HUMAN-COMPUTER INTERACTION, 2002, Volume 17, pp. 63–139

Comparison of Four Primary Methods for Coordinating the Interruption of People in Human–Computer Interaction

Daniel C. McFarlane
Lockheed Martin Advanced Technology Laboratories

ABSTRACT

Interruptions can cause people to make mistakes or errors during human–computer interaction (HCI). Interruptions occur as an unavoidable side-effect of some important kinds of human computer-based activities, for example, (a) constantly monitor for unscheduled changes in information environments, (b) supervise background autonomous services, and (c) intermittently collaborate and communicate with other people. Fortunately, people have powerful innate cognitive abilities that they can potentially leverage to manage multiple concurrent activities if they have specific kinds of control and interaction support. There is great opportunity, therefore, for user-interface design to increase people's ability to successfully handle interruptions, and prevent expensive errors. The literature contains very little concrete design wisdom about how to solve the interruption problems in user interfaces (UIs). Coordination support, however, is identified as a most important design topic. This article presents the results of an empirical investigation to compare basic design solutions for coordinating human interruption in computer-based multitasks. A theory-based taxonomy of human interruption is used

Daniel McFarlane is a computer scientist with an interest in intelligent command and control systems; he is a Senior Member of the Engineering Staff in the Advanced Technology Laboratories of Lockheed Martin.

CONTENTS

to identify the four primary methods for coordinating human interruption. An experiment with 36 participants compares these four different design solutions within an abstracted common user multitasking context. The results show important design tradeoffs for coordinating the interruption of people in HCI and support some UI design guidelines. Negotiation support is the best overall solution except where small differences in the timeliness of handling interruptions is critical and then immediate is best.

1. INTRODUCTION

Information technologies continue to improve and increase in functionality, and to expand people's ability to simultaneously (a) monitor dynamic information environments, (b) supervise autonomous services, and (c) conduct intermittent ongoing communication and collaboration with other people. These useful technologies provide concurrent multitasking support, including mixed-initiative interaction; support for delegation and supervisory control of automation, for example, intelligent agents; and many kinds of distributed, backgrounded services and technologies that increase intermittent human-human communication.

These advances are driven by powerful consumer markets of people who want–need tools to increase their ability to control their many simultaneous activities. People have a natural ability and predisposition to multitask (Cherry, 1953; Cypher, 1986; Woods, 1995). Systems that support mixed-initiative user multitasking, however, have an interaction requirement for computers to interrupt their users. Trends in technological progress, therefore, make human interruption a central HCI design problem for the future. This is a problematic user interface (UI) design requirement because although people have natural abilities to handle interruptions, they can only leverage these strengths if given specific control and interaction support. (See McFarlane &

Latorella, 2002[1] for an in-depth review of literature related to the scope and importance of human interruption in human–computer interaction [HCI].)

Without appropriate support, interruptions can cause people to make tragic errors. For example, a Northwest Airline crew preparing to fly out of Detroit was interrupted during their preflight checklist by an air traffic control (ATC) operator with new taxiing instructions and a warning about possible windshear. After the crew finished talking to ATC they failed to resume their checklist. They took off without properly setting the aircraft's flaps, and a flight emergency occurred shortly after takeoff because the flaps were in the wrong position. The crew mistakenly interpreted the problem as windshear and crashed (NTSB, 1988). A professional and highly motivated crew crashed a perfectly functioning aircraft because they did not recover from the ATC interruption and resume the preflight checklist.

The literature shows that interrupting people does not always cause them to make errors (Lee, 1992, p. 81). The *taxonomy of human interruption* (McFarlane, 1997, 1998; McFarlane & Latorella, 2002) identifies coordination support for users as a critical design problem with potential to empower people in interrupt-laden multitask environments. Appropriate coordination support can enable users to handle interruptions effectively without it causing them to make errors on other tasks.

Unfortunately, there exists little concrete design wisdom and few UI design guidelines about how to solve this problem. And it is a complex design problem as evidenced by several existing examples of computer systems with ineffective ad hoc solutions to the human interruption problem (McFarlane & Latorella, 2002).

1.1. Goals and Overview

This article reports the findings of an empirical comparison of four primary methods of coordination support to enable users to handle human interruption in HCI. The goal is for the results to reveal critical UI design issues for this problem and to support the creation of UI design guidelines. Section 2 reviews basic research that describes the effects of interruptions on human performance in a variety of contexts and individual differences that may mediate these effects. It provides evidence for the importance and ubiquity of the interruption-management problem, and reviews the existing design guidance for incorporating interruptions in multitasking situations. Section 3 focuses on methods to coordinate interruptions and discusses the theoretical foundation

[1] McFarlane and Latorella (2002) contains a detailed review of the literature on human interruption in HCI. It appears in this issue of *HCI.*

for the experiment. It discusses the four basic coordination solutions identified in McFarlane's taxonomy of human interruption: immediate, negotiated, mediated, and scheduled. Section 4 describes the experimental methods including the experimental multitask. Section 5 reports the results. The data show important design tradeoffs for coordinating the interruption of people in HCI. Section 6 proposes UI guidelines for solving the human interruption problem in HCI. These guidelines are based on the statistically significant findings from the experiment. Section 7 discusses the observations of the experimenter that could not be statistically tested with these data and speculates on potentially useful research topics for future work.

2. BACKGROUND

Researchers have observed that interrupting people affects their behavior. This is the basis of the classic Zeigarnik Effect in psychology (Van Bergen, 1968). First identified in 1927, the Zeigarnik Effect describes a finding that people have selective memory relative to interruption, that is, that people are able to recall the details of interrupted tasks better than the details of uninterrupted tasks. Results from many studies of this effect have produced somewhat inconsistent results. However, two findings seem universal: (a) interrupting people affects their behavior and (b) the interruption of people is a complicated process.

