12-2012

Understanding the Relative Attentional Demands of the Dimensions of Interface Consistency

Jeremy Mendel
Clemson University, mendel.clemson@gmail.com

Recommended Citation

Mendel, Jeremy, "Understanding the Relative Attentional Demands of the Dimensions of Interface Consistency" (2012). *All Dissertations*. 1076.
https://tigerprints.clemson.edu/all_dissertations/1076

UNDERSTANDING THE RELATIVE ATTENTIONAL DEMANDS OF THE DIMENSIONS OF INTERFACE CONSISTENCY

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctorate of Philosophy
Human Factors Psychology

---

by
Jeremy Mendel
December 2012

Submitted to:
Dr. Richard Pak, Committee Chair
Dr. Joel Greenstein
Dr. Leo Gugerty
Dr. Christopher Pagano

ABSTRACT

A consistent interface is thought to be beneficial because it allows users to draw on previous training and experience when operating a new interface. Design guidelines like the eight golden rules of interface design argue that a highly consistent interface improves system usability (Shneiderman, 1987). However, interface consistency is not monolithic; instead it is a complex, multidimensional construct. I refer to the two dimensions of interface consistency as perceptual consistency (the appearance) and conceptual consistency (the functionality) of an interface. Perceptual consistency considers aspects like interface layout and orientation; conceptual consistency considers how the system operates or responds. I sought to understand how combinations of these dimensions might affect performance and user perceptions of a system. For example, what if a system looks the same but operates differently? Results indicate that both an inconsistent appearance and an inconsistent functionality can hurt performance. Forcing consistency, however, may not be beneficial either. When there was a mismatch between dimensions (i.e., one was consistent and the other inconsistent) performance was worse than that of an entirely inconsistent version. Specifically, participants in the conceptual inconsistency and perceptual consistency condition (operates differently but looks the same) performed worse and reported higher workloads. Designers should encourage interface consistency by making systems that function similarly also share a similar appearance; however, when the systems are functionally disparate (i.e., they do different things) designers should take care to avoid implying similarities where they do not exist.

ACKNOWLEDGEMENTS

Thank you to Dr. Richard Pak, my committee chair and advisor during my time in graduate school. You taught me more than I ever imagined; your guidance made this all possible. I would also like to thank my committee members: Dr. Joel Greenstein, Dr. Leo Gugerty, and Dr. Christopher Pagano for helping me improve my skills as a researcher. Thank you to my fantastic officemates Brock, Margaux, Natalee, and Nicole for the camaraderie. To Margaux and Nicole, we made it!

I also want to thank Shana for putting up with my occasionally unconventional hours while I worked on this dissertation. Shana, you helped me get through all the tough times, even when I was convinced I would never finish. We can be “real adults” together now that I am done with school at last.

Finally, and most importantly, thank you mother for your unwavering support (even though I keep moving farther away) and for helping me grow into the person I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Although usability experts (e.g., Norman, 1988; Shneiderman, 1987) support interface consistency and include it as a core part of design guidelines, empirical evidence is not as unanimous. Further, interface consistency is a complex multi-dimensional construct. Models break interface consistency into either two dimensions (e.g., Tanaka, Eberts, & Salvendy, 1991) or three dimensions (e.g., Rhee, Moon, & Choe, 2006). The goal of the current study was to examine the unique contributions of the dimensions of interface consistency on the attentional demands of an interface.

Review of the Effects of Interface Consistency

When examined empirically, the performance effects of consistency are unclear (see Table 1 for a summary). Some studies found that consistency improved performance (e.g., AlTaboli & Abou-Zeid, 2007). Additional studies have found limited or no effects of consistency (e.g., Ozok & Salvendy, 2000; Rhee et al., 2006). Other studies even demonstrated at least partially detrimental effects of a consistent interface (Finstad, 2008; Satzinger & Olfman, 1998).

Interface consistency can be thought of as a system that encourages similar behaviors by designing that system with analogous situations or task objectives. Generally speaking, a consistent interface is thought to be beneficial because it allows users to leverage previous knowledge when using a new system (Brown, 1999; Nielsen, 1989; Norman, 1988). An early theory of transfer suggests that the amount of transfer between tasks is determined by the amount of similar content (Thorndike & Woodworth, 1901). Thorndike and Woodworth argue that the more overlap in stimulus-response pairings between two systems, the more transfer. A consistent interface attempts to facilitate this transfer through system design (e.g., Polson & Lewis, 1990). Rieman, Lewis, Young, and Polson (1994) provide evidence that consistent interfaces can help users learn a new system by encouraging analogical reasoning from a previously learned system.

Most researchers in the interface consistency literature argue that it can improve user outcomes. The outcomes of a consistent interface include reduced task completion time, fewer errors, and improved user satisfaction (e.g., Mendel, Pak, Drum, 2011). Specifically, researchers suggest that a consistent interface decreases working memory demand, increases efficiency, enhances visual search, and reduces the learning process (Bayer, 1992; Polson, 1988; Proctor & Vu, 2006). It is important to note however, that not all researchers believe that a consistent interface is beneficial. Grudin (1989) criticized the concept of interface consistency as being too vague when specifying what makes an interface consistent. He argued that the time spent attempting to make an interface consistent could be better used to pursue more effective improvements. It is important to first understand how researchers describe interface consistency before discussing the findings from previous research.

Models of Interface Consistency

Usability guidelines describe consistency as though it is a scale ranging from consistent to inconsistent. Developing a consistent interface is not a single design choice (e.g., the layout); instead, interface consistency is multifaceted and comprised of different “dimensions” or types of consistency (e.g., Kellogg, 1987). In an effort to further clarify the concept of consistency, researchers have operationalized specific dimensions of user interaction and how they contribute to a consistent interface. Previous research modeled a consistent interface either using a three dimensional model (e.g., Rhee et al., 2006) or a two dimensional model (e.g., Tanaka et al., 1991).

Three dimensional model

A three dimensional model was first described by Kellogg (1987). Much of the research during the last dozen years organized the components of a consistent interface using this three dimensional model (AlTaboli & Abou-Zeid, 2007; Ozok & Salvendy, 2000; 2001; 2003; 2004; Rhee et al., 2006). Kellogg operationalized interface consistency as three dimensions of user interaction with the system: physical, conceptual, and communicational. Using a multidimensional model allowed Kellogg, and later researchers, to describe the components of a consistent interface and study how those components affected user outcomes.

Physical consistency considers the visual or graphical appearance of an interface or object including details like color, location, orientation, and arrangement of interface elements (e.g., AlTaboli & Abou-Zeid, 2007). The physical interface aspects influence the visual aesthetics of the system and can affect the way a user perceives the system. Physical interface consistency serves two important purposes: first, it is the most easily perceivable of the three dimensions allowing it to signal users to presence of consistency; second, a consistent location and arrangement of interface elements, major components of physical consistency, can help by reducing the visual search required to use the interface. One example of physical consistency is the location of an automobile speedometer. Nearly all automobiles have an analog speedometer with a similar design right above the steering column. This standard is so prevalent that drivers are able to gauge speed in a different car without first searching for the speedometer or studying its design.

Conceptual consistency can be thought of in terms of Norman’s conceptual models (1988). This dimension describes how the system image (i.e., system operation, responses, and documentation) informs the user’s model (i.e., user’s expectations or understanding of the system). Conceptual consistency can be thought of as the consistency of the user’s model and how that model represents components of an interface. Better understanding of the system makes it easier for users to convert task goals into system procedures (Kellogg, 1987). Systems with higher conceptual consistency allow users to better transfer existing knowledge from a previously learned system to a new system. Conceptual consistency is not always readily apparent to users; instead, other forms of consistency, like physical consistency, may be needed to help indicate conceptual consistency to users.

An example of conceptual consistency is the menu bar found in most windows-based applications. This menu bar uses similar menu organizational structures between programs with each menu containing similar commands such as File>Save or Edit>Copy. Expert users of these products have a detailed representation of how these various functions are organized and can generalize this knowledge between most programs. Newer versions of Microsoft Office restructured the menu into a less hierarchical, tab-based menu. Changes like completely removing the edit menu and redistributing commands to other areas forced expert users to relearn the conceptual structure of the system.

Communicational consistency is the consistency of the human-system interface (e.g., Rhee et al., 2006). Communicational consistency includes both the way a user interacts with the system (input) and how the system presents information to the user (output). Communicational inconsistency results from switching a mouse for a touchscreen or using an auditory command prompt rather than text. Communicational aspects tend to coincide with physical and/or conceptual aspects of the interface. Previous research used manipulations of communicational consistency that are confounded with either one or both of the other two dimensions. For example, one study attempted to assess the effects of the three dimensions by manipulating each one independently (Ozok & Salvendy, 2000). In that study, the researchers manipulated communicational consistency by altering the location of task elements and requiring scrolling with either a scroll bar or a text link. That manipulation affects both the communicational and physical consistency of an interface. This issue suggests that communicational consistency may not be an independent form of consistency.

Research using the three dimensional model

Although researchers like Ozok and Salvendy (2000) advocate the benefits of a consistent interface, empirical results are less conclusive. Three studies using similar methodologies sought to explore the effects of the three dimensions of consistency (AlTaboli & Abou-Zeid, 2007; Ozok & Salvendy, 2000; Rhee et al., 2006). AlTaboli and Abou-Zeid only examined the effects of physical consistency (i.e., the appearance) while the studies by Ozok and Salvendy along with Rhee et al. studied the effects of all three dimensions. Participants in these studies completed a series of web-based tasks including point-and-click tasks (find a link and click it), data entry, reading comprehension, and word searches.

Results from two of the studies suggested that physical consistency sometimes can improve performance. Consistent element location, an aspect of physical consistency, reduced error-rate (AlTaboli & Abou-Zeid, 2007; Ozok & Salvendy, 2000); however, none of the studies found any effect on task completion time. The other two dimensions of interface consistency (i.e., conceptual and communicational consistency) had no significant effect on performance or user satisfaction. Further, the study by Rhee et al. (2006) found no significant effect from any of three dimensions of interface consistency. These studies produced three main conclusions. First, physical interface consistency had a larger effect than communicational or conceptual consistency. Second, element location was the most effective manipulation of physical consistency (AlTaboli & Abou-Zeid, 2007). Third, error-rates seem to be more influenced by interface consistency than task completion time.

Table 1. Studies Examining Interface Consistency's Effect on Performance.

Study	Task Domain	General Findings	Interface Consistency
AlTaboli & Abou-Zeid (2007)	Web-based (e.g., point & click, reading comprehension, form filling).	No significant effect on task completion time. Some significant effects on error-rate and user satisfaction.	Positive: some types of physical consistency resulted in fewer errors and better user satisfaction. Location consistency was the most effective.
Finstad (2008)	Web Browser Applications	Poorly implemented consistency might lead novice users to inappropriately generalize knowledge between systems. Improper generalizations might hurt performance more than inconsistent interfaces.	Negative: one form of consistency resulted in longer task completion time and more errors.
Ozok & Salvendy (2000)	Web-based (e.g., point & click, reading comprehension, form filling).	Limited effects of consistency. One type of consistency (physical) reduced errors. No effects on completion time or satisfaction.	Positive & No effects: physical consistency reduced error-rate. Others had no effect.
Rhee, Moon & Choe (2006)	Web-based (e.g., point & click, reading comprehension, form filling).	No significant effects on performance at the .05 level.	No effects.
Satzinger & Olfman (1998)	Scheduling and Communications Applications	Some forms of consistency improved performance, others hurt performance.	Mixed: one form of consistency improved user efficiency while the other form increased error-rate.
Tanaka, Eberts, & Salvendy (1991)	Menu-interactive Tasks	Both dimensions of consistency improved performance. Did not consider combinations of the dimensions. Inconsistency resulted in worse knowledge retention a week later.	Positive: higher levels of each of the two dimensions improved performance and retention.

Problems with the three dimensional model

One major goal of previous research (e.g., Ozok & Salvendy, 2000; Rhee et al., 2006) was to examine the relative effects of each dimension of consistency. The interdependent nature of communicational consistency makes it difficult to accurately assess the relative effects of each dimension of interface consistency. Additionally, it makes classifying design manipulations as a certain form of consistency more challenging. For example, how do the two systems in Figure 1 differ? The vehicle on the left uses a knob and the vehicle on the right uses up and down buttons. According to the three dimensional model, this change is both a manipulation of physical and communicational consistency. Perhaps, a model of interface consistency might be more parsimonious if communicational consistency was incorporated into the other two dimensions.

Figure 1. Two example vehicle climate control systems.

Two dimensional models

Other studies based manipulations of interface consistency on some variation of a two dimensional model (e.g., Finstad, 2008; Satzinger & Olfman, 1998; Tanaka et al., 1991). These models differ from the three dimensional models by collapsing communicational consistency into the other two dimensions. Tanaka et al. (1991) refers to the two dimensions as *display consistency* and *cognitive consistency*. Display consistency is roughly equivalent to the physical dimension used in the three dimensional model. Cognitive consistency considers what the users knows and is closest to the conceptual dimension of the three dimensional model. Others studies used similar two dimensional models but with variations on the names of the dimensions (Finstad, 2008; Satzinger & Olfman, 1998). In this paper I refer to the two dimensions as conceptual consistency (organized knowledge structures related to the organization of system functions) and perceptual consistency (the outward visual appearance of the system including color, layout, and visual organization).

Research using two dimensional models

One pair of studies used two dimensional models to examine the effects of interface consistency (Finstad, 2008; Satzinger & Olfman, 1998). Both studies found mixed performance effects in that some forms of interface consistency were detrimental to performance (longer completion time and more errors). Finstad found that conceptual inconsistency coupled with perceptual consistency performed worse than complete inconsistency. Finstad argued that these errors were due to participants over-generalizing prior knowledge to the new interface. The second study found that interface consistency could be either beneficial or detrimental depending on the form of consistency used (Satzinger & Olfman, 1998). Specifically, Satzinger and Olfman found that conceptual consistency reduced the steps needed to complete a task but that perceptual consistency reduced accuracy. An issue in both studies, however, is that some of the “consistent” interfaces actually demonstrated what I would consider incomplete consistency. Incomplete consistency occurs when there is a mismatch in consistency between dimensions (i.e., one consistent and one inconsistent). In these studies, the combination of conceptual inconsistency and perceptual consistency may have led participants to erroneously perceive the system as consistent resulting in inappropriate generalizations. Conflicting results from these studies make it unclear exactly under what conditions interface consistency is helpful or harmful.

