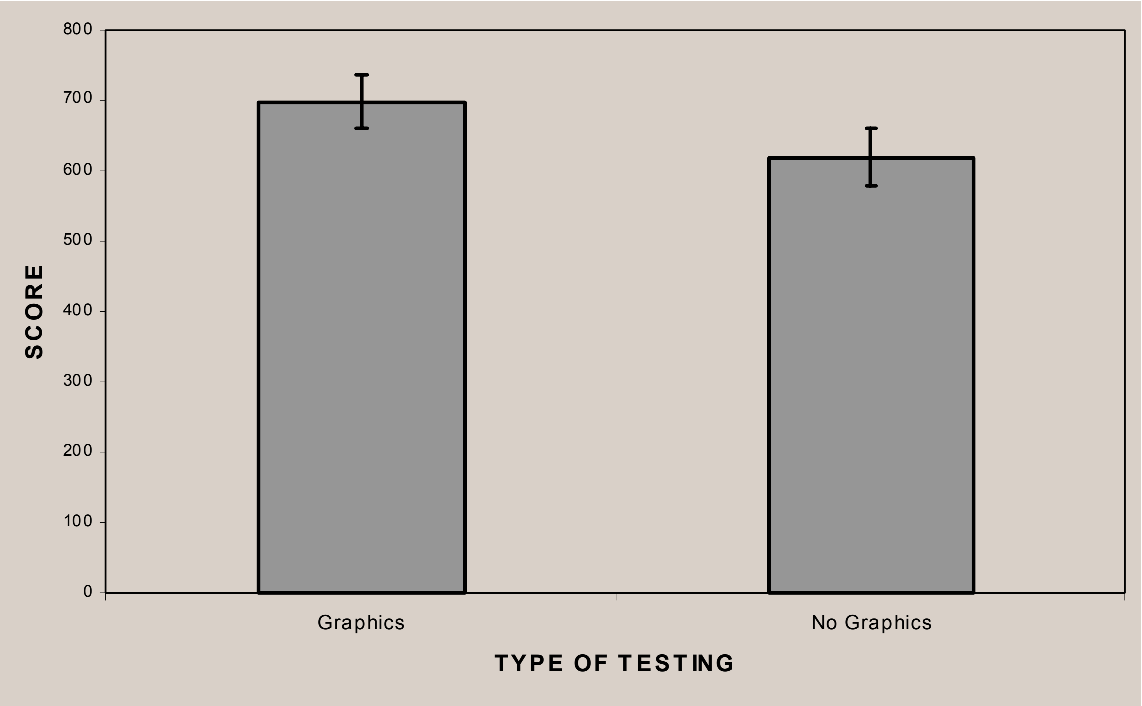

Figure 5.
Type of testing main effect for Score. The graphics conditions outperformed the no-graphics conditions.

The training x testing interaction was significant, $F(2, 197) = 3.83$, $p = .023$, with a very small effect, $\text{eta}_p^2 = .04$, which was evident despite low power, $\beta = .69$. Though the effect was small, the interaction was evident regardless of the covariates and throughout the changes made during data screening. The difference was primarily between the GN group and the other groups; $F(1,197) = 22.75$, $p < .001$, $\text{eta}_p^2 = .10$, $\beta > .99$; see Figure 6. The GN group (the group that saw the graphics during training but not during testing) performed significantly worse than the other experimental groups—particularly the GG group, which saw the graphics during both training and testing.

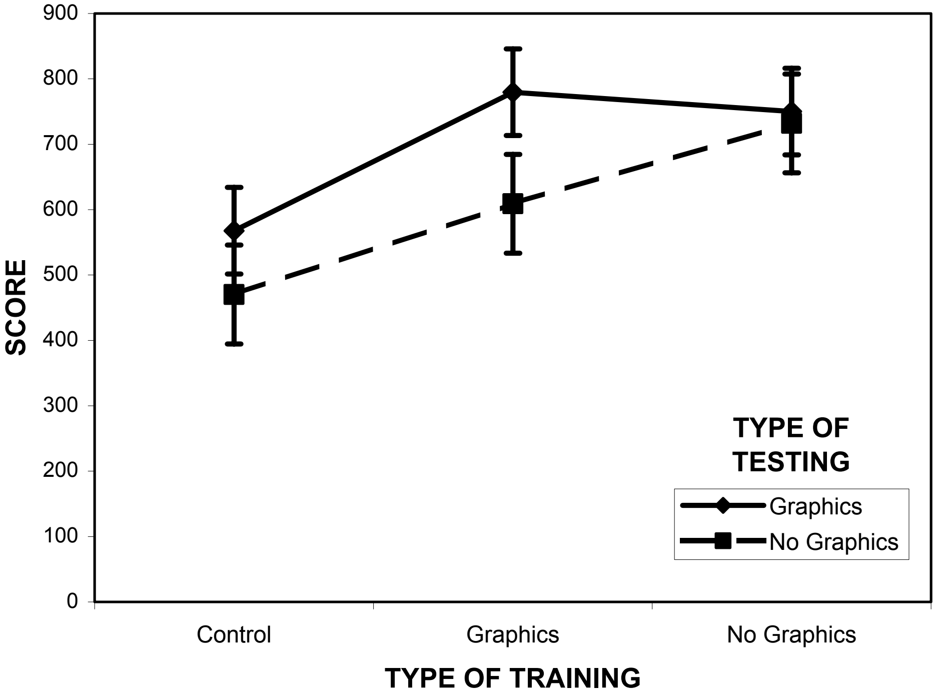

Figure 6.
Type of training x type of testing interaction for Score. The GN group performed worse than the other experimental groups.

Summing up, presenting the support graphics during training had no effect on performance during the testing sessions, as determined by participants' overall Score. However, presenting them during training did provide a performance advantage on the training sessions. As well, a significant, though weak, training x testing interaction was evident. Contrasts show that the interaction was primarily between GN and the other experimental groups; in other words, showing the graphics during training but then taking them away during testing resulted in the worst performance among the experimental groups.

Sub-scores

The testing data were analyzed using a 3 (training sessions) x 2 (testing session) between-participants MANCOVA on WaitTime[6], NumExits, PlnCon, NavCon, and TimeCon, as well as CourseChanges and AltitudeChanges. (As with Score, each DV was averaged across the two sessions to produce a single measure; see Appendix B for descriptive statistics.) Adjustment was made for two covariates: gender and the type of location in which the experiment was performed. Type of testing (graphics vs. no graphics) and type of training (control vs. graphics vs. no graphics), entered in that order, served as independent variables. SPSS MANOVA with Method 1 adjustment for unequal cell sizes was used to analyze the data.

The distributions for the five averaged DVs within the six conditions were generally skewed. There were no multivariate outliers detected; however, within each condition, several of the DVs contained univariate outliers, as determined via box plots. NavCon was generally the most problematic in this regard, as was NumExits. Given the problems with skewness and outliers, square-root and logarithmic transformations were applied to the distributions in an attempt to normalize them and to remove the outliers. For NavCon, a logarithmic transformation removed the skew and all but three of the outliers. For the other DVs, though, the transformations either did not help or instead worsened

[6] Two months into the experiment, a problem was discovered with the measurement of the WaitTime variable. The problem was fixed and so did not affect any of the experimental groups, just the control groups. To salvage the data, the score for each participant in each of the two control groups was set to the mean of the two corresponding experimental groups (i.e., CN = mean of NN + GN, CG = mean of GG + NG). Doing so would of course make it impossible to determine support for TAP, because such support depends on the relative performance of the control groups to the experimental groups, but it was better than the alternative of losing the data altogether. At any rate, the analyses were performed both with and without the changed data, and the change simply changed some of the significance levels and effect sizes by a hundredth of a point or two.

the situation, and so they were handled as described in the sub-scores section for the training data.

Robustness to violations of the homogeneity of variance-covariance matrices was a concern, however, given the unequal sample sizes and because Box's *M* test was significant at p < .001. An inspection of the DVs within each cell showed that the larger cells often but not always had larger variances than the smaller cells, which could make the significance tests too liberal. Hence, Pillai's criterion was used in the following analyses because of its greater robustness to unequal cell sizes and to violations of the homogeneity of variance-covariance matrices.

Each of the IVs and the interaction between them had a significant multivariate effect on the set of DVs. For type of training (control vs. graphics vs. no graphics), *F*(14, 384) = 11.96, *p* < .001, with a large effect, $\text{eta}_p^2 = .30$; power was high at β > .99. For type of testing (graphics vs. no graphics), *F*(7, 191) = 4.35, *p* < .001, with a modest effect, $\text{eta}_p^2 = .14$; β = .99. For the testing x training interaction, *F*(14, 384) = 2.15, *p* = .009, with a small effect, $\text{eta}_p^2 = .07$; β = .97.