Researchers have since documented other effects of interruption. Cohen (1980) found that unpredictable and uncontrollable interruptions induce personal stress that can negatively affect performance after interruptions. Interruptions can cause an initial decrease in how quickly people can perform post-interruption tasks (Gillie & Broadbent, 1989; Kreifeldt & McCarthy, 1981). They also can cause people to make mistakes, reduce their efficiency, or both (Cellier & Eyrolle, 1992; Gillie & Broadbent, 1989; Kreifeldt & McCarthy, 1981; Latorella, 1996a, 1996b, 1998).

People also have individual differences in their ability to accommodate interruptions while they multitask (Braune & Wickens, 1986; Joslyn & Hunt, 1998; Morrin, Law, & Pellegrino, 1994), in their ability to recall information about interrupted tasks (Husain, 1987), in their performance on interrupted tasks (Cabon, Coblentz, & Mollard, 1990; Weiner, 1965), and in how they handle interruptions in human–human communication (e.g., Lustig, 1980; West, 1982).

People, however, have some natural abilities to dynamically adapt their behaviors to accommodate interruptions. The normally deleterious effects of interruptions can be mitigated when an operational environment allows flexibility in task performance, a variety of methods for responding to interruptions, specific training, or both (Chapanis, 1978; Hess & Detweiler, 1994;

Jessup & Connolly, 1993; Karis, 1991; Lee, 1992; Zijlstra & Roe, 1999). Speier, Valacich, and Vessey (1997) found work contexts where the introduction of interruptions actually increases human performance. They also found, however, that this phenomenon does not hold for complex or cognitively demanding tasks; in these contexts interruptions decreased performance.

Some research has identified the aspects of multitasking situations that influence the effects of interruption on peoples' performance. Czerwinski, Chrisman, and Rudisill (1991) found an inverse relation between task similarity—between the primary and the interruption tasks—and people's ability to remember information about the interrupted task after interruption. Gillie and Broadbent (1989) found weak evidence that the similarity between the interruption and current tasks and the complexity of the interruption task directly affected the disruptiveness of interruptions. They also found that allowing users to review their foregrounded activity prior to handling interruption did not necessarily help them recover that activity after interruption. They asserted that the negative effect of interruption on memory was caused by memory interference created by interruption tasks that were complex or similar to the pre-interruption task. Speier et al. (1997) found a negative relationship between interruption frequency and human performance on complex tasks. Storch (1992) found that interruptions expressed as on-screen messages were more disruptive to people performing a computer data-entry task than interruptions expressed as telephone calls or as human visitors.

McFarlane and Latorella (2002) provided in-depth examples of real-world systems that have the UI design problem of human interruption. They described the negative effects of human interruption in commercial aircraft cockpit or flight deck systems and in an important Naval warship command and control system. They also identified other real-world examples contained in the literature, including intelligent tutoring systems (Galdes & Smith, 1990), computer-mediated communication (McCarthy & Monk, 1994), telephone communications (Katz, 1995), U.S. Navy's Multi-Modal Watchstation (Obermayer & Nugent, 2000; Osga, 2000), office environments (Rouncefield, Hughes, Rodden, & Viller, 1994; Speier et al., 1997; Zijlstra & Roe 1999), and Internet instant messaging (Czerwinski, Cutrell, & Horvitz, 2000a, 2000b).

There are few sources of design wisdom on how to best support human interruptions in HCI. Some UI design research has included user interruption to increase the realism of the experimental domain. These studies have not directly studied the interruption problem, but the results contain relevant findings of the potential utility of some UI solutions for supporting interruption. Kreifeldt and McCarthy (1981) used interruption to investigate the "stress tolerance" of alternative UI designs for calculators. Field (1987) used interruptions to compare two different database navigation tools. Williams (1995) used interruptions to evaluate the utility of a digital data link for complementing voice communications between aircraft and ground stations. McDonald and Stevenson (1996) used interruptions to compare three alternative text structures on hypertext navigation performance.

Smith and Mosier (1986) proposed guidelines for UI support that combine good interruption presentation to minimize disruption, with more user control over handling interruptions. Burton and Brown (1979) said that the UI design problem of when to interrupt is critical to the success of any intelligent computer-assisted instruction (ICAI) system. Their UI design guidelines focus on minimizing the frequency of interruptions and using context to determine when to interrupt users. Also for ICAI, Galdes and Smith (1990) analyzed the teaching behaviors of expert human tutors and identified successful interruption strategies employed by people. Galdes and Smith then presented these strategies as UI design guidelines to build an ICAI tutorial system. These guidelines, like those of Burton and Brown, said that timing of the interrupt must be context sensitive.

Cooper and Franks (1993) identified human interruption as a complex cognitive process that can be used as a formative example to design cognitive models. They identified useful dimensions of interruption in their framework: source, effects, content, applicability, duration, mechanism for recovery, and state space of the underlying system (pp. 76–78). Obermayer and Nugent (2000) presented a list of UI-design guidelines to create alerting and attention management systems in Navy command and control systems. Their guidelines focus on minimizing interruption frequency, matching the degree of attention-getting cues to the degree of importance of the alert, and providing users ultimate control over when to handle interruptions.

McFarlane and Latorella (2002) propose two theoretical tools for solving UI design for human interruption. The first is Latorella's Interruption Management Stage Model (IMSM; Latorella, 1996b, 1998; McFarlane & Latorella, 2002). This is an HCI process model that shows the stages of cognitive-information processing that people exhibit and the kinds of management strategies that they use to handle interruptions for a class of identified problems. The model structures a discussion of human information processing to extract task, operator, and environment factors that will likely determine the degree to which an interruption will have deleterious effects. This information highlights where different kinds of performance problems can happen, and it is useful for researchers who need to work in or support people working where the stated interruption pattern is appropriate. It also provides insights into the HCI process that can guide the design of human-interruption management support.