The literature reviewed suggests a conflicted view of consistency. Although a consistent interface is theoretically beneficial, empirical results of consistency are unclear. In testing, interface consistency studies found positive effects (e.g., AlTaboli & Abou-Zeid, 2007), non-significant effects (e.g., Rhee et al., 2006), and even detrimental effects (e.g., Finstad, 2008).

One possible explanation for these mixed results can be seen in how the dimensions of interface consistency were manipulated. Research to date attempted to manipulate each dimension of consistency separately, treating each dimension as an independent component (e.g., Rhee et al., 2006). Perhaps these dimensions of interface consistency are actually connected and possibly even confounded with one another. Instead of treating these dimensions independently, as in the previous literature, combinations of the dimensions should be considered together to better understand the effect of incomplete consistency.

Model for the present study

Based on the issues presented with the three dimensional model, I favor two dimensional models. The present study framed interface consistency using a two dimensional model rather than a three dimensional model for two main reasons. First, studies using a two dimensional model found significant effects of both dimensions (e.g., Finstad, 2008). In contrast, studies using the three dimensional model found limited or non-significant effects for one or more of the three dimensions (e.g., Rhee et al., 2006). Second, the communicational dimension used in the three dimensional model does not seem to be an independent dimension. The dimension is both ill-defined in that it is not mutually exclusive and the dimension is not supported by empirical results (e.g., Ozok & Salvendy, 2000). Based on these observations I chose to use a two dimensional model for the present study. This model is similar to Tanaka et al. (1991) but instead I refer to the dimensions as conceptual consistency and perceptual consistency.

Perceptual consistency is anything that is primarily perceptual in nature; i.e., aspects of the system that users see, hear, or feel. This includes aspects of the system like color, location, sound cues, and vibrations. Perceptual consistency is relatively easy for users to notice and serves as a crucial cue to indicate consistency. Examples include changing color schemes, rearranging the layout of the interface, or replacing a knob with buttons.

Conceptual consistency is the consistency of the user's model (Norman, 1988). Designers communicate system information to the user through the interface (e.g., an underlined blue word communicates "click me"). These design choices can influence user expectations (e.g., a user expects a certain response from the system or expects the system to respond in a specific way to a certain action sequence). Some conceptual consistencies may not be immediately obvious to a user; in that case, some perceptual consistencies might be helpful to cue the users. Other conceptual manipulations might coincide with a perceptual change; however, the distinction for these changes is that a conceptual manipulation alters the expectations or understanding of a system. Systems with conceptual consistency make it easier for a user to transfer existing knowledge to a new system.

Previous research, like Satzinger and Olfman (1998), assessed the relative effects of each of the dimensions of consistency as though they operate in isolation. While it is important to understand the independent effects of the dimensions, it is also important to consider how the dimensions of interface consistency interact with one another. Ignoring the relationship between the dimensions of consistency might have contributed to the unanticipated negative effects observed in previous research (Finstad, 2008; Satzinger & Olfman, 1998). Another important consideration in interpreting previous research is the discrepant task characteristics between studies, specifically the amount of task workload (Mendel et al., 2011). A task must be sufficiently challenging for interface consistency to have an effect.

Using this model of consistency, it is important to note that the dimensions of interface consistency can be manipulated independently. I describe mixed levels of consistency (i.e., one dimension high consistency and the other low) as incomplete consistency. Although no research to date has specifically described incomplete consistency, previous research demonstrated the adverse effects of incomplete consistency (e.g., Finstad, 2008). Further, based on previous research (Mendel et al., 2011), I expected that the effects of incomplete consistency would be greatest when workload is high.

An example of the interplay between these dimensions can be illustrated by comparing a traditional internal combustion engine vehicle to an electric vehicle. In this case, the two vehicles are mostly perceptually consistent with the primary exception of engine noise. The vehicles are also mostly conceptually consistent in that users can easily transfer existing knowledge of driving a combustion-based vehicle into the ability to drive an electric vehicle. A major conceptual inconsistency, however, is the differences in the transmission systems and the maintenance required for each system. An electric vehicle never needs an oil change since there is no combustion engine to lubricate. Instead, maintenance for electric vehicles consists of non-drivetrain related issues like replacing tires or changing brake pads. Drivers should appreciate how an electric vehicle differs to understand the different maintenance requirements between the two vehicle types. Instead, users may see the perceptual consistencies between the two vehicles and as a result, fail to appreciate the conceptual inconsistencies that are less salient. When two systems appear identical but operate in different ways, users will likely generalize expectations inappropriately just as in some of the previously discussed studies (e.g., Satzinger & Olfman, 1998).

Remaining Issues in the Literature

Outcomes of previous research ranging from positive (e.g., AlTaboli & About-Zeid, 2007) to negative (Finstad, 2008) obscure the conditions in which a consistent interface is beneficial. I identified two key issues of interface consistency that remain unresolved. First, the research to date has only focused on measuring the relative performance of interface consistency rather than assessing the differential attentional demands resulting from consistency (or inconsistency). Second, previous research (e.g., Rhee et al., 2006) manipulated individual dimensions of consistency without considering the interdependent nature of these dimensions. As a result, some conditions might have resulted in incomplete consistency. Without careful manipulation, the relative effects of the dimensions of consistency remain unclear.

Interface consistency's effect on attentional demand

As suggested by previous research (Mendel et al., 2011), other studies may not have used tasks that were sufficiently challenging (e.g., Rhee et al., 2006). If the tasks were too easy and thus not resource-limited, then these studies may not have effectively assessed the impact of interface consistency. One possible solution to this issue is to utilize a multitask approach. A multitask approach can assess the relative cognitive capacity required to complete a task (e.g., Fisk, Derrick, & Schneider, 1986; Wickens & Hollands, 2000). Multitask procedures can effectively create differential levels of attention allocation within a study (e.g., Gopher, 1993). In the case of interface consistency, a multitask approach could help to elucidate the relative demands of different combinations of interface consistency. The idea is that in a resource-constrained situation (i.e., multi-task), a consistent interface would result in reduced attentional demand. If participant performance varies as a function of interface consistency, then the change can be attributed to differential attention requirements (e.g., McLaughlin, Rogers, & Fisk 2009).

Conflicting dimensions of consistency and incomplete consistency

Previous studies attempted to study the effects of the dimensions of interface consistency by treating each as an independent construct (e.g., Ozok & Salvendy, 2000). For example, a manipulation of communicational consistency might actually coincide with a conceptual manipulation (e.g., Rhee et al., 2006). This must be minimized to assess the real effects of each dimension. Further, past research has not accounted for the possible detrimental effects of incomplete consistency. Research must carefully manipulate each dimension of consistency while considering the possible effects of incomplete consistency. Forms of incomplete consistency must be examined to understand how combinations of the dimensions help or harm users.

Current Study

Past research either focused on the differential effects of the dimensions of interface consistency (e.g., Rhee et al., 2006) or on the impact of task workload (e.g., Mendel et al., 2011). The present study sought to explore how conceptual and perceptual interface consistency influences the workload of a task. Specifically, the goal of this study was to understand the relative attentional demands of the dimensions of consistency under high task workload. Further, I wanted to assess the effects of the various combinations of the dimensions of consistency (i.e., perceptual consistency and conceptual inconsistency). The two hypotheses for the present study were as follows:

1. The effects of a consistent interface would vary as a function of time with the greatest effects occurring immediately after implementing the manipulations. As time passes, the effects of an inconsistent interface would diminish.
2. Although a consistent interface would improve performance, incomplete consistency conditions (i.e., when one dimension is consistent and the other is inconsistent) would be detrimental. From this I expected that the condition of complete inconsistency (i.e., both dimensions are inconsistent) would do better than incomplete consistency for at least some performance measures. Additionally, I expected that the detrimental effects would be especially pronounced for the combination of conceptual inconsistency and perceptual consistency (i.e., when the system operates differently but looks the same).

The current study utilized a multitask approach to assess the attentional demands of the interface design (e.g., Fisk et al., 1986). Participants completed tasks using a simulated, novel control task (i.e. a futuristic spaceship control panel). The panel required participants to perform three separate but interrelated tasks: constantly managing the power allocation in the ship, continuously providing course corrections, and completing tasks as assigned by the ship’s captain. Participants were instructed that the power allocation and course correction tasks were to be emphasized.

Participants completed tasks with the simulator during two separate 30 minute task phases. Versions of the system used for each of the two task phases depended on randomly assigned participant conditions. Based on the design of the study, there are a total of four possible experimental conditions (Figure 2). An example participant might be assigned to be in the low conceptual consistency and high perceptual consistency condition. In that case, the two task sessions would be perceptually consistent and conceptually inconsistent with one another. All conditions were counterbalanced and manipulations were all between-group.

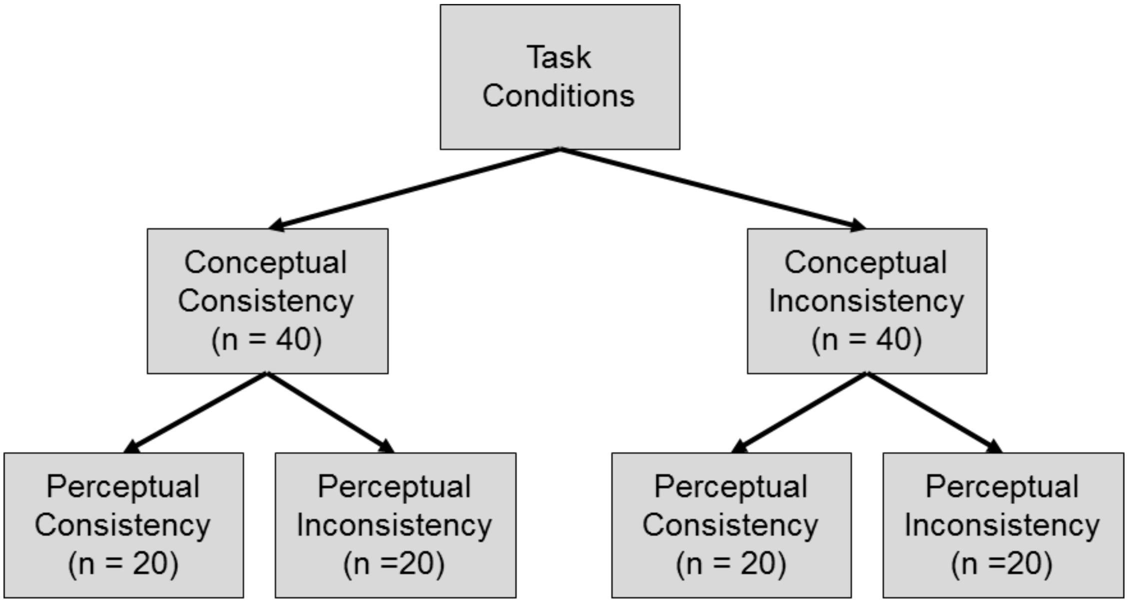


Figure 2. Experimental conditions.

METHOD

Participants

Eighty undergraduate students from Clemson University participated in the study. Participants received course credit for participation. Three participants with missing data were removed from the final results. Missing data was a result of participants not performing any actions with the system for three or more minutes. A total of 77 participants were included in the analysis. The experiment lasted approximately two hours.

Table 2. Participant demographic frequencies by condition.

Condition	Complete Consistency	Concept. Incon. Percept. Con.	Concept Con. Percept Incon.	Complete Inconsistency
Mean Age (SD)	19.8 (*SD* = 1.7)	19.9 (*SD* = 1.6)	19.8 (*SD* = 1.7)	20.4 (*SD* = 1.9)
Male	9	9	7	7
Female	10	11	13	11

Note: Pearson Chi-Squared showed no significant differences between groups.

Of the 77 participants, 32 were male and the average age was 20 (*SD* = 1.7). I used chi-squared tests to check that all four conditions did not vary in terms of demographics, cognitive abilities, or self-reported videogame experience. Participants in the four conditions did not differ in terms of age (*p* = .101) and sex (*p* = .857). These four conditions also did not differ significantly in cube comparison scores (*p* = .878), paper folding scores (*p* = .059), reverse digit span scores (*p* = .654), digit symbol substitution reaction times (*p* = .436), and digit symbol substitution scores (*p* = .653). The four conditions also did not vary in terms of whether participants considered themselves active gamers (*p* = .490), gaming frequency (*p* = .744), and gaming skill (*p* = .065).

Materials/Apparatus

Seven computer workstations running Windows XP were used in the study. The experimental program was an imagined version of a futuristic spaceship. I chose this domain to allow for freedom to manipulate the system and the tasks required. The simulator was created using RealStudio. Participant performance was continuously recorded step-by-step as they worked through tasks.

Additionally, participants completed computer administered versions of standardized abilities tests along with the NASA-TLX workload survey (see Hart, 2006). These included a paper folding test of spatial visualization (Ekstrom et al., 1976), a reverse digit span test of working memory (Wechsler, 1997), and a digit symbol substitution test to measure perceptual speed (e.g., Wechsler, 1981). Additionally, participants reported videogame experience by completing a questionnaire adapted from Maclin et al. (2011).

Starship simulator

The starship simulator was an experimental tool designed to measure the attentional demands of the dimensions of interface consistency using a multitask approach. The starship simulator bypassed constraints of existing systems in terms of design while also removing the possibility of participants having prior experience. The simulator consisted of six separate screens used to control four subsystems (navigation, shields, phasers, and life support; Figure 4).

Operating the simulator required participants to manage two separate, on-going tasks along with completing a series of discrete tasks presented by the simulator (referred to as “Captain’s Orders Task”). One on-going task involved managing the power allocation throughout the system. The other on-going task required participants to provide constant course corrections (Figure 3). Participants were be instructed to emphasize performance for the two on-going tasks (i.e., power allocation and course corrections).