Type of training significantly affected all DVs except for Plane Conflicts (PlnCon);[7] see Table 8 and Figures 7-9. In all cases, however, contrasts showed that the difference was between the control groups and the experimental groups, rather than between the graphics and no-graphics experimental groups. Because the control groups, unlike the

[7] WaitTime also did not contribute to the main effect for type of training. However, the adjustments made to the WaitTime scores for the control groups (see earlier footnote) eliminated any differences between them and the experimental groups, and such differences were the primary reason for the training main effect.

experimental ones, were untrained, the fact that they performed the worst is not particularly surprising.

- NumExits: univariate $F(2, 197) = 34.21$, $p < .001$, $\text{eta}_p^2 = .26$, $\beta > .99$.
- NavCon: univariate $F(2, 197) = 7.27$, $p = .001$, $\text{eta}_p^2 = .07$, $\beta = .93$.
- TimeCon: univariate $F(2, 197) = 10.18$, $p < .001$, $\text{eta}_p^2 = .09$, $\beta = .99$.

Table 8
Means and Standard Deviations of NumExits, NavCon, and TimeCon by Type of Training

	Type of Training		
	Control	No Graphics	Graphics
NumExits	4.65 (3.17)	10.55 (4.16)	9.86 (5.25)
NavCon	8.03 (4.88)	5.11 (3.55)	6.65 (5.91)
TimeCon	16.18 (5.62)	12.47 (4.03)	13.22 (4.76)