The second is McFarlane's Definition and Taxonomy of Human Interruption (McFarlane, 1997, 1998; McFarlane & Latorella, 2002). These are an attempt to map the total design space and identify a broad array of potential influences of user performance with tie-ins to relevant design literature for addressing these factors. The taxonomy shows areas of the problem, where specific technologies could be introduced to give people richer support for handling interruption. McFarlane's taxonomy identifies eight major dimensions of the problem of human interruption: (a) source of interruption, (b) individual characteristic of person receiving interruption, (c) method of coordination, (d) meaning of interruption, (e) method of expression, (f) channel of conveyance, (g) human activity changed by interruption, and (h) effect of interruption.

There is a useful common ground on support for interruption coordination in both Latorella's IMSM and McFarlane's Definition and Taxonomy of Human Interruption. Both works provide treatments of when to interrupt the user and what kind of user control should be supported in the UI design. This question is a paramount design topic for supporting human interruption. *Method of coordination*, the third factor of the taxonomy, identifies four solutions or methods for UI coordination support for interruption: (a) immediate, (b) negotiated, (c) mediated, and (d) scheduled.

3. APPROACH

The purpose of our research is to determine the relative strengths and weaknesses of the four different methods of coordination UI solutions for human interruption identified earlier (immediate, negotiated, mediated, scheduled).[2] An experiment was performed to compare these solutions within a common context.

Coordination support determines when interruptions are presented to the user and what kind of control the user is given to deal with them. Example: A person concurrently performs two tasks: (a) supervising the navigation of an unmanned air vehicle (UAV)[3] and (b) using eight intelligent agents to monitor changes in the larger dynamic information environment for events that could impact the UAV's mission. These agents are tasked to search for information relative to the UAV's mission and send alerts to interrupt the human operator whenever something is discovered. An immediate solution would have the agents interrupt the person immediately regardless of what they are doing in a way that insists that the operator immediately pause the UAV-navigation task and interact with the agent. A negotiated solution would have the agents announce their need to interrupt and then support a negotiation with the per-

[2] See McFarlane and Latorella (2002) for detailed review of the existing UI design literature relative to coordinating human interruption.

[3] UAV's are used by military and other groups for cost-effective aerial reconnaissance.

son. This would give the user control over when or whether to deal with the interruption. A mediated solution would have the agents indirectly interrupt and request interaction through a personal broker like a personal digital assistant (PDA). The PDA would then determine when and how the agents would be allowed to interrupt. A scheduled solution would restrict the agents' interruptions to a prearranged schedule, such as a 15-min cycle.

There is not enough design knowledge in the literature to say which method of coordination would be best for specific work contexts, and different UI designers have very different intuitive answers. Prior studies have looked at topics related to each of the four methods of coordination.

Many of the detrimental effects of interruptions are related to people's difficulty resuming the original task after handling the interruption. One cost that has been identified for the "immediate" solution is that people experience a troublesome initial decrease in performance called automation deficit when they try to resume interrupted tasks (Ballas, Heitmeyer, & Pérez, 1992a, 1992b). Some attempts to help users more easily resume interrupted tasks have been investigated. Providing brief warnings of impending interruption has utility (Czerwinski, Chrisman, & Rudisill, 1991; Czerwinski, Chrisman, & Schumacher, 1991). Awareness of "backgrounded" tasks can be heightened with sonification (Gaver, 1989). Reminders can prepare people to resume interrupted tasks (Davies, Findlay, & Lambert, 1989). Tools can be devised to help people quickly review and resume interrupted tasks (Field, 1987).

The "negotiated" solution is an attempt to exploit people's natural ability to negotiate for changes in their activities. Clark (1996) said that people normally negotiate human–human interruptions. Unlike the immediate interruption method of coordinating interruption, people usually have choices about whether to allow interruptions and how and when to handle them (Clark, 1996). Clark said that in normal human–human language usage people have four possible responses to interruption: (a) accept with full compliance, (b) accept with alteration, (c) decline, or (d) withdraw. Woods (1995) said that people have a natural ability to manage their own attention. While people concentrate on a single task with focused attention, they also subconsciously gather information about peripheral activities and use this information to effectively guide their focus of attention between competing tasks. Raby and Wickens (1991) said that people naturally manage their own attention while performing multitasks. Some papers have investigated the usefulness of presenting interruptions in ways that allow people to ignore them if they choose (Lieberman, 1997; Oberg & Notkin, 1992). Katz (1995) found that there are overhead costs related to negotiating interruptions, and that users sometimes prefer immediate interruption solutions when that overhead cost is not justified. Rich (1996) investigated the utility of using a moving hand-shaped icon as an attention-getting technique for interaction with an intelligent agent.

The "mediated" solution is an attractive but controversial approach. Delegating the interruption problem to a mediator begets a new task of supervising the mediator (Kirlik, 1993). There are five main approaches for mediation: (a) predict people's interruptibility (Czerwinski, Cutrell, & Horvitz, 2000a, 2000b; Miyata & Norman, 1986), (b) implement intelligent UIs for supervision tasks (Chignell & Hancock, 1988; Lieberman, 1997), (c) automatically calculate users' cognitive workload for dynamic task allocation (Berger, Kamoun, & Millot, 1988), (d) apply human factors techniques for supervisory control (Sheridan, 1987), and (e) use cognitive models to guide interaction (Hammer & Small, 1995). Interruption by proxy can be another mediation solution. If the user has a proxy that can act in their behalf, then the proxy can receive interruptions and handle them for the user.