The power allocation task required participants to constantly monitor the power of each of the four subsystems. Combined, the total power of these four subsystems equaled the overall system power (see Figure 4). Tasks presented by the simulator had a range of different power requirements that participants attended to. For example, firing phasers requires a certain level of phaser power and shield power. Power both drains at a constant, steady rate and as participants use it to complete tasks. Participants had to convert fuel into power to meet the power demands.

Course stability of the ship randomly fluctuated throughout the task. These fluctuations required participants to perform regular checks of the current course of the ship and make course corrections. The ship could be centered using the left and right arrow buttons to make course corrections (Figure 3). The navigation system must have power or the participant could not make course corrections. If a participant ignored this task, the ship would drift into the red area.

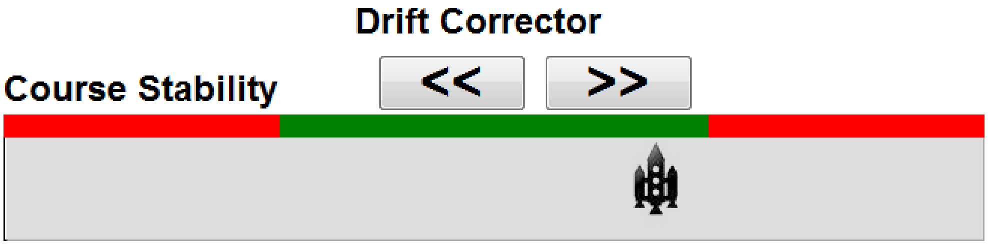
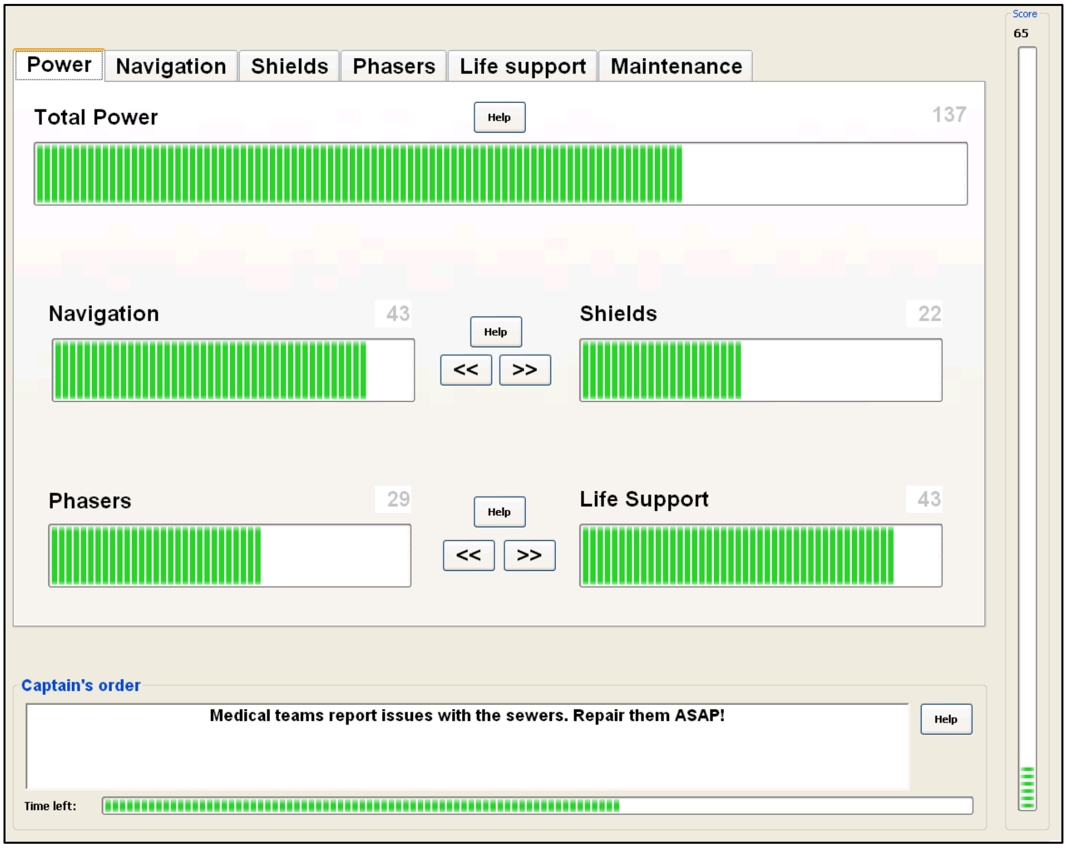


Figure 3. Course corrections task.

Failure to adequately monitor the ship’s status caused the ship to go into danger mode. Danger mode alerted participants with a pervasive warning of red highlighting along the top and bottom of the screen along with a large “Danger” label. Danger mode caused power to drain much faster and the participant’s score to decrease over time. Danger mode resulted from three different events. First, if the overall system power exceeded 200 units (i.e., participants add too much fuel). Second, if power to the Life Support system went below 10 units (i.e., participants does not add enough fuel). Third, if the course became unstable and the ship drifted into the red area (i.e., participants ignore the course stability task). These events are interdependent, meaning that a participant must first diagnose the cause of the danger mode then take appropriate actions to fix the problem(s). For example, if a participant ignored the course stability task then the ship would drift into the red area. This causes the ship’s energy to drain rapidly. Course corrections cannot be made unless the navigation system has power. To remedy this situation, a participant first must add enough fuel to the system and then quickly make course corrections to stabilize the ship.

The participants also completed a series of discrete tasks or “Captain’s orders” (see the bottom of Figure 4). These tasks required participants to complete specific orders as instructed by the “Captain” (i.e., the text at the bottom of the screen). Captain’s orders remain on the screen until completed as described; once completed, participants receive 10 points toward a final score and a new one immediately appears. Participants were instructed to complete these tasks as quickly and efficiently as possible while still placing priority on the other two tasks to avoid danger mode. All tasks required sufficient power available in the relevant subsystem. For example, firing phasers drains power from the phaser system and also requires a minimal amount of shield power before firing.

Finally, previous work (e.g., Mendel et al., 2011) suggested that the effects of a consistent interface are most apparent for highly demanding tasks. To ensure that the ship simulator task was sufficiently challenging, I increased the difficulty associated with the course corrections task by increasing the amount of the ship's navigational drift (i.e., the ship required more course corrections). This change required participants to check and adjust the ship's course routinely throughout the task. The difficulty of the course corrections task remained constant between both phases and across all four participant conditions.



Medical teams report issues with the sewers. Repair them ASAP!

Figure 4. Example starship simulator power tab.

System knowledge

System knowledge was assessed at the end of the study using a ten question multiple choice quiz about the starship simulator. Questions focused on either the way the ship worked (conceptual) or the location and arrangement of interface elements (perceptual). An example of a conceptual question is “Which of the following is the correct sequence to perform a navigation task for the first system?” An example of a perceptual question is “Which edge of the screen contained the shield frequency number pad for the first starship system?” The answer to both questions varied based on a participant’s assigned experimental condition.

Task

Participants used the starship simulator to complete tasks. The simulator consisted of three separate tasks: the captain’s orders, ship’s power allocation, and course corrections. The current captain’s order was constantly displayed until participants completed that task; the next task was displayed immediately after completing the prior task. As an example, in Figure 4, the current task reads “Medical teams report issues with the sewers. Repair them ASAP!” This task required participants to repair the sewers by clicking a button within the Life Support tab. Participants were instructed to prioritize managing the ship’s system (i.e., the power allocation and course stability tasks) while completing as many captain’s tasks as they could before time ran out. Time remaining could always be seen at the bottom of the screen as a bar labeled “Time left”.

Design

The study design was a 2 (perceptual consistency, high/low) x 2 (conceptual consistency, high/low) factorial manipulated between participants. I randomly assigned participants to one of four possible conditions with counter-balancing to control for order effects. Participants were tasked with operating versions of the starship simulator during two separate time-limited phases of 30 minutes each.

Participants began with a series of practice tasks designed to introduce them to the simulator. Participants then managed the ship's power allocation and course while working to complete as many tasks as they could. Participants worked during the first simulator phase for 30 minutes with one version of the simulator to gain a basic level of proficiency with the simulator system.

During the second phase, participants worked with a version of the starship simulator as determined by the randomly assigned experimental condition. The purpose was to determine if participants could successfully transfer skills gained from the first system during phase one when using this second system. I measured the relative attentional demands based on the performance of other tasks (e.g., the course correction task) and how they differ between each experimental condition.

Independent variables

Conceptual consistency and perceptual consistency served as the independent variables and both were manipulated between-subjects. I manipulated conceptual consistency between systems by altering the control order. One version of the system gave participants direct control over converting fuel into energy (Figure 5). The alternate version gave participants control over the rate that fuel gets converted to energy (Figure 6). The goal of this manipulation was to force participants to approach a task differently depending on the system used. The other manipulation of conceptual consistency was a change in the task sequence required. For example, one version of the system requires a participant to select the shield frequency then the shield pattern while the other version requires the opposite sequence.

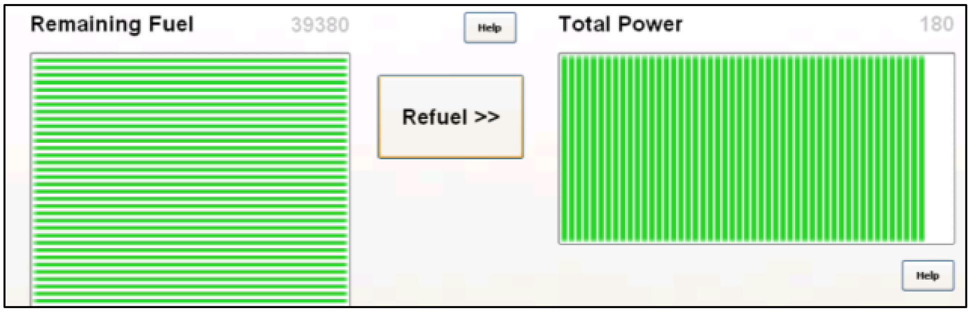


Figure 5. Fuel conversion from one conceptual variation with a single button.

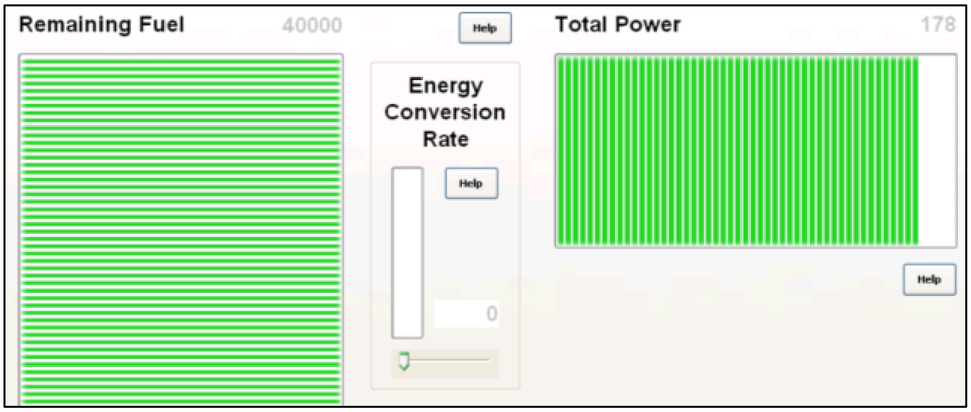
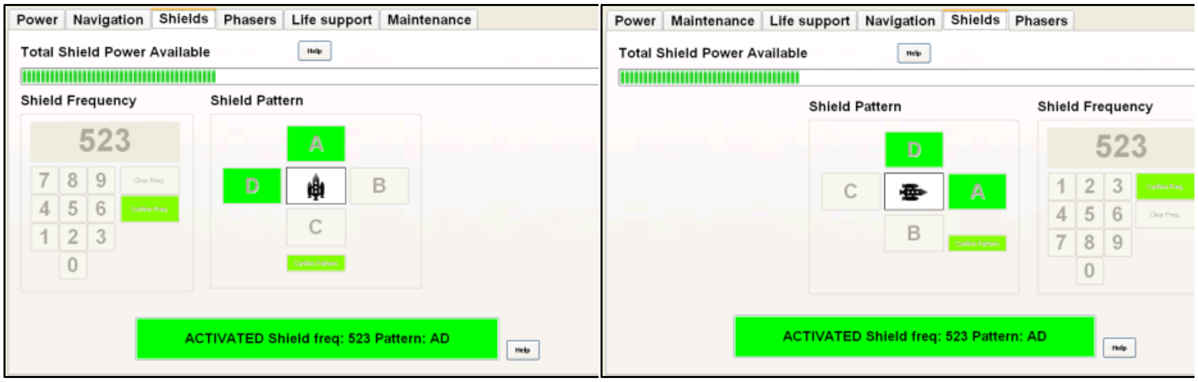


Figure 6. Fuel conversion from another conceptual variation with a rate slider.

For perceptual consistency, I focused on the location of interface elements. Previous research suggested that manipulating the location is the most influential form of perceptual consistency (AlTaboli & Abou-Zeid, 2007). In the present study, I manipulated the location and arrangement of interface elements in each of the screens (Figure 7).



ACTIVATED Shield freq: 523 Pattern: AD

ACTIVATED Shield freq: 523 Pattern: AD

Figure 7. Comparison of perceptual manipulations on the shield tab.