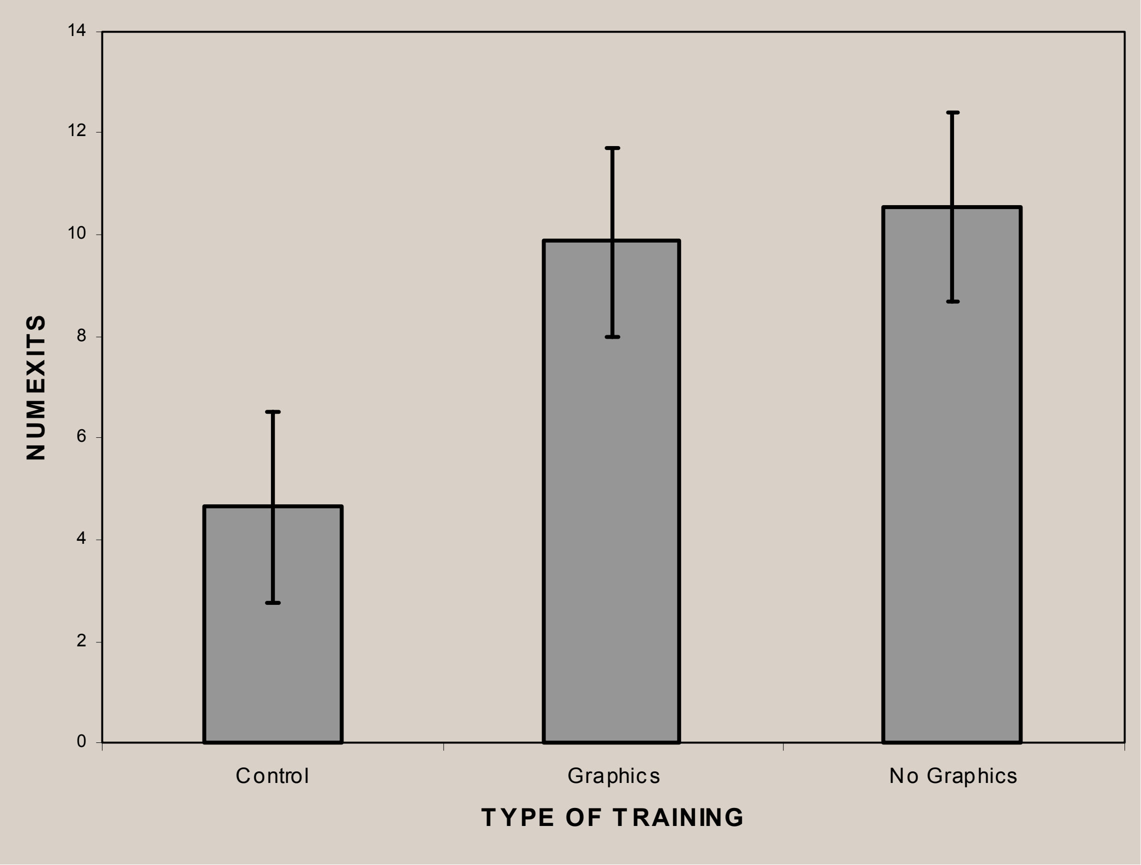

Figure 7.
Type of Training Main Effect for NumExits

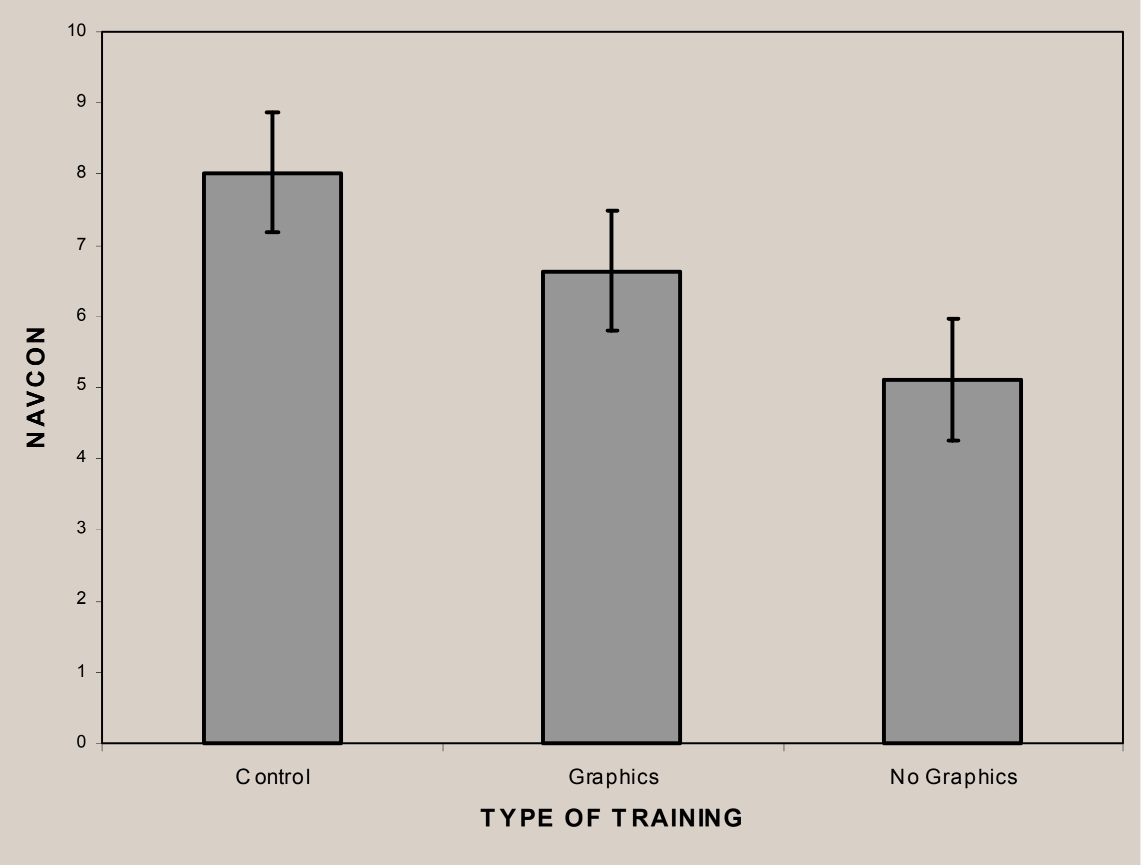

Figure 8.
Type of Training Main Effect for NavCon

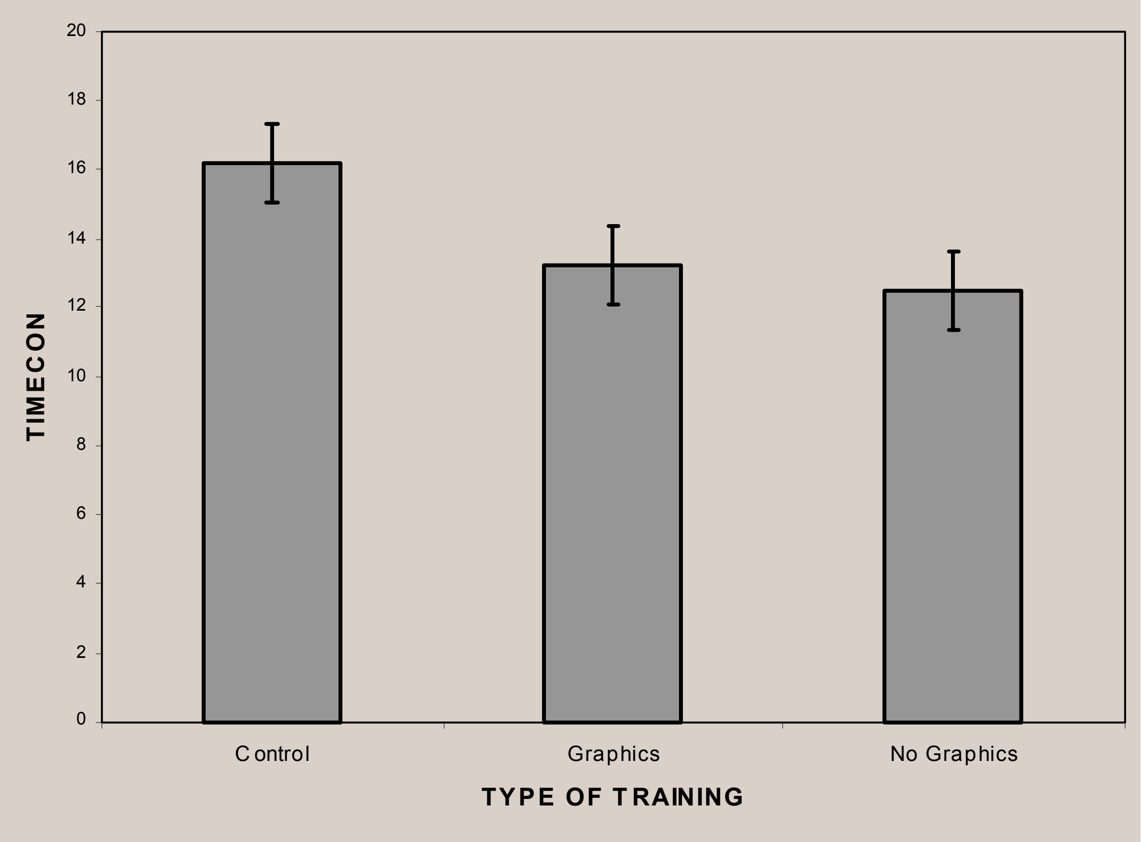

Figure 9.
Type of Training Main Effect for TimeCon

Type of testing affected primarily PlnCon, NumExits, and Navcon—essentially, any DV not directly involving timing issues; see Table 9 and Figures 10-12. In all cases, the differences were attributable to better performance in the groups that had the support graphics during testing.

- NumExits: univariate $F(1, 197) = 12.42, p = .000$, $\text{eta}_p^2 = .06$, $\beta = .94$.
- PlnCon: univariate $F(1, 197) = 9.36, p = .003$, $\text{eta}_p^2 = .05$, $\beta = .86$.
- NavCon: univariate $F(1, 197) = 12.35, p = .001$, $\text{eta}_p^2 = .06$, $\beta = .94$.

Table 9
Means and Standard Deviations for NumExits, PlnCon, and NavCon by Type of Testing

	Type of Testing	
	Graphics	No Graphics
NumExits	9.16 (4.92)	7.93 (5.05)
PlnCon	24.92 (10.72)	29.85 (12.66)
NavCon	5.66 (3.75)	7.46 (5.98)

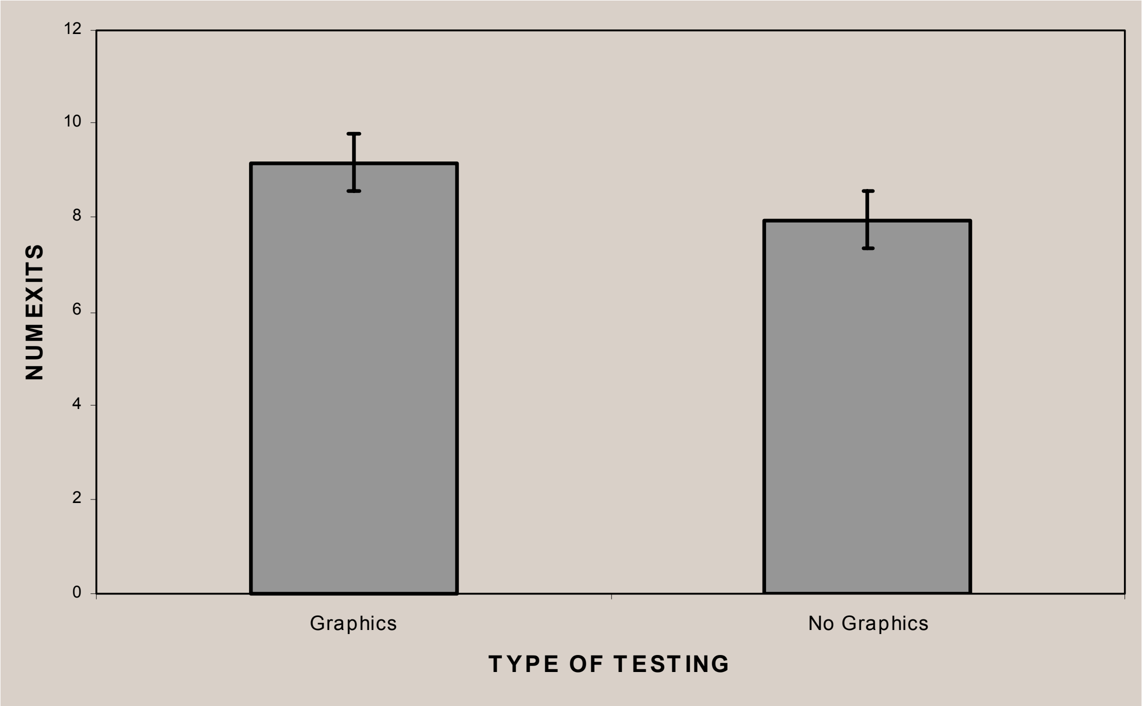

Figure 10.
Type of training main effect for NumExits.

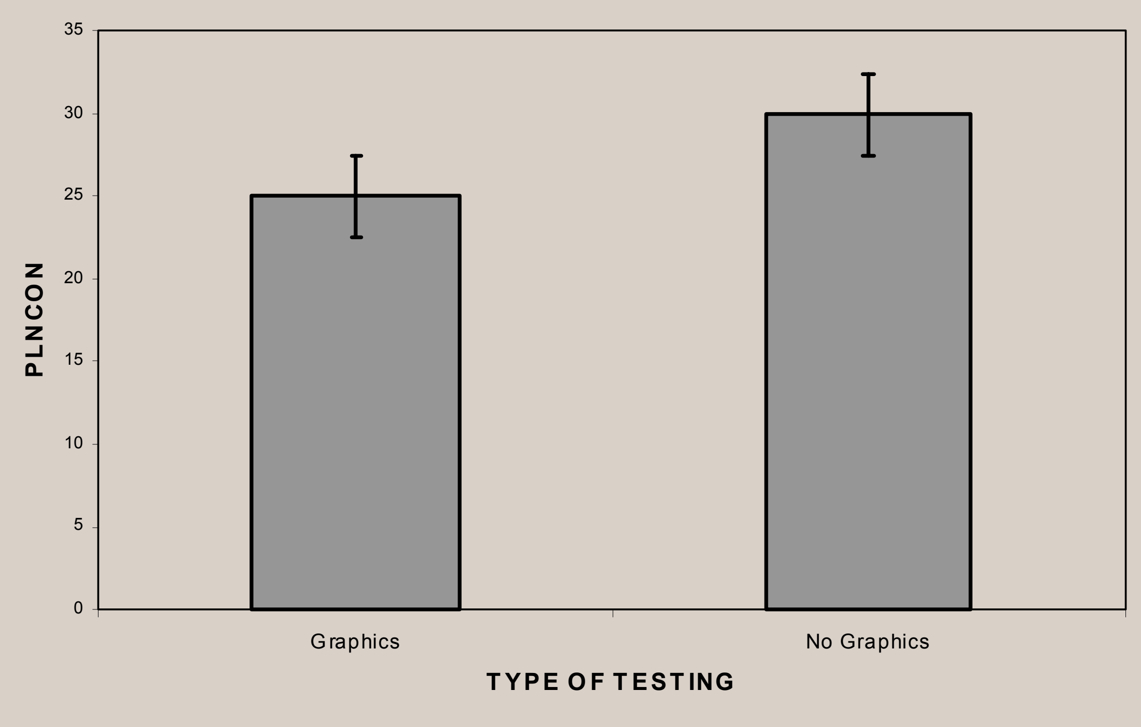

Figure 11.
Type of training main effect for PlnCon.