The "scheduled" solution is an attempt to give a degree of reliable expectation to a user about when they will be interrupted. If people had foreknowledge of the *when, what, where, why,* and *how* of incoming interruptions, they could plan their other activities to minimize the negative effects of interruptions. Clark (1996) said that people are very familiar with two useful kinds of scheduling techniques for normal human–human activities: explicit agreement and convention. Explicit agreement is a technique that people use to prearrange the coordination of a one-time event, like a meeting. Convention is a technique that people use to prearrange the coordination of a recurring event, like a regularly scheduled weekly group meeting. In many ways, scheduling times for unexpected activities transforms interruptions into normal planned activities. Time management training has been found to have a positive effect on people's abilities to manage interruptions (Hall & Hursch, 1982). "Constant interruption" is another form of scheduled work solution. If a person knows that they will receive a constant, unending, stream of interruptions, then none of the interruptions interfere with other work in unexpected ways (Rouncefield et al., 1994).

4. EXPERIMENT

Four different methods for coordinating interruption described in the previous section (immediate, negotiated, mediated, and scheduled) are empirically compared. An experiment was conducted to test a hypothesized causal relationship between these alternative UI design solutions and people's behavior during interruption.

A conservative first question is "Does it matter which coordination method is chosen as a solution to this UI design problem?" If the answer is, "Yes, it does matter," then the relative strengths and weaknesses of the different solutions can be compared. Findings may lead to the creation of UI design guidelines.

4.1. Hypothesis

Hypothesis (H_a): Four separate UI design solutions for coordinating the interruption of people in HCI—immediate, negotiated, mediated, and scheduled—will differentially affect users' performance on interrupt-laden, computer-based multitasks.

4.2. Participants[4]

Thirty-six volunteers successfully participated as participants in this experiment (18 men and 18 women). Participants had a median age of 21 ($M = 24.7$, minimum 18, maximum 47). All but two participants were recruited from e-mail broadcasts to students in The School of Engineering and Applied Science at George Washington University (28) and employees of the Navy Center for Applied Research in Artificial Intelligence (6). The recruitment message did not reveal the purpose of the experiment, but portrayed the experimental task as "fun" and "similar to a video game," and offered a "$20 reward." Self-selection of participants was constrained to equal numbers of male and female volunteers. This method for population sampling is less than random and therefore not optimal. However, it was judged adequate because of the exceptional diversity of the GWU student population, and because of the motivation of the monetary reward.

4.3. Design

A single-factor, within-subjects, Latin square design was chosen as an appropriate design for this experiment. Six treatments were devised—four experimental treatments and two base-case control treatments. Each of the experimental treatments represented one of the four methods for coordinating interruption.

Each treatment condition used a different version of a UI (the independent variable). The computer-based multitask was not varied between treatment conditions. Participants' performance (the dependent variable) on the multitask was observed and recorded under the six treatment conditions.

All participants received all six treatments. However, each participant was assigned to one of six groups that defined the counterbalanced ordering (digram-balanced) of the presentation of the six treatments. The presentation of each treatment was divided into two contiguous trials to avoid the confounding influences of fatigue and boredom.

[4] This article uses the term "participants" instead of subjects as per the APA style guidelines.

Each participant performed a total of 24 trials of the computer-based multitask. Each trial was 4.5 min long, followed by a brief rest period with a masked screen. Rest periods were a minimum of 25 sec each; therefore the total time for a participant to complete the experimental task was about 2 hr. For all participants, the first 12 trials were practice (~ 1 hr) and the following 12 trials were experimental (~ 1 hr). The length of the practice period and the size of trials were determined based on the results from pilot testing.

Participants received the same counterbalanced ordering of trials on practice trials as they did on experimental trials. For example, participants assigned to the Latin squares order Group 2 received their 24, 4.5-min trials of treatments (1–6) in the following order (Figure 1, from left to right).

A digram-balanced Latin squares ordering was chosen as the counterbalanced grouping scheme because it ensures that each condition preceded and followed all other conditions exactly once (Keppel, 1991, p. 339; Wagenaar, 1969). This ordering was used to try to control for possible differential carryover effects (Figure 2). Male and female participants were randomly assigned to groups with three men and three women in each group.

4.4. Multitask

An abstract interruption-laden, computer-based multitask was created as a testbed for this experiment. It is targeted at maximizing external validity of the results for an important class of real world multitasks similar to the US Navy's Aegis Identification Supervisor (IDS) task. The experimental dualtask is a simplified model of a class of common multitasks that require people to perform a continuous nonpausable, computer-based task while they simultaneously process arbitrary external interruptions. These continuous tasks are composed of multiple discrete subtasks that overlap in time and require some degree of concurrent attention from the human operator. The kind of multitask modeled here is composed of different, possibly unrelated, concurrent tasks.

Examples of jobs that require people to routinely perform these kinds of multitasks include 911 emergency dispatch operators, air traffic controllers, military radar operators, UAV operators, commercial and military aviators, and nuclear power plant control room engineers. The Navy's Aegis System, for example, requires an operator to concurrently maintain the accuracy of information about tracks of several objects[5] as they appear and change over time. Each of the individual tracks is a subtask that requires actions, and the

[5] A track is a computer representation of an object in the environment, like an airplane, or ship. It is usually supported with radar and/or other sensor data.