Dependent variables

Based on simulator performance and participant feedback, I gauged performance and workload using the following variables:

- Performance Measures
    - Number of captain's tasks completed: the total number of tasks that a participants completes in the allotted time. More were better. Each task counted equally as one task completed.
    - Number of course corrections: the amount of times a participant adjusted the ship's course for the course correction task. Fewer corrections suggested that a participant might have neglected the course correction task. I expect that more would be better but that will be tested by comparing the number of course corrections to the accuracy of the navigation task. This number was derived by counting the total number of times that a participant clicked the left or right arrow buttons (Figure 3).
    - Average variance in the ship's overall power: the average amount of variation in the power of each system between each task. If the participant ignores the power

    allocation task then the power levels would be unstable (e.g., dropping rapidly from one task to another). Ideally participants should monitor the power allocation closely and keep the levels stable. Less variability was better. This was calculated by comparing the ship's overall power at the end of each task completed to determine how large of a movement in power levels that occurred.
  - Amount of time in danger mode: the amount of time that a user spends in danger mode (a system warning displayed). This danger mode warning increases system power drain and reduces the participant's score. Lower was better. This was calculated by counting the total duration that the ship was in danger mode.
- Workload metric:
  - Subjective workload: assessed after each phase using the NASA-TLX. Lower was better.
- System knowledge:
  - Simulator system knowledge: assessed using a questionnaire designed to test a participant's understanding of the functioning of the simulator system. Higher was better.

Procedure

Experiment sessions included up to seven participants and lasted about two hours each. Participants were randomly assigned to one of four interface conditions (not including counterbalancing). I gave participants a brief overview of the study before beginning. Participants were told that the continuous tasks (i.e., power allocation and ship course corrections) should be the main focus in an effort to avoid danger mode.

Participants completed a series of introductory practice tasks designed to familiarize them with all aspects of the starship simulator. Upon completing the practice tasks, participants worked for 30 minutes to complete as many tasks as possible while managing power allocation and the course of the ship.

After the first session, participants completed a NASA-TLX survey about the perceived workload of the task. Next, participants completed a battery of computer-based abilities tests. These tests included paper folding, reverse digit span, and digit symbol substitution.

Participants then worked during a second session, again for 30 minutes, to complete tasks using another starship simulator system. The design of this second system relative to the first session’s system depended on the randomly assigned experimental condition (e.g., perceptually consistent and conceptually inconsistent between sessions). Again at the end of the session, participants filled out a NASA-TLX survey regarding the perceived workload of the task. Finally, participants completed a videogame experience questionnaire and a brief quiz designed to assess system knowledge.

RESULTS

A total of four dependent variables were used to measure performance: tasks completed, number of course corrections, ship's overall power variability, and time in danger mode. Additionally I measured system knowledge and subjective workload (NASA-TLX).

I assessed the effects of conceptual and perceptual consistency on performance in three ways. First, I compared performance during the entirety of phase 2. Second, I analyzed performance during the first and last six minutes of phase 2 to understand how manipulations of interface consistency differed as a function of time (e.g., how the effects of conceptual and perceptual consistency might vary over time). Third, and finally, I compared performance for the condition of conceptual inconsistency with perceptual consistency (hypothesized to be the worst) to that of complete consistency and also complete inconsistency. Additionally I included analyses for performance immediately before and immediately after the manipulations in APPENDIX F**: Additional Analyses**.

In addition to performance data, I also analyzed system knowledge scores and subjective workload (NASA-TLX). System knowledge scores were measured at the end of the study and were compared between interface conditions. Subjective workload was analyzed both within-groups (i.e., how workload changed from phase 1 to phase 2) and between-groups (i.e., how subjective workload varied by condition).

Before beginning the analyses I wanted to determine how the course correction variable related to the actual performance on the course stability task (i.e., was more course corrections associated with better performance). I ran a correlation between the two variables and found that more course corrections was significantly correlated with better performance on the course stability task $r(75) = .31$, $p = .006$. This relationship indicated that more course corrections tended to occur with better performance in keeping the ship's navigation centered.

Performance during Phase 2

These analyses were designed to assess how conceptual and perceptual consistency manipulations influenced performance during phase 2 (i.e., once the manipulations were implemented). To assess condition differences in performance during phase 2, a 2 (conceptual consistency; high/low) × 2 (perceptual consistency; high/low) between-groups ANOVA was used. Descriptive statistics for performance during phase 1 and phase 2 are in Table 3 and Table 4 respectively.

Tasks completed

Conceptual consistency did not affect the number of tasks completed during phase 2 ($p = .481$). Perceptual consistency had a significant effect in that the perceptually consistent group completed significantly more tasks during phase 2 than the perceptually inconsistent group $F(1,73) = 15.1$, $p < .001$, $\eta_p^2 = .17$. The interaction of conceptual consistency × perceptual consistency was not significant ($p = .569$).

Number of course corrections

There was a significant main effect of conceptual consistency $F(1,73) = 7.7$, $p = .007$, $\eta_p^2 = .10$. Participants in the conceptually consistent group made more course corrections than the conceptually inconsistent group. Perceptual consistency did not have a significant effect on the number of course corrections during phase 2 ($p = .931$).

Additionally, the conceptual consistency × perceptual consistency interaction was significant $F(1,73) = 5.6$, $p = .021$, $\eta_p^2 = .07$. A follow-up analysis indicated that participants in the conceptual inconsistency and perceptual consistency condition made fewer course corrections than participants in the conceptual consistency and perceptual consistency condition $F(1,37) = 15.7$, $p < .001$, $\eta_p^2 = .30$. In contrast, conceptual consistency had no effect on the number of course corrections for the perceptual inconsistency condition ($p = .783$).

Ship's power variability

Neither conceptual consistency ($p = .527$) nor perceptual consistency ($p = .475$) had a significant effect on the ship's power variability during phase 2. The interaction of conceptual consistency × perceptual consistency was also non-significant ($p = .774$).

Time in danger mode

Neither conceptual consistency ($p = .617$) nor perceptual consistency ($p = .650$) had a significant effect on the amount of time in danger mode during phase 2. The interaction of conceptual consistency × perceptual consistency was also non-significant ($p = .636$).

Table 3. Descriptive Statistics Phase 1.

	Conceptually Inconsistent				Conceptually Consistent			
	Perceptually Inconsistent		Perceptually Consistent		Perceptually Inconsistent		Perceptually Consistent	
Measure	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Tasks Completed	115.3	31.9	125.2	34.1	116.3	26.7	111.4	22.5
Course Corrections	555.2	138.9	503.0	147.2	532.8	153.6	563.1	184.9
Overall Power Variability	19.7	7.9	22.8	6.9	22.7	6.6	20.8	5.1
Time in Danger Mode (seconds)	258.4	177.2	243.6	177.6	256.0	132.5	227.5	180.9

Table 4. Descriptive Statistics Phase 2.

	Conceptually Inconsistent				Conceptually Consistent			
	Perceptually Inconsistent		Perceptually Consistent		Perceptually Inconsistent		Perceptually Consistent	
Measure	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Tasks Completed	105.1	23.6	136.0	38.5	114.0	29.1	136.9	27.7
Course Corrections	547.8	156.3	462.9	156.7	562.6	169.3	641.5	122.1
Overall Power Variability	28.5	28.1	31.4	40.5	22.3	11.7	29.1	31.1
Time in Danger Mode (seconds)	292.9	161.4	293.8	166.3	339.7	209.3	295.1	280.3

Performance at the Beginning and the End of Phase 2

The goal of these analyses was to determine how the effects of conceptual and perceptual consistency changed as a function of time. To measure this, I contrasted the effects of conceptual and perceptual consistency during the initial six minutes of phase 2 and the final six minutes of phase 2. This gave an indication of how the immediate effects of consistency compared to the effects after a longer exposure. I used a 2 (conceptual consistency; high/low) × 2 (perceptual consistency; high/low) × 2 (time segment; beginning of phase 2/end of phase 2) mixed factorial ANOVA to compare performance between these two time segments (conceptual and perceptual manipulations were between-group). Descriptive statistics for performance during the beginning of phase 2 and the end of phase 2 are in Table 5 and Table 6 respectively.

Tasks Completed

Data for the number of tasks completed showed a significant interaction of conceptual consistency × time segment $F(1,73) = 7.8$, $p = .007$, $\eta_p^2 = .10$. The perceptual consistency × time segment interaction was not significant (p = .068). The interaction of conceptual consistency × perceptual consistency was not significant (p = .872).

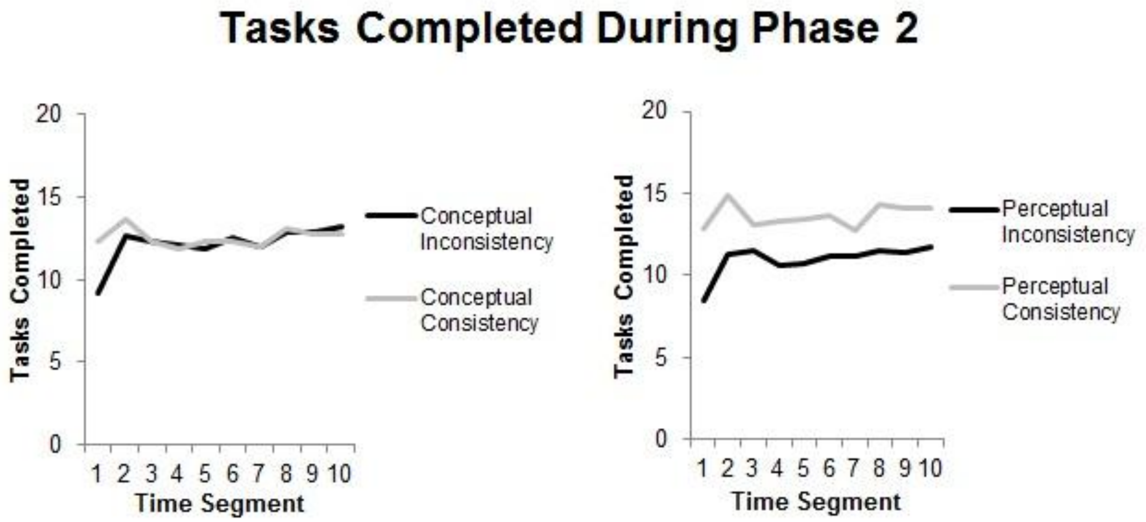


Figure 8. Tasks completed during phase 2 by type of consistency manipulation.
Note: each time segment is three minutes in duration.

The source of the conceptual consistency × time segment interaction was that the conceptual inconsistency group completed fewer tasks during the beginning of phase 2 $F(1,36) = 9.4$, $p = .004$, $\eta_p^2 = .21$. During the last six minutes of phase 2, however, conceptual consistency had no effect in terms of tasks completed (p = .637). There was also a main effect of perceptual consistency in that the perceptual consistency group completed significantly more tasks regardless of time segment $F(1,73) = 20.3$, $p < .001$, $\eta_p^2 = .22$. Additionally, the main effect of time segment was significant, indicating that participants completed more tasks during the last six minutes of phase 2 than during the first six minutes of phase 2 $F(1,73) = 5.0$, $p = .028$, $\eta_p^2 = .07$.

Number of course corrections

Results for the number of course corrections indicated a significant interaction of conceptual consistency × perceptual consistency $F(1,73) = 3.2$, $p = .039$, $\eta_p^2 = .04$ (1-tailed). A follow-up analysis found that participants in the conceptual inconsistency and perceptual consistency condition made fewer course corrections than participants in the conceptual consistency and perceptual consistency condition $F(1,37) = 16.8$, $p < .001$, $\eta_p^2 = .31$. In contrast, conceptual consistency had no effect on the number of course corrections for the perceptual inconsistency condition ($p = .700$).

The conceptual consistency × time segment interaction ($p = .094$) and the perceptual consistency × time segment interaction ($p = .164$) were non-significant. There was a significant main effect of conceptual consistency in that the conceptual consistency made more course corrections than the conceptual inconsistency group regardless of time segment $F(1,73) = 6.0$, $p = .016$, $\eta_p^2 = .08$. The main effect of perceptual consistency was not significant ($p = .220$). The main effect of time segment was significant, in that participants made more course corrections during the last six minutes of phase 2 than during the first six minutes of phase 2 $F(1,73) = 141.1$, $p < .001$, $\eta_p^2 = .66$.

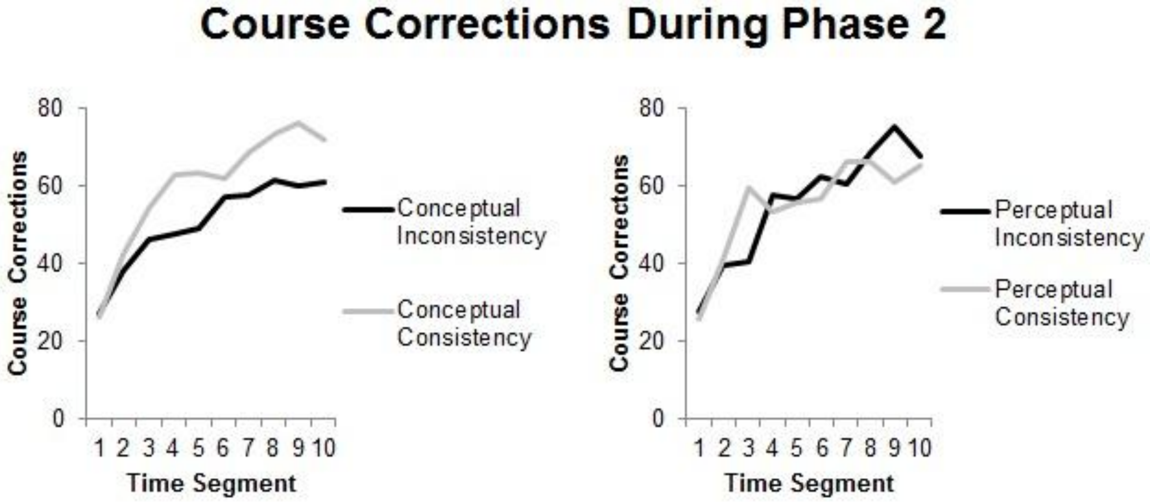


Figure 9. Course corrections during phase 2 by type of consistency manipulation.
Note: each time segment is three minutes in duration.