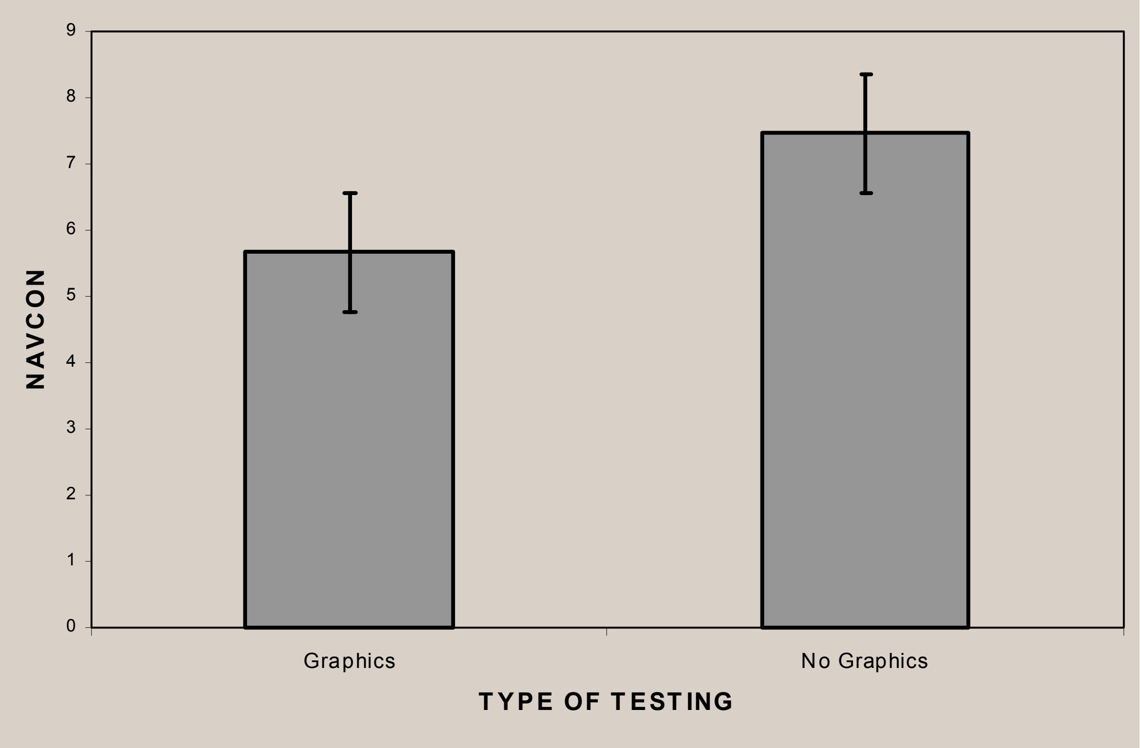

Figure 12.
Type of training main effect for NavCon.

The multivariate interaction between type of training and type of testing was primarily on NumExits and NavCon; see Tables 10-11 and Figures 13-14. However, the effect sizes were very small.

- NumExits: univariate $F(2, 197) = 4.92$, $p = .008$, $\text{eta}_p^2 = .05$, $\beta > .80$.
- NavCon: univariate $F(2, 197) = 6.33$, $p = .002$, $\text{eta}_p^2 = .06$, $\beta = .90$.

In all cases, the interaction was primarily attributable to the GN group (the group that saw the graphics during training but not during testing) performing significantly worse than the other experimental groups—particularly, the GG group, which saw the graphics during both training and testing. The contrasts between GN and the other non-control groups for each significant DV were as follows:

- NumExits: univariate $F(1, 197) = 17.89$, $p < .001$, $\text{eta}_p^2 = .08$, $\beta = .99$.
- NavCon: univariate $F(1, 197) = 23.22$, $p < .001$, $\text{eta}_p^2 = .10$, $\beta > .99$.

Table 10
Means and Standard Deviations for NumExits by Type of Training x Type of Testing

	Type of Testing		
Type of Training	Graphics	No Graphics	*Grand Total*
Control	5.32 (2.92)	3.57 (3.33)	*4.65 (3.17)*
Graphics	11.91 (3.83)	8.18 (5.69)	*9.86 (5.25)*
No Graphics	10.46 (5.06)	10.67 (2.74)	*10.55 (4.16)*
Grand Total	*9.15 (4.92)*	*7.93 (5.05)*	*8.59 (5.01)*

Table 11
Means and Standard Deviations for NavCon by Type of Training x Type of Testing

	Type of Testing		
Type of Training	Graphics	No Graphics	*Grand Total*
Control	6.86 (4.02)	9.89 (5.63)	*8.03 (4.88)*
Graphics	4.48 (2.70)	8.42 (7.16)	*6.65 (5.91)*
No Graphics	5.49 (3.96)	4.64 (2.97)	*5.11 (3.55)*
Grand Total	*5.66 (3.75)*	*7.46 (5.98)*	*6.50 (4.98)*

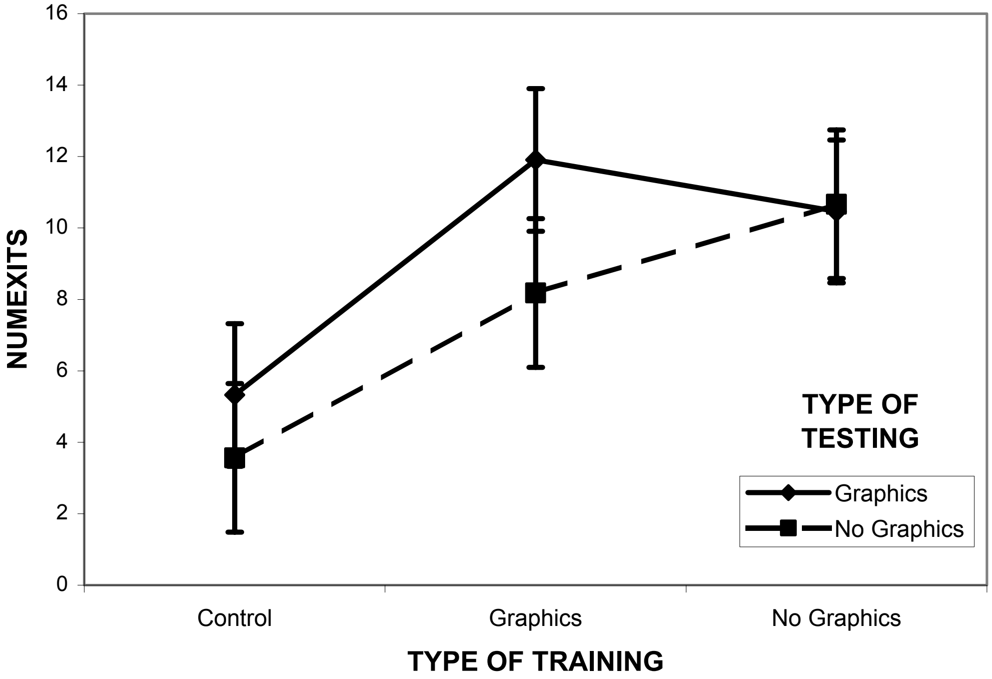

Figure 13.
Type of training x type of testing interaction for NumExits.

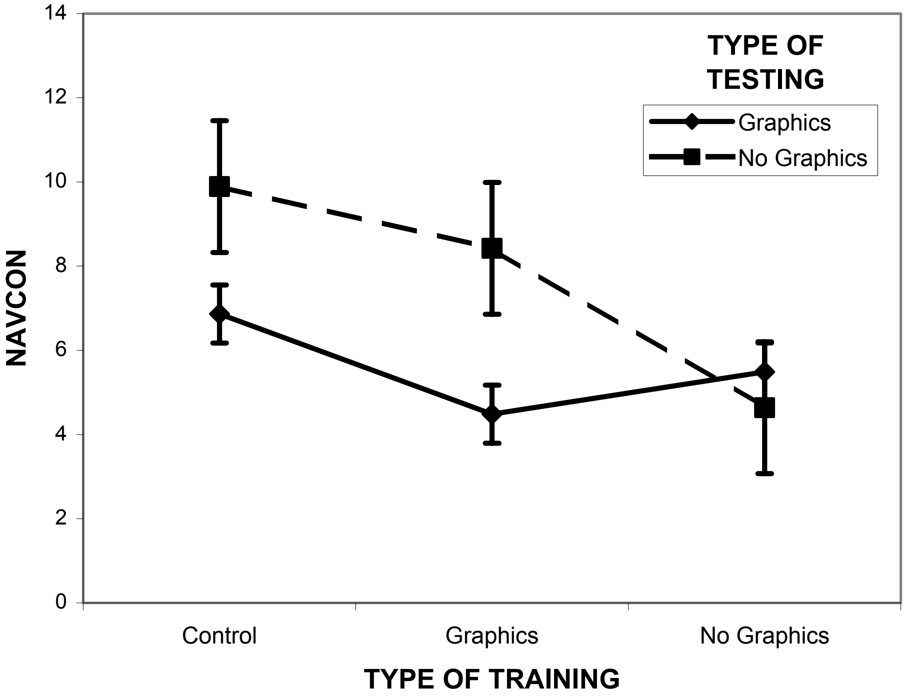

Figure 14.
Type of training x type of testing interaction for NavCon.

Summing up, presenting the support graphics during training had no effect on performance during the testing sessions. However, presenting them during training did provide a performance advantage during the training sessions, primarily because of improvements in Effectiveness (NumExits), PlnCon, and NavCon. Finally, a significant, though weak, training x testing interaction was evident, mostly because of the effect of the graphics on Effectiveness (NumExits) and NavCon. Contrasts show that the interaction was primarily between GN and the other experimental groups; in other words, the group that had the graphics during training but not during testing performed worse than the other groups, not including the controls.