Figure 1. **Example of trial presentation sequence.**

Practice Trials	Experimental Trials
2 2 4 4 1 1 6 6 3 3 5 5	2 2 4 4 1 1 6 6 3 3 5 5

Figure 2. **Counterbalanced treatment order by participants' group.**

	Treatment Condition Order					
	First	**Second**	**Third**	**Fourth**	**Fifth**	**Sixth**
Group 1	1	2	3	4	5	6
Group 2	2	4	1	6	3	5
Group 3	3	1	5	2	6	4
Group 4	4	6	2	5	1	3
Group 5	5	3	6	1	4	2
Group 6	6	5	4	3	2	1

actions for the track subtasks must be done concurrently in an unpredictable, possibly overlapping order. These overlapping subtasks can only be acted on one at a time, but the operator must maintain simultaneous awareness of all subtasks. The operator must also be available for arbitrary interruptions by their leaders for direct requests of information. McFarlane and Latorella (2002) contains a detailed review of the human interruption issues in the Aegis IDS task.

The multitask used for this experiment is a *dualtask* (a two-task multitask) composed of a continuous game task and an intermittent matching task. The game task is modeled after a video game by Nintendo™ Corporation called "Fire" that was originally released in 1980 and 1981 as a version of the Nintendo Game & Watch product series (Nintendo, 1980–1981, 1997). The matching task is modeled after the matching tasks used in experiments of the Stroop Effect (Jensen & Rohwer, 1966; Stroop, 1935). The dualtask is conceptually simple and yet can be very difficult for people to perform. The results of pilot studies confirmed that this dualtask elicits the kind of human errors associated with the interruption phenomenon.

The tasks comprising the dualtask are dissimilar. This experimental design was taken to avoid any potentially confounding task interaction effects. We postulate that the issue of degree of task similarity in the experimental dualtask is not critical to the investigation of best overall UI design solution for coordinating interruption. The effects of variance in degree of task similarity across interruptions is beyond the scope of this study.

The many subtle low-level cognitive mechanisms involved in human interruption were not directly investigated here. The potentially interesting small-scale effects of these mechanisms were ignored and isolated from the high-level effects by imposing randomness into several aspects of the experimental design. These randomization sources equalize the effects of these small-scale mechanisms across the different treatment conditions and allow these small-scale influences to be ignored in the data analysis.

Game Task

The object of the game task required participants to direct stretcher-bearers to catch other game characters as they jumped from a building. Each falling character had to be successfully bounced three separate times at three different locations. If a character was missed at any of the three bounce points, then it was lost. The game task is trivial when game characters jump one at a time. However, when multiple characters jump in quick succession it becomes a difficult game of juggling (Figure 3). The game scenario involved Marine stretcher-bearers saving diplomats jumping from an overrun U.S. embassy.

All subtasks (individual jumping characters) required the same three bounces. For each subtask, the time it took from its start jump until its third (and last) bounce was 13.7 sec. After a character had been successfully bounced its third time, it was on the screen another 3.2 sec until it fell safely into the military truck (total time on the screen for a saved diplomat was 13.7 + 3.2 = 16.9 sec). Note that this experimental task has some similarities with the marching boxes tasks used by Tulga and Sheridan (1980) and subsequently by Moray, Dessouky, and Kijowski (1991).

Pilot studies revealed that this timing for subtasks did not require constant attention from participants, only a few well-timed actions. This ensured that participants could potentially switch between tasks but that they would have to maintain a significant amount of situational awareness to multitask successfully.

The level of difficulty of the game had to be contrived so that it was complex enough to attack participants' vulnerability to interruption, but simple enough not to cause participants to despair of performing well. Through testing with pilot participants, it was discovered that 59 game subtasks per trial were appropriate.

The results of the pilot studies also revealed the need to have two different levels of complexity for the practice trials. It was found through pilot testing that an introductory period of easy play was necessary to give participants time to learn everything they needed to know. The number of jumpers for the 24 trials for each participant is shown in Figure 4 (except for the "match only" base case treatment condition, which had no game task). Note, the cells of the table in Figure 4 contain the pairs of trials for each of the six treatment conditions.

Figure 3. **Game task.**

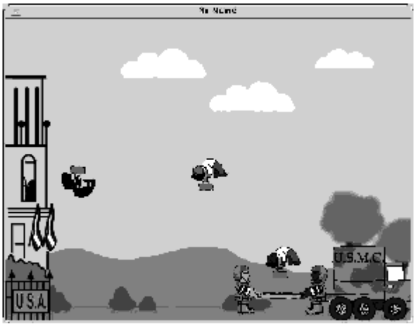


Figure 4. **Number of jumpers for each trial of game task.**

	First	Second	Third	Fourth	Fifth	Sixth
Practice trials	38 59	38 59	38 59	38 59	38 59	38 59
Experimental trials	59 59	59 59	59 59	59 59	59 59	59 59

Each subtask (a “jumper”) was completely independent. This had some useful consequences that facilitated experimentation: (a) participant’s performance on completed subtasks could be easily classified as success or failure, (b) errors made while performing one subtask did not automatically cause errors on other subtasks, (c) each experimental trial could have a unique randomly scheduling of subtasks to prevent predictability across trials, and (d) the overall complexity of the game task to be conveniently manipulated by specifying the number of jumpers to occur within the fixed 4.5-min trials.

The six treatments or different versions of the UI were designed to be simplistic representations of the four different coordination solutions plus two control conditions. These were made intentionally minimalistic in an attempt to expose the natural differences between the four approaches, and allow for a clear comparison. The following kinds of feedback were intentionally not implemented: sound, performance scores, animation of secondary events, alerts of impending events, and information of the state of the hidden task. These additional feedback stimuli could have been powerful sources of confounding influence on participants’ performance.

Matching Task

The second task of this dualtask was used as the interruption task. This task was an intermittent graphical matching task loosely based on the textual matching tasks reported in investigations of the Stroop effect (Jensen & Rohwer, 1966; Stroop, 1935). The interruption task required participants to make matching decisions either based on color or shape. Participants were presented with a colored shape upper center of the window, and instructed to choose one of the bottom two colored shapes according to the matching rule displayed in the center. The matching rule instructed participants to either "match by shape" or "match by color" (Figure 5).