Ship's power variability

Data on the amount of overall power variability showed that the conceptual consistency × time segment interaction (p = .390), the perceptual consistency × time segment interaction (p = .194), and the conceptual consistency × perceptual consistency (p = .373) were all non-significant. Additionally, the main effects of conceptual consistency (p = .158) and perceptual consistency (p = .226) were also non-significant. The main effect of time segment, however, was significant $F(1,73) = 5.0$, $p = .028$, $\eta_p^2 = .07$. Participants kept ship power more stable (i.e., less variability) during the last six minutes of phase 2 than during the first six minutes of phase 2.

Time in danger mode

Data for the amount of time in danger mode of showed a significant interaction of conceptual consistency × time segment $F(1,73) = 6.6$, $p = .012$, $\eta_p^2 = .08$ and a significant interaction of perceptual consistency × time segment $F(1,73) = 4.4$, $p = .040$, $\eta_p^2 = .06$.

The interaction of conceptual consistency × perceptual consistency, however, was not significant (p = .671).

Additional analyses of the conceptual consistency × time segment interaction indicated that its source was that the conceptual consistency group spent less time in danger mode than the conceptual inconsistency group, but only during the initial six minutes $F(1,73) = 3.5$, $p = .032$, $\eta_p^2 = .05$ (1-tailed). By the end of phase 2, during the last six minutes, performance was equivalent regardless of conceptual consistency (p = .495). Similarly, the source of the interaction of perceptual consistency × time segment was that the perceptual consistency group spent less time in danger mode than the perceptual inconsistency group but only during the first six minutes of phase 2 $F(1,73) = 4.5$, $p = .038$, $\eta_p^2 = .06$. Both the main effects for conceptual consistency (p = .593) and perceptual consistency (p = .302) were non-significant. The main effect of time segment was significant with participants spending more time in danger mode during the last six minutes of phase 2 than during the first six minutes of phase 2 $F(1,73) = 19.5$, $p < .001$, $\eta_p^2 = .21$.

Table 5. Descriptive Statistics for First Six Minutes of Phase 2.

	Conceptually Inconsistent				Conceptually Consistent			
	Perceptually Inconsistent		Perceptually Consistent		Perceptually Inconsistent		Perceptually Consistent	
Measure	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Tasks Completed	18.2	6.0	25.0	8.9	21.2	6.1	30.7	4.8
Course Corrections	65.1	26.2	60.6	21.3	66.5	25.3	71.3	21.8
Overall Power Variability	63.5	133.7	30.9	10.7	32.7	25.8	26.1	8.0
Time in Danger Mode (seconds)	68.0	56.2	50.5	38.0	52.9	48.0	26.6	37.3

Table 6. Descriptive Statistics for Last Six Minutes of Phase 2.

	Conceptually Inconsistent				Conceptually Consistent			
	Perceptually Inconsistent		Perceptually Consistent		Perceptually Inconsistent		Perceptually Consistent	
Measure	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Tasks Completed	22.4	7.0	29.3	9.8	24.0	8.4	27.1	6.5
Course Corrections	139.6	60.3	104.9	37.3	146.9	57.5	149.2	34.1
Overall Power Variability	23.8	15.2	23.1	22.0	18.1	12.1	19.6	7.4
Time in Danger Mode (seconds)	66.7	45.7	71.8	51.5	79.5	44.0	76.2	73.7

Comparison by Interface Conditions

I compared participant performance in the different interface consistency conditions using two separate ANOVAs to compare the performance of the conceptual inconsistency paired with perceptual consistency to the performance of complete consistency and complete inconsistency. In both cases, participant performance was assessed using a 2 (interface condition) × 2 (time segment; beginning of phase 2/end of phase 2) mixed factorial ANOVA (interface condition was between-group). Descriptive statistics are available in Table 7 and Table 8.

Tasks completed

Participants in the complete consistency condition completed significantly more tasks than those in the conceptual inconsistency/perceptual consistency condition but only during the beginning of phase 2 $F(1,37) = 6.4$, $p = .016$, $\eta_p^2 = .15$. In comparison, the participants in the conceptual inconsistency/perceptual consistency condition completed more tasks than those in the complete inconsistency condition across all of phase 2 $F(1,36) = 9.2$, $p = .004$, $\eta_p^2 = .20$.

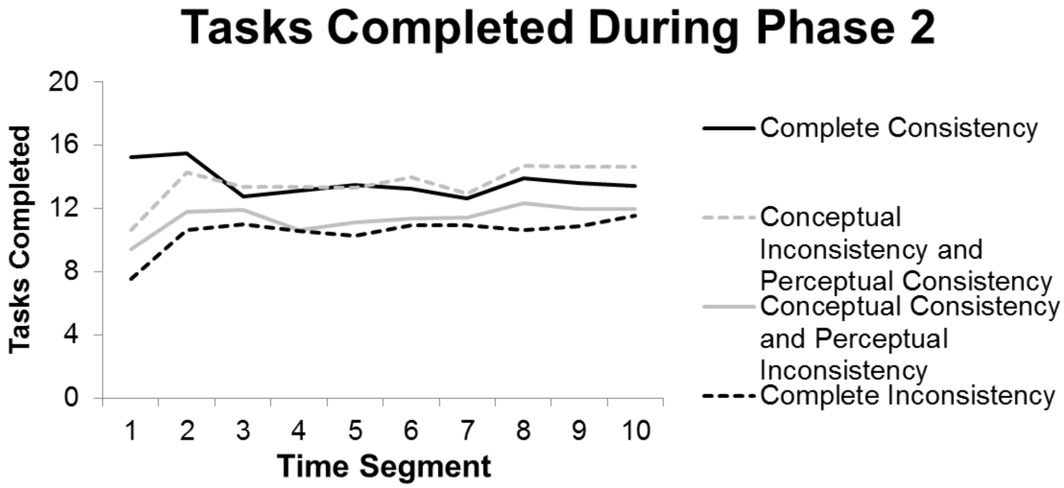


Figure 10. Tasks completed during phase 2 by interface condition.
Note: each time segment is three minutes in duration.

Number of course corrections

Participants in the complete consistency condition made more course corrections than those in the conceptual inconsistency/perceptual consistency condition during all of phase 2 $F(1,37) = 16.8$, $p < .001$, $\eta_p^2 = .31$. Similarly, participants in the complete inconsistency also made more course corrections than those in the conceptual inconsistency/perceptual consistency condition but only during the end of phase 2 $F(1,36) = 4.6$, $p = .038$, $\eta_p^2 = .11$.

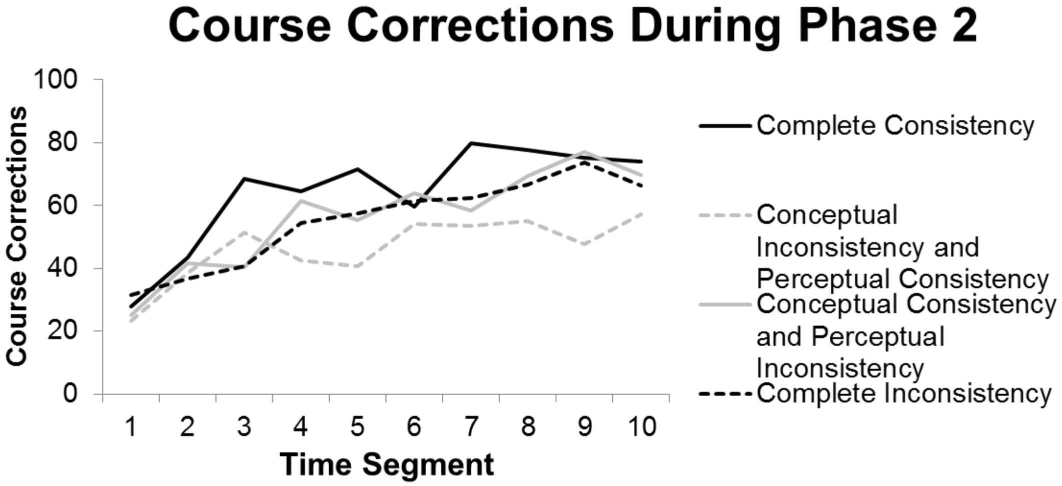


Figure 11. Course corrections during phase 2 by interface condition.
Note: each time segment is three minutes in duration.

Ship's power variability

There were no significant differences detected neither between complete consistency and conceptual inconsistency /perceptual consistency nor between complete inconsistency and conceptual inconsistency/perceptual consistency in terms of the ship's power variability ($p > .05$).

Time in danger mode

Participants in the complete consistency condition spent less time in danger mode than those in the conceptual inconsistency/perceptual consistency condition but only during the end of phase 2 $F(1,37) = 3.9$, $p = .028$, $\eta_p^2 = .10$ (1-tailed). There was no difference in terms of time in danger mode for participants in the complete inconsistency and conceptual inconsistency/perceptual consistency condition ($p > .05$).

Table 7. Descriptive Statistics by Interface Condition for First Six Minutes of Phase 2.

	Complete Consistency		Conceptual Incon. Perceptual Con.		Complete Inconsistency	
Measure	Mean	SD	Mean	SD	Mean	SD
Tasks Completed	30.7	4.8	25.0	8.9	18.2	6.0
Course Corrections	71.3	21.8	60.6	21.3	65.1	26.2
Overall Power Variability	26.1	8.0	30.9	10.7	63.5	133.7
Time in Danger Mode (seconds)	26.6	37.3	50.5	38.0	68.0	56.2

Table 8. Descriptive Statistics by Interface Condition for Last Six Minutes of Phase 2.

	Complete Consistency		Conceptual Incon. Perceptual Con.		Complete Inconsistency	
Measure	Mean	SD	Mean	SD	Mean	SD
Tasks Completed	27.1	6.5	29.3	9.8	22.4	7.0
Course Corrections	149.2	34.1	104.9	37.3	139.6	60.3
Overall Power Variability	19.6	7.4	23.1	22.0	23.8	15.2
Time in Danger Mode (seconds)	76.2	73.7	71.8	51.5	66.7	45.7

System Knowledge

System knowledge scores were assessed using a 2 (conceptual consistency; high/low) × 2 (perceptual consistency; high/low) between-group ANOVA. The conceptual consistency manipulation had no effect on overall system knowledge scores ($p = .982$) or on the scores for only the conceptual questions ($p = .625$). Perceptual consistency had no effect on the overall scores ($p = .437$); however, when considering only the perceptual questions, the perceptual consistency group answered significantly more questions correctly than the perceptual inconsistency group $F(1,73) = 5.3$, $p = .024$, $\eta_p^2 = .07$.

Subjective Workload

Subjective workload scores were analyzed using a 2 (conceptual consistency; high/low) × 2 (perceptual consistency; high/low) × 2 (phase; phase 1/phase 2) mixed factorial ANOVA (conceptual and perceptual manipulations were between group). I conducted analyses for the weighted overall workload measure along with each of the separate component measures (e.g., mental workload). I analyzed differences both within-group (i.e., did participant workload change between phases) and between-group (i.e., did the manipulations influence subjective workload for phase 2).

Table 9. Descriptive Statistics for Subjective Workload after Phase 1.

	Conceptually Inconsistent				Conceptually Consistent			
	Perceptually Inconsistent		Perceptually Consistent		Perceptually Inconsistent		Perceptually Consistent	
Workload Measure	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Computed Overall	64.4	12.0	56.4	18.5	58.9	16.4	66.4	13.7
Mental	73.1	19.0	65.0	25.4	60.5	21.5	70.0	20.3
Physical	32.2	25.3	30.5	22.7	34.0	27.5	34.7	22.8
Temporal	75.0	15.9	67.3	23.3	69.5	20.4	80.8	14.1
Effort	61.7	19.8	56.0	25.0	59.5	21.4	67.9	18.0
Performance	49.2	21.7	36.8	22.0	38.0	21.4	51.3	26.2
Frustration	56.7	23.9	49.3	28.3	50.0	21.7	61.6	25.1

Table 10. Descriptive Statistics for Subjective Workload after Phase 2.

	Conceptually Inconsistent				Conceptually Consistent			
	Perceptually Inconsistent		Perceptually Consistent		Perceptually Inconsistent		Perceptually Consistent	
Workload Measure	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Computed Overall	67.1	11.0	63.4	20.5	65.6	20.6	61.3	14.6
Mental	73.3	16.4	66.8	20.9	66.3	25.6	62.1	18.8
Physical	32.5	22.4	40.3	31.0	39.3	31.5	42.6	26.8
Temporal	68.3	19.0	62.8	25.4	69.0	23.4	72.9	19.2
Effort	71.1	20.0	59.8	26.2	66.5	24.9	62.9	20.8
Performance	54.2	23.4	51.8	30.0	46.8	22.7	44.5	25.9
Frustration	59.2	29.4	69.5	24.7	62.0	28.6	61.8	24.2

Computed overall workload

There was a three-way interaction of conceptual consistency × perceptual consistency × phase for the total workload measure $F(1,73) = 5.0$, $p = .029$, $\eta_p^2 = .06$. The source of the three-way interaction was an increase in overall computed workload from phase 1 to phase 2, but only for the group with the combination of conceptual consistency and perceptual inconsistency $F(1,19) = 8.7$, $p = .008$, $\eta_p^2 = .31$ (Figure 12). The two-way interactions were all non-significant for total workload ($p > .05$). The main effects for conceptual consistency, perceptual consistency, and phase were also all non-significant ($p > .05$).