CONCLUSION

The Graphics Advantage

This study found an overall beneficial effect of the decision-support graphics on performance. Whenever they were present, regardless of training, overall performance improved (hypothesis 6). The effect was small but consistent, appearing regardless of the changes made during data screening. In much human-factors research, even small effects are desirable, given that human lives are often at stake. In this case, the support graphics helped the participants get a few more aircraft safely to their destinations.

That finding is important for several reasons. First, it helps fill the gap in the literature regarding empirical studies of the use of graphics to improve training and decision making in complex tasks. Research in the area of decision support generally focuses on more-complex tools than the graphics used here, and because of the complexity of those tools, it is difficult to say what role the graphics (or a particular graphic) played in any performance improvements. In the current study, though, the support tool was simple enough that conclusions concerning the causes of any performance changes could be limited to the set of graphics involved. As a result, the research provides support for the implementation of such tools in complex systems. More importantly, although the graphics used were relatively simple, the research also provides a starting point for further investigating why such tools are helpful and which aspects of them are particularly beneficial—which can lead to a refinement and perhaps elaboration of decision-support systems.

Second, the research already suggests reasons why the graphics might have improved performance. In particular, the graphics were specifically designed to represent explicitly information that normally is represented mentally in air traffic control and similar tasks. It was believed that in doing so, the graphics would facilitate at least part of the mental simulation required to anticipate the trajectories of several aircraft. More specifically, it was argued that they would transform some of the conceptual processing of mental simulation into perceptual processing, thus capitalizing on human perceptual capabilities and improving performance in high-workload situations—in this case, situations characterized by numerous aircraft. The graphics actually did seem to improve performance, which supports the conclusion that using external simulations or representations to facilitate the formation and use of mental ones is a successful strategy. That conclusion also points out directions for future research on the role, use, and elaboration of such external simulations, as well as on the possible consequences of using them. It is possible that users could become overreliant on them, for example, or that they could impair situation awareness. A more theoretical direction in which to extend this research would be to examine the assumption that such external simulations work through the transformation of conceptual into perceptual processing.

The use of multiple performance measures also helped to pinpoint the effects of the support graphics. In other words, measuring several aspects of performance, rather than a single composite measure, made it possible to trace the effects of the support graphics more precisely. In particular, performance only improved on the elements of the task that were primarily spatial or perceptual in nature: Plane Conflicts, Navigational Conflicts, and Effectiveness (NumExits). The graphics did not affect Workload (WaitTime) or Timing Conflicts at all, both of which involved regularly occurring changes to the task. As mentioned earlier, given that the graphics were specifically designed to facilitate mental simulation of aircraft position and movement, it was expected that they would have the greatest effect on Plane Conflicts; Effectiveness (NumExits) and Navigational Conflicts were also expected to benefit at least indirectly from the graphics because of the relatively greater salience of relevant information. However, the only way in which Workload (WaitTime) and Timing Conflicts were expected to benefit was if the graphics freed up enough attention that participants could actually learn more about the task—specifically, the timing of regular events. Because the graphics had no effect on either of those measures, it would seem that they were not successful in creating any sort of attentional slack—at least, not enough.

Combining that conclusion with the finding that providing the graphics during training did not result in any overall performance improvement during testing, and it would seem that decision-support graphics benefit performance only in an immediate sense—as predicted in hypothesis 6. More specifically, they provided immediate performance benefits on the task by highlighting imminent plane conflicts and procedural errors. The high number of planes successfully exited by participants viewing the testing graphics may also be an indication that those participants were better able to plan ahead to a small degree. However, in doing so, the graphics did not free up enough extra attention for participants to learn less-salient elements of the tasks, such as timing.

Using performance as a crude indicator of learning, then, it appears that the support graphics did not help learning at all; rather, they were useful only when present.

Evidence of Overshadowing

Although the graphics did not have an overall effect on testing performance when collapsed across testing conditions, when training condition was considered, the results indicated they were both an asset and a liability during training. As mentioned, they served as a performance aid when performing the task, and participants who saw them throughout training and testing generally had the best scores. However, as was predicted in hypothesis 5, that aid evidently became a crutch, because when the graphics were taken away during testing from participants who were trained on them (the GN group), performance dropped significantly below all other groups (except the controls). That was the case for NavCon and NumExits, as well as for the composite variable, Score. Those variables demonstrated an interaction in which presenting the graphics during training improved performance when the graphics were retained during testing, but worsened performance when the graphics were removed during testing. In contrast, participants who did not see the support graphics during training performed fairly equivalently during testing, regardless of whether they saw the graphics at that time. The suggestion is that, for NavCon and NumExits, at least, the support graphics apparently improved performance, as demonstrated in the high scores of the GG group; however, that improvement came at the cost of learning, as demonstrated by the low scores of the GN group.

Given that the graphics seemed to provide immediate performance benefits in terms of maintaining plane separation and following navigational rules, but that they apparently provided no learning benefits, a few explanations are possible for the performance drop in the GN group. First, it may be that participants trained without the graphics learned certain strategies for handling the aircraft, such as flying them close to the border rather than straight through the middle of the airspace. As demonstrated by the poor scores overall, such strategies, if used, were not very effective, and the performance advantage provided by the graphics more than made up for the absence of such strategies in graphically trained participants—until the graphics were removed.

Alternatively, in the absence of the graphics, participants may have been forced to develop a better situation awareness of the tasks—in other words, to maintain a general idea of where the aircraft were, which ones were near each other and near exits, whether the ones near exits were at the right altitude and speed, and so on. Their improved situation awareness may have led to a better ability to anticipate potential aircraft conflicts, as well as better prospective memory for such things as remembering to change aircraft altitude to accommodate the exit rules. The graphics, however, somewhat alleviate the need to maintain situation awareness, at least as it relates to aircraft separation and navigation; for example, they provide perceptual cues for impending aircraft conflicts, meaning participants need not devote much attention (relatively speaking) to such conflicts until the cues indicate such a conflict is about to happen. In other words, the graphics were actually designed to reduce the need for mental simulation, an aspect of situation awareness. They perhaps succeeded in that respect, but in doing so, they kept participants from developing that ability naturally—which was not an issue until the aid was removed. Future research could examine that possibility in more detail by using Endsley's (1995a) Situation Awareness Global Assessment Technique (SAGAT). The basic procedure in SAGAT involves freezing or blanking the computer screen at intervals, and then testing participants' memory for the state of various elements in the task immediately preceding the freeze. If participants using the graphics demonstrated worse memory than those not using them, that would be evidence that the graphics aid performance at the cost of situation awareness.

However, given that number of exits and number of navigational errors (violations of exiting rules and so on) were the only measures showing an overshadowing effect, a simpler explanation may be more likely. It is possible that reading and understanding the alphanumeric symbols indicating speed and altitude takes some experience. The participants trained with the graphics never needed to use those symbols, so when the graphics were removed, their performance suffered as they tried to get used to the alphanumerics. Or, similarly but more likely, it may be that participants trained with the graphics learned to associate colors and length of trajectory lines with exits—blue and long for gates, green and short for airports; participants trained without the graphics, though, learned the associations in terms of altitude and speed—high and fast (or, alphanumerically, 3 and F) for gates, low and slow (1 and S) for airports. When the colors and trajectory lines were removed, then, the graphically trained participants were forced to learn new associations. That process likely would have resulted in several exiting violations (the NavCon measure). In turn, because such violations result in crashed planes, they reduce the number of aircraft that can be successfully exited (the NumExits measure). If so, then the overshadowing effect found for NavCon and NumExits was primarily the result of the graphically trained participants having to relearn the exiting rules.

At any rate, in the area of decision support, the results indicate that support graphics may be beneficial to include during training if those graphics will always be present on the job: Participants who had access to the support graphics throughout the experiment performed the best, though the difference was not significant. If there are times when the graphics might not be available, though, it would seem prudent *not* to present them during training because of potential overshadowing effects. However, it is quite likely that the overshadowing effect would disappear over time. Hence, it would be worthwhile in future projects to have more than two testing sessions to see if that is actually the case and how long it takes for the effect to disappear. Four sessions would likely be enough to see the overshadowing effect at least begin to dissipate. Another issue to examine is whether it would be beneficial to present the graphics only during part of training—just the second half, for example. It may be that partially exposing trainees to the graphics can both eliminate the overshadowing effect and preserve the performance advantage demonstrated by the participants who saw the support graphics throughout the experiment.

Some Support for IE

The trained experimental groups performed better during testing than the untrained control groups, which unfortunately means that hypothesis 1, transfer-appropriate processing, was not supported. Such support would have taken the form of some trained groups (GN and NG) performing worse than—or at least equal to—the corresponding untrained group (CN and CG, respectively). At least two conclusions are possible here. First, it may be that TAP cannot be extended to naturalistic cognitive tasks. TAP is generally used to explain differences in performance on implicit vs. explicit memory tests. However, it has been successfully applied to somewhat more-real-world tasks such as memory for television news, as well as areas outside of memory performance—namely, physical performance. Given that, one of the purposes of the current study was to see if the theory can predict performance on skilled cognitive tasks. Perhaps it can, but the results described here do not support such an extension of the theory.