This task was chosen because each individual matching task requires a definite but minimal focus of attention that cannot be automated through "overlearning." Pilot studies confirmed that, although this matching task is conceptually simple, pilot participants were not able to automate this task through overlearning even after 2.5 hr. It was discovered that 80 matching tasks per trial were appropriate.

The graphic nature of the matching task was chosen to correspond with the graphic nature of the game task. Matching subtasks had to be done one at a time from a first-in-first-out queue, so there were no interruptions of interruptions. Individual matching tasks were independent and participants' choices were easily judged right or wrong. The left–right choice was conveniently mapped to a left–right keyboard selection.

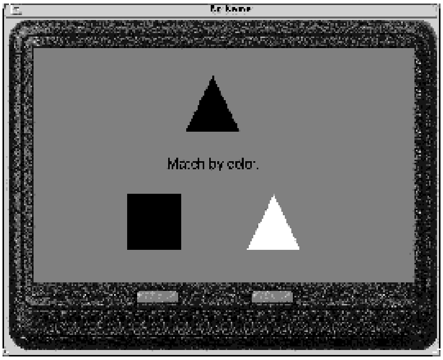
As with the game task, pilot studies revealed the need to have a simplified introduction version of the matching task for the first trial of each pair of practice trials. The number of matching tasks for the 24 trials for each participant is shown in Figure 6 (except for the "game only" base case treatment condition, which had no matching task).

Each of the 864 trials in this experiment (36 Participants × 24 Trials each) was provided with a unique and unpredictable schedule for the multitask. A constrained randomization scheme was used to schedule the onset of events. Two standard sets of wait times were calculated (one for scheduling the 59 game task subtasks, and one for scheduling the 80 individual matching tasks) so that the intervals between wait times increase linearly from 0 to a maximum number such that all the intervals summed to ~4.5 min. Before each trial, these sets were randomly resorted. This scheme, therefore, did not affect the frequency domain of the intervals between subtasks.

4.5. Treatments

Participants performed the multitask with one-handed keyboard key presses on an isolated group of six keys of a common extended computer keyboard. Participants performed the game task by pressing the Delete and Page Down keys with one hand to control the back and forth movement of the stretcher-bearers. Participants performed the matching task by using the same hand to press the Insert and Page Up keys to choose either the left or right

Figure 5. **Matching task.**

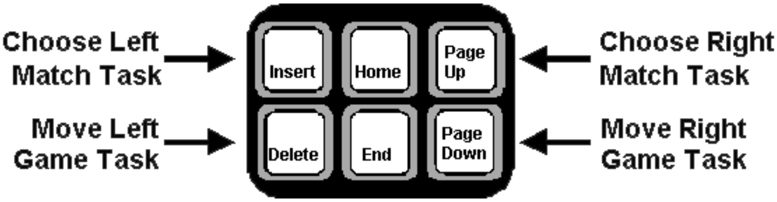


Match by color.

Figure 6. **Number of matches per trial of matching task.**

	First	Second	Third	Fourth	Fifth	Sixth
Practice trials	40 80	40 80	40 80	40 80	40 80	40 80
Experimental Trials	80 80	80 80	80 80	80 80	80 80	80 80

Figure 7. **Keys used for performing the experiment.**



shapes (Figure 7). The Home and End keys were only used in the "negotiated" treatment condition.

Whenever a matching task was in the foreground, it appeared in the same window as the game task and totally obscured the view of the game. The game task continued to run without possibility for pause regardless of whether participants could see it or not. In all treatments, except negotiated, once the multitask was switched to the matching task participants had to perform all queued matching tasks before they could resume the game task in progress. Whenever a user completed the last queued matching task, the multitask switched back to the game task in progress. The following are descriptions of the six treatment conditions. The subtask scheduling scheme was the same for all six treatment conditions.

- Treatment 1: “Game only” base case implemented the game task in isolation with no matching tasks.
- Treatment 2: “Match only” base case implemented the matching task in isolation with no game task.
- Treatment 3: “Immediate” treatment condition presented matching tasks directly whenever they occurred regardless of the state of the game task.
- Treatment 4: “Negotiated” treatment condition gave participants control over when they would handle interruptions. When a matching task occurred, its arrival was immediately announced with a flash of a blank matching task for 150 msec and then the game task display resumed. Participants then had to decide when to begin the queued matching task. Participants could use the Home and End keys at any time to show the queued matching tasks in the foreground or to hide them in the background. If more than one matching task was queued, participants did not have to perform all of the tasks together, but instead could switch back and forth between the game task and the queued matching tasks at will.
- Treatment 5: “Mediated” treatment condition dynamically calculated a simple function of participants’ workload that measured how many jumping diplomats were currently visible on the screen.[6] In general, interruptions were automatically held until workload metric was low. When workload metric was high and interruptions were being held, no notification of the arrival of interruptions was presented. In practice, an algorithm was used to ensure that the interruption queue did not exceed six, even when the workload was high.[7]
- Treatment 6: “Scheduled” treatment condition held all interruptions without notifying participants, and only switched from the game task to the matching task on a prearranged schedule of once every 25 sec.

4.6. Apparatus

All participants performed the computer-based dualtask on a PC laptop (an HP OmniBook 5700CTX, 166MHz Pentium; Windows 95). The built-in monitor was used as the display. It was a 12.1-in. backlit liquid-crystal XGA

[6] The workload level definitions: 0–1 current jumpers → “easiest” workload; 2 current jumpers → “easy” workload; 3 current jumpers → “hard” workload; and 4 or more current jumpers → “hardest” workload.