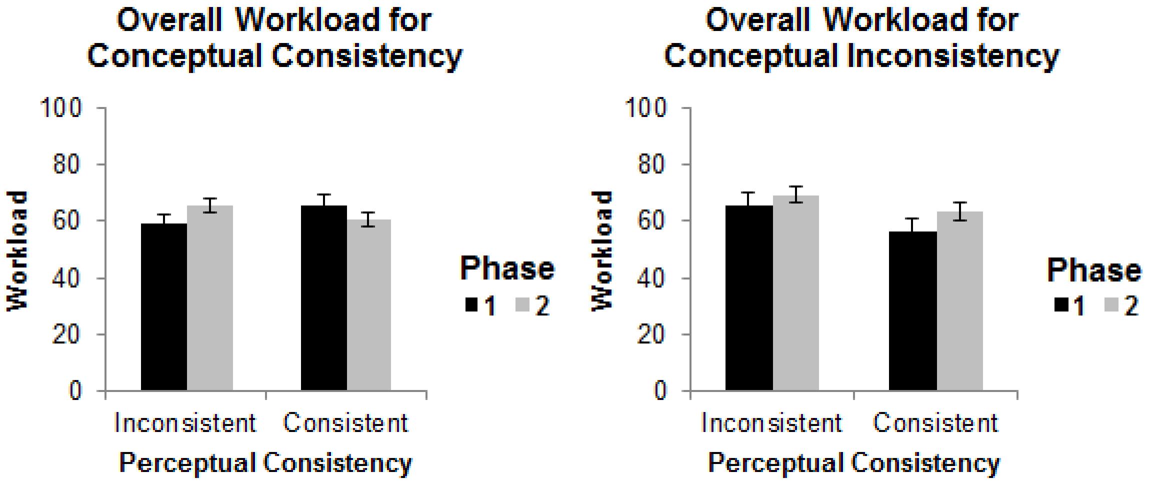


Figure 12. Interactions of perceptual consistency and phase for overall workload split by conceptual consistency.

NASA-TLX subscales

There was a three-way interaction of conceptual consistency × perceptual consistency × phase for the frustration workload measure $F(1,73) = 4.0$, $p = .049$, $\eta_p^2 = .05$. The source of the three-way interaction was an increase in frustration workload from phase 1 to phase 2, but only when the system was conceptually inconsistent and perceptually consistent *F*(1,19) = 7.5, *p* = .013, η_p^2 = .28 (Figure 13). There was also a significant main effect of phase indicating that participants rated frustration workload higher for phase 2 than for phase 1 *F*(1,73) = 5.6, *p* = .020, η_p^2 = .07. The two-way interactions were non-significant for frustration workload ($p > .05$). Main effects for conceptual consistency and perceptual consistency were also non-significant ($p > .05$).

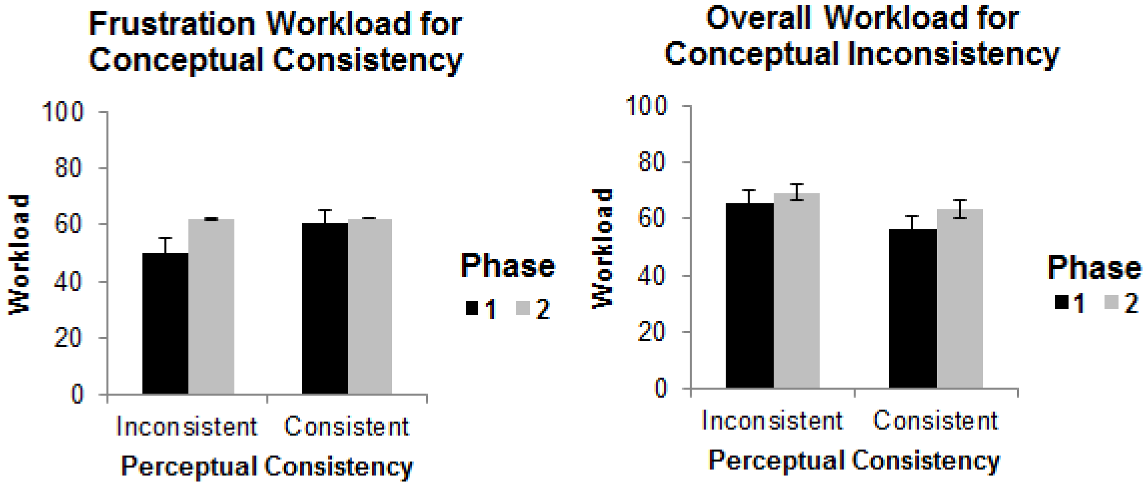


Figure 13. Interactions of conceptual consistency and perceptual consistency for frustration workload split by phase.

There was a main effect of phase for physical workload *F*(1,73) = 5.8, *p* = .019, η_p^2 = .07 indicating that participants rated physical workload higher for phase 2 than for phase 1. There was also a main effect for temporal workload *F*(1,73) = 5.1, *p* = .027, η_p^2 = .07 with participants rating temporal workload lower for phase 2 than for phase 1. Effects for mental workload, effort workload, and performance workload were all non-significant ($p > .05$).

DISCUSSION

This study was motivated by the contradicting results from previous research. Past studies found results ranging from beneficial effects (e.g., AlTaboli & Abou-Zeid, 2007), detrimental effects (e.g., Satzinger & Olfman, 1998), and limited or non-significant effects (e.g., Rhee et al., 2006) of interface consistency. The present study sought to explore the effects of the two dimensions of consistency. Past research failed to accomplish this by ignoring the interrelated nature of the two dimensions; instead, the detrimental effects of incomplete consistency may have contaminated the results of past studies. The present study clarified the effects of the two dimensions while considering how the forms of incomplete consistency contribute to worse performance and higher workload. Additionally, previous research only measured outcomes in terms of performance. In contrast, the present study utilized a multitask approach to more accurately assess the effects of interface consistency. This approach allowed for a direct comparison of the relative demands required to cope with different forms of interface consistency (e.g., conceptual inconsistency and perceptual consistency versus conceptual inconsistency and perceptual inconsistency).

As a summary, in the current study participants completed two 30 minute sessions using a starship simulator designed with three interdependent tasks. Participant’s primary tasks were the course correction and power management tasks; the secondary task was to complete captain’s orders. Depending on the experimental condition, the interface of the simulator was manipulated between the two sessions. I manipulated both the conceptual consistency and perceptual consistency of the simulator between the 30 minute sessions. I manipulated conceptual consistency by altering the control order of one of the simulator's systems. I manipulated perceptual consistency by altering the layout of interface elements.

Summary of Effects

Interactions of conceptual and perceptual consistency

For performance variables, the only interaction of the two dimensions occurred for the number of course corrections. The number of course corrections was an important indicator of the attentional demands of the system; an overloaded participant would be unable to make constant course corrections. Results indicated that the combination of conceptual inconsistency and perceptual consistency performed the worst in regard to the number of course corrections. This means that a system that operates differently but looks the same was more demanding. One explanation for this is that the participants expected the system to operate similarly since it looked analogous. Instead, the design of the system might have misled participants into making inappropriate generalizations. In addition to performance issues, participants also rated this version of the system as having the highest level of overall workload.

Operates differently but looks the same

As expected, participants in the complete consistency condition performed better than those in the condition with the system that operated differently but looked the same (i.e., conceptual inconsistency with perceptual consistency). Throughout all of phase 2, the complete consistency condition made more course corrections. Also, during the first six minutes participants in the the complete consistency condition completed more tasks and spent less time in danger mode.

Interestingly, at least for the course correction task, the system that operated differently but looked the same seemed to be even worse than the completely inconsistency system (i.e., both dimensions were inconsistent). This finding suggests that forcing a system to appear consistent when it is functionally inconsistent is more demanding than leaving the system completely inconsistent. From this it seems that perceptual consistency alone can be detrimental.

Conceptual consistency's effects

The performance effects of a conceptually consistent interface were most prominent during the beginning of phase 2 (i.e., immediately after manipulations). Performance in terms of course corrections benefited from conceptual consistency during all of phase 2. In contrast, tasks completed, and time in danger mode only benefited from conceptual consistency during the early part of phase 2. A possible explanation for these findings is that the initial struggle of using a conceptually inconsistent system resulted in reduced performance overall. As participants continued to use the new system, participants recovered in terms of the secondary task performance (i.e., the captain's orders task). Coping with the conceptually inconsistent system, however, seemed to increase the attentional demands of the system, causing participants to neglect the primary task of keeping the ship on course. This would explain why the conceptual inconsistency condition had relatively good performance on the secondary task but fewer course corrections overall.

Conceptual consistency had no effect on system knowledge scores. Participants in the conceptually consistent groups, on average, did not do any better answering system knowledge questions, overall and just for the conceptually-based questions. The lack of effect might be because both conceptual variations of the system required the same basic understanding of the system. Maybe the differences were not enough for the questions to detect; conversely, maybe the conceptually-based questions did not focus well enough on the differences between the two conceptual variations of the system.

Perceptual consistency's effects

Perceptual consistency improved performance for the secondary task (i.e., the captain's orders task) throughout all of phase 2. The perceptual consistency condition performed better in terms of tasks completed. Performance for the primary tasks (i.e., course corrections and power management tasks) were unaffected by perceptual consistency manipulations. The only exception to this is that the perceptual consistency group spent less time in danger mode than the perceptual inconsistency group, but only during the beginning of phase 2. Based on these results, it seems that perceptually-based manipulations only affected performance for the task that relied more on rapid visual searches. It is also interesting that the effects of a perceptually consistent interface seem to persist relatively longer. In this case, the performance effects continued throughout the 30 minutes.

The perceptual consistency group did no better in terms of overall system knowledge scores; however, the perceptual consistency condition did perform better for only the perceptually-based questions. This makes sense since the perceptually-based questions tested participant knowledge on the arrangement of the system; if the system arrangement varied then it would be harder to correctly recall interface element locations.

Effects on subjective workload

As expected, conditions resulting in incomplete consistency had worse (i.e., higher workload) NASA-TLX scores. Participants felt that the combination of conceptual inconsistency and perceptual consistency demanded higher levels of overall workload. This is when the system looks the same but functions differently. Participants also reported higher frustration with the other form of incomplete consistency, when the system looks different but operates the same. These negative perceptions support the notion that forcing consistency inappropriately (i.e., incomplete consistency) might be worse than a system that is entirely inconsistent.

Overall, I expected greater effects on the subjective workload given that previous work found greater beneficial effects of a consistent interface (Mendel et al., 2011). One possible explanation for the results from the subjective workload scores is that participants might have considered the task so demanding to begin with that any additional demands like coping with inconsistency seemed relatively minimal (i.e., workload going from high to slightly higher). Another explanation is that the already high workload of phase 1 left little room on the scale for increases due to inconsistency in phase 2 (similar to a ceiling effect).

Study Hypotheses

Hypothesis 1

I hypothesized that the effects of a consistent interface would be greatest immediately after the interface consistency manipulations. Further, I expected that those effects would diminish with time. This hypothesis was partially supported. Conceptual consistency influenced participant performance for number of captain's orders completed and time in danger mode only at the beginning of phase 2 (i.e., immediately after the manipulations of interface consistency). By the end of phase 2, those effects of a conceptually inconsistent were no longer significant. The effect on course corrections remained constant throughout phase 2. Similarly, the effects of perceptual consistency remained constant throughout phase 2.

These findings were surprising as I expected that participants would better cope with the inconsistencies as they used the systems. Instead, even after 30 minutes, participants were still struggling with the inconsistencies, especially the perceptual inconsistency manipulations. Based on this study, it is unclear how long it would take participants to recover from the two different forms of inconsistency. A longer study could help to determine the duration of these effects.

The short-lived effects of conceptual consistency could explain why past studies (e.g., Rhee et al., 2006) did not observe any effects of a consistent interface. In Rhee et al. the authors sampled performance across the entire session. Since some effects of inconsistency seem to be short-lived, averaging performance across a single time-period could have obscured the effects of interface consistency. Conceptual consistency manipulations may be especially prone to this since they appear to dissipate faster than perceptual consistency manipulations. This could also help explain why previous studies (e.g., Ozok & Salvendy, 2000) found a significant effect of perceptual consistency but not conceptual consistency.

Hypothesis 2

The final hypothesis was that incomplete consistency (i.e., one dimension is consistent and the other is inconsistency) would be harmful. I expected that participants in the incomplete consistency condition would perform even worse than those in the complete inconsistency condition (i.e., both dimensions are inconsistent) for some performance measures. This hypothesis was supported in the present study. As expected, participants using the combination of conceptual inconsistency and perceptual consistency (i.e., the system operates differently but looks similar) performed especially poorly. Participants in that incomplete consistency condition had worse outcomes than the participants in the complete inconsistency condition. The condition of conceptual inconsistency and perceptual consistency had the lowest performance in terms of course corrections and reported the largest increases in overall workload from phase 1 to phase 2.

These results could help to explain why some prior research found that a consistent interface could result in worse performance (e.g., Finstad, 2008; Satzinger & Olfman, 1998). In those studies, however, only specific forms of interface consistency seemed to be detrimental. Satzinger and Olfman concluded that conceptual consistency improved task efficiency (i.e., fewer steps required) but that perceptual consistency reduced accuracy (i.e., more errors). If the two dimensions of interface consistency were not properly controlled then perceptual consistency may have appeared to be detrimental because it was paired with conceptual inconsistency, resulting in incomplete consistency. As seen in the present study, participants using a system with this form of incomplete consistency (i.e., operates differently and looks the same) performed especially poorly; for some performance variables, participants in that condition did even worse than those in the complete inconsistency condition (i.e., operates differently and looks differently). Based on the findings of the present study it seems that a consistent interface is beneficial only if the consistency is properly implemented (i.e., completely consistent). If, however, the dimensions of interface consistency are not considered in conjunction, then the effects of incomplete consistency (i.e., one dimension is consistent and the other is inconsistent) could make it appear as though interface consistency was detrimental.

Role in the Literature

Previous studies found limited (Ozok & Salvendy, 2000) and even no effects of interface consistency (Rhee et al., 2006). Further, in past research, perceptual consistency tended to have a relatively greater influence on performance. The findings from the present study did not follow this pattern. For example, Ozok and Salvendy (2000) found that a perceptually consistent interface improved performance while conceptual manipulations had no significant effects. Other studies focused solely on the role of perceptual consistencies and found that a perceptually consistent interface was beneficial for performance (AlTaboli & Abou-Zeid, 2007; Mendel et al., 2011).

One possible explanation for this discrepant finding is that many previous studies (e.g., Ozok & Salvendy, 2000; Rhee et al., 2006) used the three-dimensional model of interface consistency as originally described by Kellogg (1987). The three dimensional model may not be as effective in measuring the effects of conceptual consistency since it splits conceptual consistency into two separate, ill-defined dimensions. Splitting up the dimension of conceptual consistency in such a manner may have obscured or mitigated the benefits associated with conceptual consistency for the studies by Ozok and Salvendy and Rhee et al. In contrast, the three-dimensional model's perceptual consistency equivalent (referred to as physical consistency) is much more clearly defined. The better operationalization of physical consistency may explain why Ozok and Salvendy only found a significant effect of physical consistency and no effects from the other two dimensions.