On the other hand, though, it may be that, contrary to expectations, the support graphics simply did not change the type of processing that the participants used—at least, not to a sufficient degree. If so, then the task was not an adequate test of TAP, and in the future more direct means should be taken to ensure that different versions of the task actually do require participants to use different types of processing. For example, in previous tests of TAP, experimenters have been able to manipulate explicitly the type of processing participants used to encode a particular stimulus. When testing memory for lists of words, say, some participants may be asked to generate the words themselves, whereas others are asked to read the words from a list. And in the area of physical performance, participants can be encouraged to prepare for a task visually by asking them to rehearse it mentally. Finding an analogous method of explicitly manipulating the type of processing participants use to interact with the ATC simulation would be the most important improvement to the current study.

Although no evidence was found to support the extension of TAP to naturalistic cognitive tasks, some very weak support was found for the more general identical-elements theory, hypothesis 2. Even weak support is important here, though, given the apparent scarcity of empirical research on a theory generally assumed to be fact. In particular, participants' scores on Effectiveness (NumExits) and Procedural Errors (NavCon), as well as their overall Score, appeared to conform to what would be predicted by IE. However, they conformed only somewhat, because not all the differences were significant. In other words, the following pattern was observed, as can be seen in Figures 6 and 9:

1) GG > NG > CG
2) NN > GN > CN

Statistically speaking, though, the only real difference was between the GN and NN groups; the NG and GG groups did not differ. However, that the pattern of results for Score, NumExits, and Navcon matched, albeit nonsignificantly, what would be predicted by IE suggests that perhaps the asymmetric argument is the best fit here. As mentioned, some of the various hypotheses under consideration were not mutually exclusive with either TAP or IE, meaning the pattern of results could be asymmetric if more than one of them happened to be at work. For example, if the manipulations produced both IE and overshadowing effects, then the difference between the GN and NN groups would be expected to increase compared to IE alone. Similarly, if there was also a graphics main effect, the gap between the GG and NG groups would be expected to decrease compared to IE alone, potentially eliminating any statistical differences between them. Both of those effects were actually found, suggesting that an IE effect was present but was modified by the nature of the graphics.

The results are interesting because they suggest how IE effects are influenced by the role of the identical elements in question. Because IE is a general theory of performance and transfer, discussions concerning it tend also to be general, meaning that particulars on what comprises an “element” are hard to come by. Goldstein and Ford (2002) provide one of the more-detailed discussions, differentiating between stimulus elements and response elements. However, they do not explicitly discuss how differences in the importance of the elements might impact IE effects. In other words, how might IE effects be influenced by stimulus elements that are central to the particular task, rather than simply environmental or peripheral to it?

The current study suggests that IE theory is susceptible to the role of the elements in question. In this case, the elements were stimuli (computer graphics) that either were or were not present during the task. However, the stimuli were central to the task in that they represented important factors required to perform it—namely, an aircraft’s trajectory and relative altitude. Moreover, they were specifically designed to aid performance on the task. That the results generally conformed to IE predictions but were asymmetric suggests that the importance of the elements to the task can influence how well IE can predict transfer.

For further clarification of the interaction between stimulus type and IE predicability, it is important to mention that of the five separate DVs, only NumExits and NavCon showed any sort of IE-related pattern of results. NumExits (perhaps somewhat arbitrarily) corresponds to effectiveness, and NavCon to procedural or navigational errors—getting too close to the border or an airport, or exiting at the wrong altitude or speed. None of the other measures—Plane Conflicts (PlnCon), Timing Conflicts (TimeCon), or Workload (WaitTime)—showed any IE patterns.

Why would only the number of planes successfully exited/landed and the number of procedural/navigational errors made show any correspondence to IE predictions? Again, the answer likely has to do with the role of the "elements" (the graphics) in question here. The support graphics were designed to indicate both an aircraft's trajectory and its relative altitude. Both of those factors are important in landing or exiting a plane correctly. In addition, the graphics likely had the side effect of making the aircraft more salient, perhaps making it easier to spot when one was too close to the border or to an airport. Thus, the graphics were central to those particular measures (NumExits and NavCon). However, because they had nothing to do with the timing elements of the task, the graphics at best only indirectly supported performance on the TimeCon and WaitTime measures. Because the graphics were central to NumExits and NavCon (which showed IE effects) but peripheral to TimeCon and WaitTime (which did not), the suggestion is that stimulus elements need to be important to the particular task in order for IE theory to have any predictability. However, that predictability will be influenced by the role of those elements—in other words, whether they support or hinder performance.[8]

Why the results did not support TAP, though, is still a question. Although originally a theory of memory performance, TAP has been successfully applied both to real-world memory situations and to physical tasks. What do those tasks have in common that a naturalistic cognitive activity such as air traffic control does not share? Likely, that common element lies in the research methods used in the previous TAP studies. Whereas those studies explicitly manipulated the type of processing used by participants, the current study instead attempted an indirect approach. Unfortunately, with the absence of positive findings supporting TAP, the question is open regarding whether that absence is because TAP cannot be applied to such situations, or because the task did not actually change the type of processing that participants used.

Training

Examining the data from the four training sessions, the composite measure, Score, showed no effect of training graphics, nor even of training session. The latter finding could indicate that little learning took place over the four sessions, but it is more likely a result of the progressive difficulty of each session. Gender, one of the covariates used in the analysis, did have an effect, though, with males scoring higher.

[8] It should be noted, though, that the graphics should have been most central to performance on the Plane Conflicts measure, yet no IE pattern was observed for PlnCon. However, PlnCon is an unusual variable in that whether or not training was provided seemed to make no difference. In other words, the only differences between the groups (both control and experimental) was between those that had the graphics during testing and those that did not. Collapsing across testing, the untrained control groups performed just as well (or as poorly) as the trained groups. That issue is dealt with later.

The analysis of the separate DVs was somewhat more interesting. A main effect of training graphics was found, due primarily to the effect of the graphics on the Plane Conflicts (PlnCon) measure, indicating that participants using the support graphics had fewer plane conflicts. Interestingly, though, when looking at the testing data, PlnCon was the only DV to show *no* effect of training; rather, the untrained control groups were not much different from their trained counterparts in their ability to maintain aircraft separation. In other words, during training, the graphics made it easier for subjects to avoid imminent aircraft conflicts. Despite that, those subjects did not perform any better than the untrained or the no-graphics groups. One possibility is that the testing sessions were so difficult that some sort of performance ceiling was reached; however, the testing graphics still resulted in better performance compared to the no-graphics groups, indicating that there was room for improvement. It is also possible that avoiding plane conflicts is simply a task that cannot be learned; however, performance improved over the testing sessions, so that explanation is unlikely. The finding is difficult to explain, but because keeping planes separated was the most salient and seemingly important component of the ATC task, the result may be related to participants focusing most of their attention on that activity.

A main effect of training session was also present, meaning that despite the progressive difficulty of the sessions, overall performance did improve. Learning was evident mainly just for NumExits and PlnCon, though. There was again a gender effect, but primarily for the NumExits measure; males exited more aircraft than did females during training. The lack of a training type x training session interaction indicates that the graphics did not help learning at all, at least as measured by performance. Rather, the indication is that they served as a perceptual cue that signaled potential aircraft conflicts, but they did not necessarily free up enough extra attention to help participants better learn how to perform the task.

Summary

The support graphics seemed generally to facilitate performance but not training. They provided immediate performance benefits on the task by highlighting imminent plane conflicts and procedural errors. However, they did not free up enough attentional resources for participants to learn less-salient elements of the tasks. An example of such an element is timing: New aircraft appeared at regular intervals in front of the gates, and the runways also switched directions at regular intervals. It is possible that such elements were simply very difficult to learn, though, so whether the support graphics did not create any attentional slack, or just not enough, is uncertain.

Though the graphics did not facilitate training overall, they did seem to have both positive and negative effects on training when type of testing was taken into account. Namely, participants who saw the graphics throughout the experiment performed the best, though not significantly. In contrast, participants who saw them only during training performed the worst, which suggests that the graphics overshadowed important information when participants were learning the task. Given the particular measures (NavCon and NumExits) that showed such effects, that overshadowed information likely involved the associations between the alphanumeric stimuli (which were redundant when the graphics were present) for each aircraft and the exits.