[7] The interruption algorithm definition: If workload is “easiest,” then interrupt now; else if workload is “easy” and the number of queued interruptions is more than 2, then interrupt now; else if workload is “hard” and the number of queued interruptions is more than 4, then interrupt now; else if workload is “hardest” and the number of queued interruptions is more than 6, then interrupt now.

display with 1024 × 768 pixel resolution and 16-bit color. The computer-based dualtask was displayed in a single 640 × 480 pixel window in the top left corner of the screen. The experimental software was implemented with sprite-based double-buffered frame animation running at 20 frames a second. A multi-threading approach was implemented to improve the reliability of timing data.

The laptop sat on a box 4.75 in. high on a tabletop in front of participants to create a comfortable viewing angle. Participants used an external extended keyboard that sat on the tabletop directly in front of them. Participants were seated on a padded chair typical of the kind used by office workers.

Potential environmental distracters were minimized by conducting the experiment in isolation. The wall behind the computer apparatus was blank. The video camera and the experimenter were located on the other side of the room well behind participants to avoid encroaching on participants' sense of personal space. The experimenter also sat facing somewhat away from participants and did not appear to be directly observing them.

4.7. Procedure

Participants participated one at a time. The experimenter followed a written script to ensure that the treatments were administered to each participant consistently. Each participant was required to sign a consent form and pass a standard color vision test (Ishihara, 1996). Participants then performed a brief entrance questionnaire (Appendix A1). Each participant was asked to read a booklet of written instructions that contained pictures of the game and matching task, and described all the treatment conditions. Treatments were identified with Greek letters so as not to imply any ranking. Participants kept these instructions for reference throughout the experiment. Each participant received 24 trials of the computer-based dualtask over the period of about 2 hr. Each trial was preceded with an on-screen message announcing which treatment condition would be next. A second on-screen message reminded participants to pace their efforts so that they would not become tired, and that the game and matching tasks were equally important. After each trial, a gray rectangle was displayed that covered the experimental task window as a mask. Detailed interaction data were unobtrusively recorded by computer throughout the experiment, and all trials were also video taped for redundancy.

After the experimental tasks were finished, participants were asked to perform an exit questionnaire (Appendix A2). Participants were then given a formal debriefing and rewarded with $20 (the six civilian employees of the Naval Research Lab could not accept the $20 reward but participated as part of their normal employment). Participants spent a minimum of about 2 hr 30 min participating in this experiment.

5. RESULTS

The main hypothesis for this experiment predicted that the four UI solutions for coordinating interruptions—immediate, negotiated, mediated, and scheduled—would cause different levels of user performance on interrupt-laden computer-based multitasks. Participants' performance is the dependent variable and is operationalized with nine objective measures, and 19 subjective measures. There are important individual difference effects. The general results show that the negotiation-based solution causes the best overall user performance. However, the immediacy-based solution causes slightly better performance on the timeliness of handling interruption tasks and would be better for cases where small differences in the timeliness of the interruption tasks are critical.

The main objective of this study was to attempt to discover valid UI guidelines. This section will report the objective data, and then determine whether there is an overall interruption effect relative to the base cases. The hypothesis is then tested and a summary of the other experimental effects is presented (Appendix B contains the details of these results for reference). The following section (Section 6) presents guidelines that are based on the empirically validated findings.

The nine objective measures do not comprise an exhaustive set. However, they cover four important kinds of user performance that are identified in the literature as being affected by interruptions: correctness (see metrics 1, 4, and 5), efficiency (see metrics 2, 3, and 9), completeness (see metric 6), and timeliness (see metrics 7 and 8). It is asserted that these nine metrics are sufficient for testing this hypothesis.

The data are reported with measures of central tendency and then analyzed in increasing levels of detail to identify the differential effects on participant performance caused by the four alternative UI design solutions. The results of the subjective measures are presented in a similar way. All results are summarized into a set of tentative UI design guidelines and presented in Section 6.

The objective measures are

1. Number of jumpers saved on the game task ("jumpers saved").
2. Number of key presses per jumper saved on game task ("G. keyed per saved").
3. Number of switches between game task and matching task in both directions ("task switches").
4. Number of matches done wrong ("matched wrong").
5. Percent of matches done wrong of those attempted ("% M. wrong of done").
6. Number of matches not done ("matches not done").

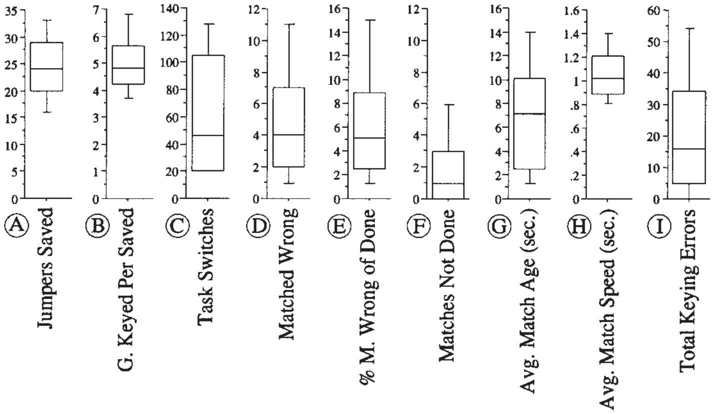
7. Average time in seconds from the scheduled onset of each matching task until it was actually completed or the trial timed out ("avg. match age").
8. Average time in seconds from display of each matching task until it was completed ("avg. matching speed").
9. Total number of keying errors on both tasks ("total keying errors"). These errors are UI manipulation errors, not task performance errors like "matched wrong." The total keying errors metric is a sum of five kinds of keying errors:
   i. Pressing a game task move left or right key when the game stretcher-bearers are already at their corresponding left or right limit ("redundant moves on game task").
   ii. Pressing the game-control keys when the dualtask is in the matching task mode ("game keys during matching task").
   iii. Pressing the matching task control keys when the dualtask is not in the matching task mode ("match keys when not matching").
   iv. Pressing the keys for switching the dualtask mode when not in the negotiated treatment condition ("illegal negotiation attempts").
   v. Pressing any key that was not part of the six key set for performing the dualtask ("unused keys").