Another possible explanation for these findings is that conceptual consistency and perceptual consistency react differently depending on the type of task. The primary tasks (i.e., the course corrections and the ship's power tasks) relied on an understanding of how the interrelated ship's systems functioned. The conceptual consistency manipulation altered the functionality of the ship's power management system by changing the control order of the fuel conversion (i.e., direct control versus rate control).If a participant failed to keep the ship's course stable then the ship went into danger mode, causing the power to drain rapidly. Conversely, if the ship's power dropped too low then the participant was unable to make course corrections until after they successfully restored power. Conceptual inconsistency seemed to reduce performance on this task since it relied heavily on a conceptual understanding of the ship’s systems. For example, if the participant did not understand how to set the fuel conversion rate slider to achieve a steady flow of power, then managing the ship’s power was extremely challenging.

In comparison, the secondary task of following the captain’s orders was relatively less complicated since participants could follow the sequence of steps described in the captain’s orders text box. As long as the ship had minimal power, participants could continue to complete captain’s orders tasks. Surprisingly, performance for the captain’s orders task was not influenced much by conceptual inconsistencies in the power management system. Perceptual inconsistencies, however, had a robust effect for performance on the captain’s orders task. This makes sense since performance on the captain’s orders task was constrained by how quickly a participant could locate the appropriate button and click it; rearranging the well-learned button layouts, as in perceptual inconsistency conditions, would make the visual search task more challenging. Perceptual inconsistency reduced performance most for the captain’s orders since this task was essentially a simple visual search tasks with well-learned action sequences. See Figure 7 for an example of the perceptual manipulations.

Implications of the Current Study

From these results, the most notable findings can be separated into three main ideas. First, in a multitasking situation, inconsistencies in one task can affect performance on another task. Second, incomplete consistency contributed to worse performance and higher subjective workload than complete inconsistency. Third, the effects of conceptual and perceptual consistency have different time courses (i.e., the effects of perceptual inconsistency persist longer).

Interface consistency's effects in multitasking

Both conceptual and perceptual inconsistency reduced performance for one of the primary tasks (the course corrections task) and for the secondary task (the captain's orders task). This suggests a greater attentional demand for inconsistent systems. Participants had to devote more resources to deal with the inconsistencies, leading to reduced performance on the primary task. Participants in the conceptually inconsistent condition made fewer course corrections.

This highlights the importance of consistency for all components in a multitask situation. For example, even inconsistency in a seemingly minor secondary system like a GPS system could disrupt performance of your primary task of driving. Inconsistencies between systems, especially conceptual inconsistencies, could be the extra distraction that contributes to an accident. Even without considering accidents, inconsistency in secondary systems could still reduce performance (e.g., more lane swerving).

Incomplete consistency worse than inconsistency

In some cases, participants in the incomplete consistency conditions performed worse than those in the complete inconsistency condition. The combination of conceptual inconsistency and perceptual consistency performed especially poorly (i.e., when the system operated differently but looked similar). Results showed that this form of incomplete consistency hurt performance for the primary course corrections task and increased subjective overall workload. The other variation of incomplete consistency, when a system operates similarly but looks different, increased participant frustration (i.e., conceptual consistency and perceptual inconsistency). The source of this increased frustration might due to the "willfully arbitrary" design meaning that the system looks different for no good reason. A functionally similar system with a seemingly arbitrary appearance might only irritate users since the perceptual inconsistencies seem irrational. Interestingly, this form of incomplete consistency did not harm performance, it only harmed user perceptions.

These results provide support for the importance of addressing both forms of interface consistency. It is not enough to make a product look the same; in fact, designers trying to inappropriately force consistency may inadvertently impair performance the system. Usability research may even overly emphasize perceptual consistency since, compared to conceptual consistency, it is easier to recognize (e.g., putting controls in the same location). A desire to save on manufacturing may also encourage perceptual consistency by reusing similar interface elements between disparate systems. Results from the present study demonstrate that this could be harmful.

In cases when conceptual consistency is implausible (i.e., two functionally different systems), perhaps designers should include perceptual inconsistencies to cue users to the conceptual inconsistencies. For example, in the case of a VCR and DVD system, including additional perceptual differences might help to cue users that the two systems operate differently (e.g., you do not rewind a DVD). This approach might help users avoid inappropriate generalizations.

Effects of interface consistency change over time

Another interesting finding was that the effects of conceptual and perceptual consistency varied differently as an effect of exposure time in phase 2. Data across multiple measured variables (e.g., tasks completed or navigational stability) suggest that conceptual inconsistencies initially harmed performance but that participants were able to adapt and perform at equivalent levels by the end of the session (with the exception of number of course corrections). Perceptual inconsistencies, in comparison, harmed performance for the secondary task (the captain's orders task) throughout all of phase 2.

These results demonstrate the importance of perceptual consistency. It appears that seemingly small changes like changing the location or orientation of an interface element (e.g., a button or a lever) can continue to hurt performance. This effect likely would be even worse if an operator switched back and forth between two perceptually inconsistent systems (e.g., Office 2003 and Office 2007). Interestingly, these findings suggest that users can overcome conceptual inconsistencies as they learn how the new system works. When first using the system, however, a user's performance would be hindered by the conceptual inconsistencies. Further, even if this effect is short-lived it may still harm initial impressions of a new system.

Limitations and Future Research

It is important to discuss some limitations of the current study. One limitation was the limited duration of the study. Participants spent a relatively short amount of time learning the initial system (about 40 minutes for practice and phase 1). Participants also spent a short amount of time using the second system (30 minutes). I expect that if participants had spent even longer with the initial systems then the effects of a consistent interface would be even more pronounced. For example, an expert Microsoft user would likely be more affected when switching from the well-learned Office 2003 to novel Office 2007. In contrast, a novice would still suffer some from the inconsistencies, but the novice would likely not be as affected. Rhee et al. (2006) attempted to study the effects of the different components of interface consistency; however, the results of that study were so limited overall that they were inadequate to address the issue of experience. Future work needs to examine how experience might moderate or exacerbate the effects of a consistent interface.

Additionally, the effects of a consistent interface on individuals with differing cognitive resources are still unexplored. An interface that is consistent with a previously learned device might be especially beneficial for individuals with limited resources (e.g., older adults) that are first learning to use a new device. In that case, the consistency would encourage them to leverage prior knowledge therefore reducing the cognitive load associated with learning a new skill. It would also be interesting to see the relative effects of the two types of consistency for both high and low ability individuals. Perhaps a perceptually inconsistent system would be even more detrimental for individuals with relatively lower perceptual speed (e.g., age-related declines in perceptual abilities). Similarly, incomplete consistency (e.g., a system that operates differently but looks the same) could be even worse since it seems to encourage inappropriate generalizations. Individuals with lower working memory tend to be even more susceptible to drawing inferences too quickly (e.g., Morrow, Leirer, Carver, & Tanke, 1998). Future research should determine how individual differences in cognitive abilities may influence the findings from the present study.

Results from the present study supported the notion that a well-implemented interface consistency (i.e., not incomplete consistency) is beneficial. As predicted, both conceptual consistency and perceptual consistency generally improved performance (although in different ways). Also, as predicted, some combinations of the two dimensions, referred to as incomplete consistency, can be detrimental to both performance and user perceptions of workload. This study should serve as a guide for future research on interface consistency by illustrating the interrelated nature of the two dimensions of interface consistency. Ultimately, knowing how to properly implement interface consistency and when it matters most will help to make systems safer, more efficient, and easier to use.

APPENDICES

APPENDIX A: Study Protocol

Protocol for Starship Simulator

Required materials for each participant:

1. This protocol
2. Copies of Informed Consent
3. Two copies of System Knowledge Questionnaire per participant

Arrive > 15 minutes before scheduled participants then:

1. Prop open lab door
2. Hang the participant running sign
3. Turn on computers and monitors
    a. Open program
    b. Type in participant number and make note of it (see number guide below)
    c. Place paper forms at workstations (i.e., consent forms and questionnaire)
4. Determine participant numbers based on condition
    a. Appearance 1
        1. Energy-Energy
        2. Rate-Rate
        3. Energy-Rate
        4. Rate-Energy
    b. Appearance 2
        1. Energy-Energy
        2. Rate-Rate
        3. Energy-Rate
        4. Rate-Energy
5. Greet participant when they arrive and verify name

Once participants have arrived:

- Hello. Thank you for agreeing to participate in this study today. You can expect this to take about two hours to complete. Before we continue, please make sure that your cell phone is set to silent.
- The purpose of this study is to examine how the design of a system's interface affects your ability to use that system. If you have any questions during the study please let me know and I will be glad to answer them.
- First, I'll need you to complete this "Informed Consent" form. This form will explain the study and inform you of your rights as a participant. Once you have read it, please sign it along with the duplicate copy; one copy is for you and one is for me.

[Hand them consent form and wait for participant to finish reading/signing consent forms]

System Introduction:

- Your task today is to pilot a pretend spaceship. Imagine that it is your job to operate the spaceship by keeping all systems running smoothly and following the commands of the ship's captain.
- First you'll work on some practice tasks that will instruct you about the basics of the starship system. Please raise your hand if you have any problems during this practice portion and I'll come help.
- Now, please enter your age and select male or female then press the begin button.

[Stand behind participants while they complete practice tasks. Assist if participants get stuck. Wait until all participants finish the practice tasks.]

- Great, everyone is finished with the practice part. As a quick review, the three things that cause the ship to go into danger mode are as follows:
- 1) Life support power is too low. 2) The ship's course is unstable in the red area. 3) You over-fuel causing the system's total power to go above 200.
- Your primary concern is to keep the ship operating safely by managing the power allocation in the ship and providing course corrections.
- When possible you should also complete as many of the Captain's tasks as you can in the allotted time. Each task you complete raises your final score. The faster you complete tasks, the higher score you'll be able to achieve. Try to avoid the danger warnings because danger warnings cause your score to decrease.
- Do you have any questions about the system?

[Wait for questions]

- Okay, for the next part you'll work to complete as many tasks as you can using the starship during a 30 minute session. The timer bar at the bottom will count down the time for you.
- Once the 30 minutes are up, some questionnaires will pop up on the computer. Please follow the instructions on the computer screen to complete these.
- One of these requires headphones so please put those on now.

  [Wait for participants to put on headphones]

- Does anyone have any questions?

[Wait for questions]

- Okay, I'm going to start each of you on the starship. You can begin as soon as I launch it for you.

[Monitor participants as needed but do not hover over them too much. Once you see dropbox updating you'll know that the simulator portion is finished. Make sure the participants move on to the NASA-TLX.]

Operating the second spaceship:

- Next, I'd like you to operate the starship one more time. Imagine that you are about to get into another spaceship.
- You will again operate this spaceship for another 30 minutes. At the end of that time you will complete the questionnaire about the difficulty of this second starship.
- Do you have any questions?

[Answer questions]

- Okay, you can begin as soon as I open the starship program for you.

[Wait ten minutes for participants to complete the task. Make sure they then finish TLX.]

Abilities Tests:

- Next I'd like you to complete a series of tests on the computer designed to assess your mental abilities.

[Launch abilities program]

Video Game Experience and System Knowledge Questionnaire:

- Next, please complete this questionnaire about your experience with video games and your knowledge about the starship system and how it works.

[Once finished]

- That concludes this experiment. Thank you very much for coming today. If you have any questions, please let me know. If not, you can expect your experiment credit to show up in the next 24 hours.

Concluding tasks:

- Collect paperwork and file it (consent form and system knowledge questionnaire)
- Record participant numbers in spreadsheet
- Assign participant credit
- Determine next conditions to run

APPENDIX B: NASA-TLX Questionnaire

Please place an “X” along each scale at the point that best indicates your experience with the display configuration. Note: participants completed an electronic version.

Mental Demand: How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc)? Was the mission easy or demanding, simple or complex, exacting or forgiving?

Low High

Physical Demand: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the mission easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

Low High

Temporal Demand: How much time pressure did you feel due to the rate or pace at which the mission occurred? Was the pace slow and leisurely or rapid and frantic?

Low High

Performance: How successful do you think you were in accomplishing the goals of the mission? How satisfied were you with your performance in accomplishing these goals?

Low High

Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?

Low High

Frustration: How discouraged, stressed, irritated, and annoyed versus gratified, relaxed, content, and complacent did you feel during your mission?

Low High

For each of the pairs listed below, circle the scale title that represents the more important contributor to workload in the display.

Mental Demand	or	Physical Demand
Mental Demand	or	Temporal Demand
Mental Demand	or	Performance
Mental Demand	or	Effort
Mental Demand	or	Frustration
Physical Demand	or	Temporal Demand
Physical Demand	or	Performance
Physical Demand	or	Effort
Physical Demand	or	Frustration
Temporal Demand	or	Performance
Temporal Demand	or	Frustration
Temporal Demand	or	Effort
Performance	or	Frustration
Performance	or	Effort
Frustration	or	Effort

APPENDIX C: Demographics Questionnaire

Note: administered electronically

Date of Birth: _____/_____/_____ ☐1 Male ☐2 Female
(month/day/year)

---

1. How many years of education did you complete?
   - ☐1 No formal education
   - ☐2 Less than high school graduate
   - ☐3 High school graduate/GED
   - ☐4 Vocational training
   - ☐5 Some college/Associate’s degree
   - ☐6 College graduate
   - ☐7 Master's degree (or other post-graduate training)
   - ☐8 Doctoral degree (PhD, MD, EdD, DDS, JD, etc)

2. Current marital status (check one)
   - ☐1 Single
   - ☐2 Married
   - ☐3 Separated
   - ☐4 Divorced
   - ☐5 Widowed
   - ☐6 Other (please specify __________________)

3. Race/ethnicity
   - ☐1 America Indian/Alaskan Native
   - ☐2 Asian
   - ☐3 Native Hawaiian or Other Pacific Islander
   - ☐4 Black/African American
   - ☐5 White
   - ☐6 Hispanic/Latino
   - ☐7 Multiracial (please specify _____________)
   - ☐8 Other (please specify _________________)

4. In which type of housing do you live?
   - ☐1 Residence hall/College dormitory
   - ☐2 House/Apartment/Condominium
   - ☐3 Senior housing (independent)
   - ☐4 Assisted living
   - ☐5 Nursing home
   - ☐6 Relative's home

☐7 Other (please specify ________________)

5. Is English your primary language?
☐1 Yes
☐2 No

5 a. If "No", What is your primary language?
☐1 English
☐2 Spanish
☐3 French
☐4 Creole
☐5 Portuguese
☐6 Other _________

Occupational Status

6. What is your primary occupational status?
☐1 Work full-time for pay
☐2 Work part-time for pay
☐3 Student
☐4 Homemaker
☐5 Retired
☐6 Volunteer worker
☐7 Seeking employment, laid off, etc.
☐8 Other (please specify): _________________________________

APPENDIX D: Video Game Experience Questionnaire

Note: administered electronically

Participant Number: ______

How often have you played the following types of video games?