Finally, the results did not support the main hypothesis (hypothesis 1), which stated that performance on the ATC task would follow the transfer-appropriate processing account. Two possibilities are likely here: 1) TAP can't be applied to naturalistic settings, or 2) this study was not an adequate test of TAP; the two explanations aren't mutually exclusive. On the other hand, the findings did appear to conform to the similar but more general theory of identical elements, hypothesis 2. Support for IE theory was weak, though, because of the apparent overshadowing effects the support graphics have on training (hypothesis 5), and because of the performance advantage those graphics provide (hypothesis 6). In other words, although matched training/testing groups generally outperformed mismatched ones, the differences were not always significant. In particular, although the overshadowing effects of the training graphics amplified the differences between the GN and NN groups, the performance benefits of the graphics mitigated the differences between the GG and NG groups, thus producing an asymmetric pattern of results.

REFERENCES

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15*(4), 657-668.

Cockrell, J. T. (1979). *Effective training for target identification under degraded conditions*. (Report No. TP 358). Alexandria, VA: US Army Research Institute for the Behavioral & Social Sciences.

Endsley, M. R. (1996). Automation and situation awareness. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance: Theory and applications. Human factors in transportation.* (pp. 163-181). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. xx, 514 pp.

Endsley, M. R. (1995a). Measurement of situation awareness in dynamic systems. *Human Factors. Special Issue: Situation Awareness., 37*(1), 65-85.

Endsley, M. R. (1995b). Toward a theory of situation awareness in dynamic systems. *Human Factors. Special Issue: Situation Awareness., 37*(1), 32-64.

Endsley, M. R., & Kiris, E. O. (Jun1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors. Special Issue: Telecommunications, 37*(2), 381-394.

Goldstein, I. L., & Ford, J. K. (2002). *Training in organizations: Needs assessment, development, and evaluation* (4th ed.). Belmont, CA: Wadsworth.

Graf, P., & Ryan, L. (1990). Transfer-appropriate processing for implicit and explicit memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* *16*(6), 978-992.

Gronlund, S. D., Canning, J. M., Moertl, P. M., Johansson, J., Dougherty, M. R. P., & Mills, S. H. (2002). *An information tool for planning in air traffic control.* (Report No. FAA-AM-02-1).

Hollan, J. D., Hutchins, E. L., & Weitzman, L. (1984). STEAMER: An interactive inspectable simulation-based training system. *AI Magazine, 5*(2), 15-27.

Horton, K. D., & Nash, B. D. (1999). Perceptual transfer in stem-completion and fragment-completion tests. *Canadian Journal of Experimental Psychology*, *53*(3), 203-219.

Kanfer, R., & Ackerman, P. L. (1989a). Dynamics of skill acquisition: Building a bridge between intelligence and motivation. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 5* (pp. 83-134). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kanfer, R., & Ackerman, P. L. (1989b). Motivation and cognitive abilities: An integrative/aptitude^treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*(4), 657-690.

Klein, G. A., & Hoffman, R. R. (1993). Seeing the invisible: Perceptual-cognitive aspects of expertise. In M. Rabinowitz (Ed.), *Cognitive science foundations of instruction.* (pp. 203-226). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. x, 239 pp.

Leshner, G., & Coyle, J. R. (2000). Memory for television news: Match and mismatch between processing and testing. *Journal of Broadcasting & Electronic Media,* *44*(4), 599-613.

Lewandowsky, S., Dunn, J. C., Kirsner, K., & Randell, M. (1997). Expertise in the management of bushfires: Training and decision support. *Australian Psychologist, 32*(3), 172-177.

Manning, C. A., & Broach, D. (1992). *Identifying ability requirements for operators of future automated air traffic control systems*. (Report No. FAA-AM-92-26).

Meier, B., & Graf, P. (2000). Transfer appropriate processing for prospective memory tests. *Applied Cognitive Psychology. Special Issue: New Perspectives in Prospective Memory, 14*, S11-S27.

Mogford, R. H. (1997). Mental models and situation awareness in air traffic control. *International Journal of Aviation Psychology, 7*(4), 331-341.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied, 6*(1), 44-58.

Morrison, J. G., Kelly, R. T., Moore, R. A., & Hutchins, S. G. (1998). Implications of decision-making research for decision support and displays. In J. A. Cannon-Bowers, & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 375-406). Washington, DC: American Psychological Association.

Muir, B. M. (1994). Trust in automation: I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics. Special Issue: Cognitive Ergonomics., 37*(11), 1905-1922.

Niessen, C., Eyferth, K., & Bierwagen, T. (1999). Modelling cognitive processes of experienced air traffic controllers. *Ergonomics. Special Issue: Cognitive Science Approaches to Process Control, 42*(11), 1507-1520.

Peynircioglu, Z. F., Thompson, J. L. W., & Tanielian, T. B. (2000). Improvement strategies in free-throw shooting and grip-strength tasks. *Journal of General Psychology, 127*(2), 145-156.

Rajaram, S., Srinivas, K., & Roediger, H. L. I. (1998). A transfer-appropriate processing account of context effects in word-fragment completion. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24*(4), 993-1004.

Roediger, H. L., Weldon, M. S., Stadler, M. L., & Riegler, G. L. (1992). Direct comparison of two implicit memory tests: Word fragment and word stem completion. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18*(6), 1251-1269.

Shanks, D. R., & Cameron, A. (2000). The effect of mental practice on performance in a sequential reaction time task. *Journal of Motor Behavior, 32*(3), 305-313.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics.* (3rd ed.). New York, NY: HarperCollins College Publishers.

Vortac, O. U., Edwards, M. B., & Manning, C. A. (1995). *Functions of external cues in prospective memory*. (Report No. FAA-AM-95-9). FAA Office of Aviation Medicine Reports.

Whitfield, D., & Jackson, A. (1982). The air traffic controller's picture as an example of a mental model. *IFAC Conference on Analysis, Design and Evaluation of Man-*

Machine Systems (pp. 37-44). London: HMSO.

Wonderlic Personnel Test (1992). *Wonderlic Personnel Test User's Manual*. Libertyville, IL: Wonderlic Personnel Test, Inc.

Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors, 43*(3), 355-365.

Young, W. C., Broach, D., & Farmer, W. L. (1997). *The effects of video game experience on computer-based Air Traffic Controller Specialist, Air Traffic Scenario Test scores*. (Report No. FAA-AM-97-04). FAA Office of Aviation Medicine Reports.

Zsambok, C. E. (1997). Naturalistic decision-making: Where are we now? In C. E. Zsambok, & G. Klein (Eds.), *Naturalistic Decision Making* (pp. 1-16). Mahwah, NJ: Lawrence Erlbaum Associates.

APPENDIX A

Results for Covariates—Training Data

Composite measure of performance

One of the covariates, gender, was significantly associated with score, $F(1, 141) = 16.42$, $p < .001$, $\text{eta}_p^2 = .10$, $\beta = .98$, with males outperforming females.

Sub-scores

The interaction between training session and the gender covariate was significant, $F(15, 127) = 2.01$, $p = .020$, $\text{eta}_p^2 = .19$, $\beta = .95$. The gender covariate also accounted for a significant amount of the variance in the set of DVs, $F(5, 137) = 7.82$, $p < .001$, $\text{eta}_p^2 = .22$, $\beta > .99$. Location of experiment was not significant.

The gender covariate also had a significant effect on NumExits, WaitTime, and TimeCon, though the effect was weak for the latter two. In all cases, males demonstrated better performance; see Table 12.

- NumExits: univariate $F(1, 141) = 21.97$, $p < .001$, $\text{eta}_p^2 = .14$, $\beta > .99$.
- WaitTime: univariate $F(1, 141) = 4.98$, $p = .027$, $\text{eta}_p^2 = .03$, $\beta = .60$.
- TimeCon: univariate $F(1, 141) = 6.33$, $p = .013$, $\text{eta}_p^2 = .04$, $\beta = .70$.

Table 12
Means and Standard Deviations for Gender Effect on NumExits, WaitTime, and TimeCon

	Gender	
	Male	Female
NumExits	8.41 (3.01)	6.12 (2.60)
WaitTime	925.24 (831.18)	1256.84 (1015.78)
TimeCon	3.31 (1.24)	3.86 (1.41)

For TimeCon, the male advantage was demonstrated primarily during the second half of training; hence, the significant training session x gender interaction, to which TimeCon was the only contributing DV; see Table 13.

- TimeCon: univariate $F(3, 423) = 4.61$, $p = .003$, $\text{eta}_p^2 = .03$, $\beta = .89$.

Table 13
Means and Standard Deviations for TimeCon x Gender over Four Training Sessions

	Gender		
	Male	Female	*Grand Total*
TimeCon1	3.83 (1.80)	3.92 (2.00)	*3.87 (1.88)*
TimeCon2	2.49 (1.41)	2.52 (1.82)	*2.50 (1.59)*
TimeCon3	2.94 (1.82)	3.54 (2.03)	*3.19 (1.92)*
TimeCon4	3.96 (2.68)	5.44 (3.08)	*4.59 (2.94)*
Grand Total	*3.31 (1.24)*	*3.86 (1.41)*	*3.54 (1.34)*

Results for Covariates—Testing Data

Composite measure of performance

One of the covariates, gender, was significantly associated with Score, $F(1, 197) = 9.66$, $p = .002$, with a small effect, $\text{eta}_p^2 = .05$, and high power, $\beta = .87$. The advantage was for males, who generally scored higher than females. Location barely approached significance, $\beta = .10$.