	Never	Seldom	Sometimes	Frequently	Often
PC Games					
Console Games (e.g., Playstation, Wii, Xbox, etc.)					
Cell phone games (e.g., iPhone or Android games)					
Online java games (e.g., popcap or yahoo games)					
Video games at an arcade					

How frequently do you play the following types of games?

	Never	Seldom	Sometimes	Frequently	Often
First person shooters (e.g., Halo, Gears of War, Half-Life)					
Strategy games (e.g., Starcraft, Age of Empires, Civilization, Sim City)					
Role playing games (e.g., World of Warcraft, Final Fantasy, Diablo)					
Casual games (e.g., online java games, card games, Popcap games, Tetris, Minesweeper)					
Simulator games (e.g., Flight games or racing games)					
Sports games (e.g., Madden, NBA Live, NCAA)					

Please list any video game systems you own (e.g., Xbox360, Playstation 3, Wii, PC, etc.)

Do you consider yourself to be an active video game player?

Yes
No

How good do you feel you are at playing video games?

No skill
Not very skilled
Moderately skilled
Very skilled

During an average week, how many hours do you spend playing video games?

Less than 1 hour
1-3 hours
3-5 hours
5-7 hours
7-9 hours
More than 9 hours

How often do you play video games?

Never
Seldom
Sometimes
Frequently
Often

If you play video games, at what age did you first begin playing?

Before age 5
Age 5-7
Age 8-10
Age 11-13
Age 14-16
After age 17
Never, I don’t play video games

Do you own a personal computer?

Yes
No, but I use a public computer (e.g., on-campus or at a library)
No, I don’t regularly use computers

APPENDIX E: System Knowledge Questionnaire

Note: administered electronically.

Which system is linked and shares power with Life Support?

Navigation
Phasers
Shields
None

Which of the following shield configurations drains the most power?

Frequency 172 using pattern A.
Frequency 392 using pattern ABC.
Frequency 2013 using pattern CD.
Frequency 27 using pattern AD.

On which edge of the screen is the shield frequency number pad for the first system you used?

Top
Right
Bottom
Left

On the list of tabs, which tab was the furthest right for the first starship system?

Navigation
Phasers
Shields
Maintenance

Which of the following is the correct sequence to perform a navigation task for the second system you used?

Press the red activate button to start the engines, then select a speed, and finally set a heading.
Set a heading, then select a speed, and finally press the red activate button to start the engines.
Increase power to the navigation system, then adjust course, and finally press the red activate button to start the engines.
Select a speed, then set a heading, and finally press the red activate button to start the engines.

Which of the following does NOT cause danger mode?

Life support power is at 3 units.
Overall power is at 225 units.
Ship's oxygen system needs repair.
Navigational course is in the red area.

To fire the phasers, which two systems must have power?

Phasers and Shields
Life Support and Shields
Phasers and Navigation
Phasers and Maintenance
Only Phasers must have power

Which system tab is the second from the left for the second system you used?

Power
Navigation
Phasers
Shields
Maintenace

If you wanted to activate the front and rear shields for the first system you used, which two segments would you select?

Right and Left
Top and Right
Top and Bottom
Bottom and Right
Bottom and Left

If the power for the phasers, navigation, and shields are each at 30 and the total system power is 130, how much power does the life support system have?

10
20
40
60
90

APPENDIX F: Additional Analyses

Performance during the End of Phase 1 and the Beginning of Phase 2

The next analyses were designed to assess how conceptual and perceptual consistency manipulations might have affected performance immediately following the manipulations. To measure this, I compared performance for the six minutes immediately before the manipulations to the six minutes immediately after the manipulations. I used a 2 (conceptual consistency; high/low) × 2 (perceptual consistency; high/low) × 2 (time segment; end of phase 1/beginning of phase 2) mixed factorial ANOVA to compare performance between these two time segments (conceptual and perceptual manipulations were between-group).

Tasks completed

Results for the number of tasks completed indicated a significant interaction of conceptual consistency × time segment $F(1,73) = 16.2$, $p < .001$, $\eta_p^2 = .18$ and perceptual consistency × time segment $F(1,73) = 12.9$, $p = .001$, $\eta_p^2 = .15$. I conducted a follow-up analysis to identify the source of these interactions.

The source of the conceptual consistency × time segment interaction was that participants in the conceptual inconsistency group completed significantly fewer tasks during the first six minutes of phase 2 than during the last six minutes of phase 1 $F(1,36) = 10.9$, $p = .002$, $\eta_p^2 = .23$. In contrast, participants in the conceptual consistency group did the opposite, completing more tasks during the first six minutes of phase 2 than during the last six minutes of phase 1 $F(1,37) = 5.4$, $p = .026$, $\eta_p^2 = .13$. The source of the perceptual consistency × time segment interaction was that participants in the perceptual inconsistency group completed significantly fewer tasks during the first six minutes of phase 2 than during the last six minutes of phase 1 $F(1,36) = 17.3$, $p < .001$, $\eta_p^2 = .32$. Participants in the perceptual consistency group completed an equivalent number of tasks during the two time segments ($p = .247$). The main effect for time segment was not significant ($p = .073$).

Number of course corrections

Results for the number of course corrections indicated that neither the interaction of conceptual consistency × time segment was significant ($p = .413$) nor was the perceptual consistency × time segment interaction ($p = .516$). There was a significant main effect of time segment, with participants making more course corrections at the end of phase 1 than during the beginning of phase 2 $F(1,73) = 107.5$, $p < .001$, $\eta_p^2 = .60$.

Ship's power variability

Data on the average amount of power variability indicated that there were no differences between the two time segments as a result of the two consistency manipulations. Neither the interaction of conceptual consistency × time segment was significant ($p = .264$) nor was the perceptual consistency × time segment interaction ($p = .214$). There was a significant main effect of time segment on overall power variability $F(1,73) = 6.3$, $p = .014$, $\eta_p^2 = .08$. The ship's overall power was more variable at the beginning of phase 2 than during the end of phase 1.

Time in danger mode

Data for the amount of time in danger mode indicated that there were no differences between the two time segments as a result of the interface consistency manipulations. The interaction of conceptual consistency × time segment was not significant (*p* = .090). The perceptual consistency × time segment interaction was also non-significant (*p* = .226). Finally, the main effect of time segment was also non-significant (*p* = .326).

Table 11. Descriptive Statistics for Last Six Minutes of Phase 1.

	Conceptually Inconsistent				Conceptually Consistent			
	Perceptually Inconsistent		Perceptually Consistent		Perceptually Inconsistent		Perceptually Consistent	
Measure	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Tasks Completed	25.1	9.2	28.2	8.9	23.7	7.6	24.4	5.2
Course Corrections	128.4	39.1	112.1	50.3	135.7	57.8	136.7	54.8
Overall Power Variability	18.4	10.8	20.0	9.5	19.8	11.9	17.6	7.3
Time in Danger Mode (seconds)	55.2	48.5	53.9	52.9	62.7	42.6	51.6	44.2

REFERENCES

AlTaboli, A. & Abou-Zeid, M. R. (2007). Effect of Physical Consistency of Web Interface Design on Users’ Performance and Satisfaction. In J. A. Jacko (Ed.), *Human-Computer Interaction: HCI Applications and Services* (pp. 849-858). Berlin: Springer.

Bayer, N.L. (1992). Consistency in Interface Design: An Overview. *IPCC 92 Santa Fe. Crossing Frontiers. Conference Record*, September 29-October 3, (pp. 412-417).

Brown C.M. (1999). Human-Computer Interaction Design Guidelines. Great Britain: Intellect Books.

Ekstrom, R B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for Kit of Factor-Referenced Cognitive Tests*. Letter Sets. Princeton, NJ: Educational Testing Service. (pgs. VZ-2, 174, 176, 179).

Finstad, K. (2008). Analogical Problem Solving in Casual and Experienced Users: When Interface Consistency Leads to Inappropriate Transfer. *Human-Computer Interaction, 23*, 381-405.

Fisk, A.D., Derrick, W.L., & Schneider, W. (1986). A Methodological Assessment and Evaluation of Dual-Task Paradigms. *Current Psychological Research & Reviews, 5,* 315-327.

Gopher, D. (1993). The skill of attentional control: acquisition and execution of attention strategies. In D.E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV* (pp. 299-322). Cambridge, MA: MIT Press.

Grudin, J. (1989). The Case against user interface consistency. *Communications of the ACM, 32,* 1164-1173.

Kellogg, W. A. (1987). Conceptual consistency in the user interface: Effects on user performance. In H. J. Bulinger & B. Shackel (Eds.), *Human-computer interaction; INTERACT ’87IFIP Conference on Human-Computer Interaction* (pp. 389–394). Amsterdam, The Netherlands: Elsevier.

Maclin, E. L., Mathewson, K.E., Low, K.A., Boot, W.R., Kramer, A,F., Fabiani, M. & Gratton, G. (2011). Learning to multitask: Effects of video game practice on electrophysiological indices of attention and resource allocation. *Psychophysiology, 48,* 1173-1183.

McLaughlin, A.C., Rogers, W.A., & Fisk, A.D. (2009). Using direct and indirect input devices: Attention demands and age-related differences. *ACM Transactions on Computer-Human Interaction (TOCHI), 16,* 1-15.

Mendel, J., Pak, R., & Drum, J.E. (2011). Designing for Consistency: Can Interface Consistency Reduce Workload in Dual-task Situations? *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting*. Las Vegas, NV: Human Factors and Ergonomics Society.

Morrow, D.G., Leirer, V.O., Carver, L.M., & Tanke, E.D. (1998). Older and younger adult memory for health appointment information: Implications for automated telephone messaging design. *Journal of Experimental Psychology: Applied. 4,* 352-374.

Nielsen, J. (1989). Executive Summary: Coordinating User Interfaces for Consistency. In J. Nielsen (Ed.), *Coordinating User Interfaces for Consistency* (pp. 1-7). San Francisco, CA: Morgan Kaufmann Publishers.

Norman, D.A. (1988). *The Design of Everyday Things*. New York, Doubleday.

Ozok, A.A. & Salvendy, G. (2000). Measuring consistency of web page design and its effects on performance and satisfaction. *Ergonomics, 43,* 443-460.

Ozok, A.A., & Salvendy, G. (2001). How consistent is your Web design? *Behavior and Information Technology, 20,* 433–447.

Ozok, A.A., & Salvendy, G. (2003). The effect of language inconsistency on performance and satisfaction on the World Wide Web: results from four experiments. *Behaviour and Information Technology*, *22,* 155–163.

Ozok, A.A. & Salvendy, G. (2004). Measuring consistency of web page design and its effects on performance and satisfaction. *Computers in Human Behavior, 20,* 149-161.

Polson, P. (1988). The Consequences of Consistent and Inconsistent User Interfaces. *Cognitive Science and Its Applications for Human Computer Interaction*, 59-108.

Polson, P.G. & Lewis, C. (1990). Theory-based design for easily learned interfaces. *Human-Computer Interaction, 5,* 191-220.

Proctor, R.W. & Vu, K.L. (2006). Location and Arrangement of Displays and Control Actuators. In W. Karwowski (Ed.), *Handbook of Standards and Guidelines in Ergonomics and Human Factors* (pp. 399-409). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Rhee, C., Moon, J., & Choe, Y. (2006). Web interface consistency in e-learning. *Online Information Review, 30,* 53-69.

Rieman, J., Lewis, C., Young, R.M., & Polson, P.G. (1994). "Why Is a Raven Like a Writing Desk?" Lessons in Interface Consistency and Analogical Reasoning from Two Cognitive Architectures. *Proceedings of CHI'94 Conference on Human Factors in Computer Systems* (pp. 438-444). New York, NY: Association for Computing Machinery.

Satzinger, J., & Olfman, L. (1998). User Interface Consistency Across End-User Applications: The Effects on Mental Models. *Journal of Management Information Systems, 14,* 167-193.

Schneider, W., & Shiffrin, R.M. (1977). Controlled and automatic human information processing: 1. Detection, search, attention. *Psychological Review, 84,* 1-66.

Shneiderman, B. (1987). *Designing the user interface-Strategies for effective human-computer interaction.* Reading, MA: Addison-Wesley.

Tanaka, T., Eberts, R.E., & Salvendy, G. (1991). Consistency of human–computer interface design: Quantification and validation. *Human Factors, 33,* 653–676.

Thorndike, E.L. & Woodworth, R.S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review, 8,* 247-261.

Wechsler, D. (1981). *The Digit Symbol Substitution Test. Wechsler Adult Intelligence Scale-Revised*. New York: Psychological Corporation/Harcourt Brace Jovanovich.

Wechsler, D. (1997). *Wechsler Memory Scale III, (3rd Ed.)*. San Antonio, TX: The Psychological Corporation.

Wickens, C.D. & Hollands, J.G. (2000). *Engineering Psychology and Human Performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.