Sub-scores

Gender had a modest effect on the combined DVs, $F(5, 193) = 6.08$, $p < .001$, $\text{eta}_p^2 = .14$, $\beta > .99$. The advantage was for males, and only for the more spatial measures of performance: NumExits, $F(1,197) = 27.14$, $p < .001$, $\text{eta}_p^2 = .12$, $\beta > .99$, and NavCon, $F(1,197) = 7.31$, $p = .007$, $\text{eta}_p^2 = .04$, $\beta = .77$; see Table 14.

Table 14
Means and Standard Deviations of NumExits and NavCon by Gender

	Gender	
	Male	Female
NumExits	9.80 (5.14)	6.94 (4.34)
NavCon	5.69 (4.52)	7.59 (5.37)

Location of experiment was also significant, *F*(5, 193) = 2.76, *p* = .020, eta_p^2 = .07, β = .82, primarily for the WaitTime, *F*(1, 191) = 7.54, *p* = .007, eta_p^2 = .04, β = .78, and NavCon, *F*(1, 191) = 6.06, *p* = .015, eta_p^2 = .03, β = .69 measures; see Table 15. For NavCon, the participants who completed the experiment in the student computing center made more such errors than did those completing it in smaller computer labs or at home, likely reflecting the computer center's more distracting environment. However, those participants were also somewhat quicker on the WaitTime measure, possibly because of increased arousal due to the busier setting.

Table 15
Means and Standard Deviations of NavCon and WaitTime for Each Location

	Location			
	Computing Center	Computer Lab	Dorm	Off Campus
NavCon	7.16 (5.41)	5.71 (5.08)	5.20 (3.61)	4.96 (3.16)
WaitTime	3465.50 (2494.37)	4573.43 (2687.16)	4445.69 (2818.34)	4628.73 (2878.81)

APPENDIX B

Descriptive Statistics for Each Testing Session

Table 16
Means and Standard Deviations for Score5 and Score6

		Dependent Variable	
Type of Training	Type of Testing	Score5	Score6
Control (C)	Graphics (G)	526.62 (206.50)	629.86 (183.45)
	No Graphics (N)	443.70 (222.13)	483.91 (207.75)
Control Total		*494.83 (214.64)*	*573.92 (204.33)*
Graphics (G)	Graphics (G)	779.69 (251.74)	760.31 (270.62)
	No Graphics (N)	592.05 (254.17)	635.51 (240.38)
Graphics Total		*676.62 (268.28)*	*691.76 (260.21)*
No Graphics (N)	Graphics (G)	766.46 (232.85)	733.66 (258.05)
	No Graphics (N)	717.27 (218.46)	725.91 (192.34)
No Graphics Total		*744.53 (226.35)*	*730.20 (229.61)*
Grand Total		*647.93 (258.69)*	*671.15 (241.59)*

Table 17
Means and Standard Deviations for PlnCon5 and PlnCon6

		Dependent Variable	
Type of Training	Type of Testing	PlnCon5	PlnCon6
Control (C)	Graphics (G)	26.22 (12.30)	22.89 (9.96)
	No Graphics (N)	34.30 (13.13)	29.61 (14.47)
Control Total		*29.32 (13.13)*	*25.47 (12.23)*
Graphics (G)	Graphics (G)	26.25 (13.52)	24.06 (12.16)
	No Graphics (N)	30.79 (16.22)	28.90 (16.53)
Graphics Total		*28.75 (15.13)*	*26.72 (14.82)*
No Graphics (N)	Graphics (G)	25.07 (12.53)	25.56 (15.09)
	No Graphics (N)	31.94 (13.20)	25.64 (12.18)
No Graphics Total		*28.14 (13.20)*	*25.59 (13.78)*
Grand Total		*28.69 (13.82)*	*25.95 (13.67)*

Table 18
Means and Standard Deviations for NumExits5 and NumExits6

		Dependent Variable	
Type of Training	Type of Testing	NumExits5	NumExits6
Control (C)	Graphics (G)	4.65 (4.44)	6.43 (4.04)
	No Graphics (N)	4.35 (5.91)	4.26 (4.21)
Control Total		*4.53 (5.01)*	*5.60 (4.21)*
Graphics (G)	Graphics (G)	12.00 (4.49)	11.00 (5.38)
	No Graphics (N)	7.62 (6.41)	8.74 (5.47)
Graphics Total		*9.59 (6.00)*	*9.76 (5.51)*
No Graphics (N)	Graphics (G)	10.46 (5.51)	10.46 (5.71)
	No Graphics (N)	11.24 (4.51)	10.39 (4.15)
No Graphics Total		*10.81 (5.07)*	*10.43 (5.04)*
Grand Total		*8.55 (5.98)*	*8.79 (5.38)*

Table 19
Means and Standard Deviations for NavCon5 and NavCon6

		Dependent Variable	
Type of Training	Type of Testing	NavCon5	NavCon6
Control (C)	Graphics (G)	7.24 (4.58)	6.49 (4.56)
	No Graphics (N)	10.39 (6.67)	9.39 (5.86)
Control Total		*8.45 (5.64)*	*7.60 (5.25)*
Graphics (G)	Graphics (G)	4.75 (3.54)	4.22 (3.26)
	No Graphics (N)	9.38 (7.17)	7.46 (7.59)
Graphics Total		*7.30 (6.23)*	*6.00 (6.22)*
No Graphics (N)	Graphics (G)	5.32 (4.14)	5.66 (4.95)
	No Graphics (N)	4.12 (2.99)	5.15 (3.95)
No Graphics Total		*4.78 (3.70)*	*5.43 (4.51)*
Grand Total		*6.73 (5.46)*	*6.26 (5.41)*

Table 20
Means and Standard Deviations for TimeCon5 and TimeCon6

		Dependent Variable	
Type of Training	Type of Testing	TimeCon5	TimeCon6
Control (C)	Graphics (G)	17.49 (7.40)	17.19 (8.02)
	No Graphics (N)	13.91 (5.32)	15.52 (5.45)
Control Total		*16.12* (*6.86*)	*16.55* (*7.14*)
Graphics (G)	Graphics (G)	13.22 (5.42)	13.47 (7.25)
	No Graphics (N)	12.67 (5.11)	13.69 (5.69)
Graphics Total		*12.92* (*5.22*)	*13.59* (*6.39*)
No Graphics (N)	Graphics (G)	11.05 (5.07)	12.98 (6.12)
	No Graphics (N)	11.91 (5.36)	14.45 (4.60)
No Graphics Total		*11.43* (*5.19*)	*13.64* (*5.51*)
Grand Total		*13.32* (*6.02*)	*14.47* (*6.44*)

Table 21
Means and Standard Deviations for WaitTime5 and WaitTime6

		Dependent Variable	
Type of Training	Type of Testing	WaitTime5	WaitTime6
Control (C)	Graphics (G)	1931.80 (1409.25)	1746.79 (1372.14)
	No Graphics (N)	1826.00 (1066.16)	1328.47 (966.40)
Control Total		*1891.24* (*1279.97*)	*1586.43* (*1240.61*)
Graphics (G)	Graphics (G)	4547.47 (2575.78)	4079.18 (2742.08)
	No Graphics (N)	3607.94 (2602.04)	3883.19 (2665.40)
Graphics Total		*4031.39* (*2614.44*)	*3971.52* (*2682.56*)
No Graphics (N)	Graphics (G)	3794.21 (3188.93)	3709.12 (3085.99)
	No Graphics (N)	3954.28 (2960.42)	4320.47 (2618.91)
No Graphics Total		*3865.60* (*3069.27*)	*3981.75* (*2884.17*)
Grand Total		*3345.16* (*2659.73*)	*3277.14* (*2660.78*)

VITA

Michael E. Stiso
Department of Psychology, Texas A&M University
College Station, TX 77843-4235

EDUCATION

August 2003	*Doctor of Philosophy*, Psychology Texas A&M University, College Station, TX
December 1998	*Master of Science*, Cognitive Psychology University of Oregon, Eugene, OR
May 1992	*Bachelor of Arts*, Psychology Purdue University, West Lafayette, IN Minors: English, Anthropology

AWARDS

August 2002 – August 2003	*NASA Marshall Space Flight Center* Graduate Student Researchers Program Fellowship Marshall Space Flight Center, Huntsville, AL
June 1998 – August 1998	*Air Force Office of Scientific Research* Graduate Fellowship Program Brooks Air Force Base, San Antonio, TX

PROFESSIONAL EXPERIENCE

May 2002 – July 2002	*Naval Research Enterprise Intern Program, SPAWAR* Human-Factors Intern
May 2000 – December 2000	*SBC Technology Resources Inc.* Human-Factors Intern
December 1998 – May 2000	*DoggettData* Cognitive Research Consultant
June 1998 – August 1998	*Air Force Office of Scientific Research* Intern/Consultant