Note that three performance measures, task switches, matches not done, and avg. match age, are not "traditional" experimental dependent variables because their value was not free to vary under participants' direct control (except in the negotiated condition). These performance measures are appropriate here, however, because these limitations on participants' performance are directly linked to the application of the different treatments and therefore illustrate how the four treatment conditions differentially affect participants' behavior.

Data from the participants' 12 experimental trials (not the practice trials) were included in these analyses. Figure 8 contains box plots for the nine chosen performance measures. The box plots display marks at the 10th, 25th, 50th, 75th, and 90th percentiles of the variable. The boxes, therefore, contain the center 50% of values with the centerline at the median. The outer brackets enclose 80% of values. The plots in Figure 8 show the general performance of participants on the experimental multitask; the data for the two single-task base cases were, therefore, not included. One of the performance measures quantifies positive performance—jumpers saved. Higher scores are better on this measure. The rest of the performance measures quantify negative performance. Lower scores are better on these measures.

Figure 9 contains bar charts for the nine performance measures split by method of coordination of interruptions. The means and error bars are in-

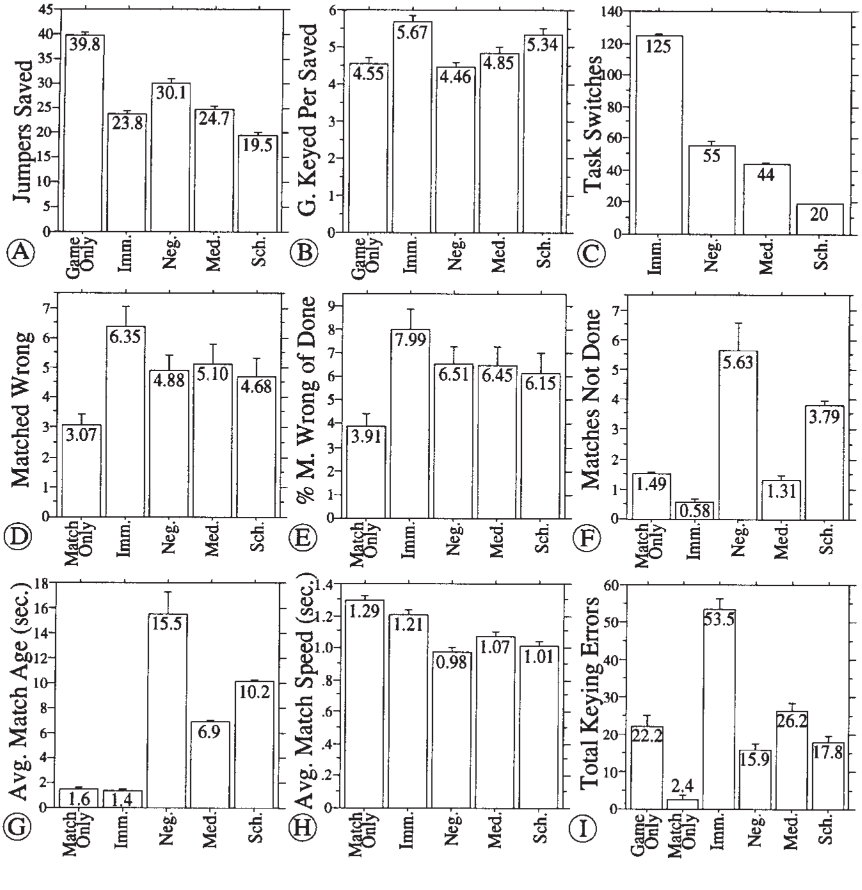
Figure 8. **Raw performance data from the experimental trials with conditions combined (not including base case data from the game or matching task only conditions). The box plots contain 50% of scores per trial A centerline at the median and outer brackets enclosing 80%. (A total of 36 participants; four experimental conditions per participant; two 270-sec trials per condition; and 59 jumpers on the game task with 80 interruptions on the matching task per trial.)**



cluded for each graphic. The error bars show one standard error. Note that, because these data are from a repeated measures experiment, caution must be practiced in using this graph to make estimations of significance between the different treatment conditions. The error bars reflect the total variance contained in the data, and not just the within-subjects variance that is relevant to testing the hypothesis. The relevant variance is therefore actually smaller than that shown. The error bars are graphically exaggerated by the inclusion of irrelevant between-subjects sources of variation, for example, differences in scale of scores from variation in participants' game playing abilities. Many differences between participants are actually controlled by the experiment through the application of repeated measures. The graphics are also slightly distorted because of the inclusion of outliers.

Because this experiment is the first to compare the four methods for coordinating user interruption, it was important to attempt to maximize the validity of the results. Nonparametric statistical tests were employed. These tests are more robust, but also more conservative than comparative parametric tests. Unlike parametric tests they do not depend on debatable assumptions about the data that can limit generalizability of findings: sampling independence, normality of distribution, and consistency of variance between conditions. Nonparametric tests, however, are less powerful than comparable parametric tests. They carry a higher risk of not finding true differences when they actually exist (type II errors, or β errors). The advantageous consequences of

Figure 9. **Raw performance data from the experimental trials, split by condition. Bar charts show the mean with error bars for one standard error. (A total of 36 participants; six experimental conditions; two 270-sec trials per condition [32.4 hr of total data]; and 59 jumpers on the game task with 80 interruptions on the matching task per trial.) Game Only = base case: game only with no interruptions; Match Only = base case: matching task only with no game; Imm. = immediate; Neg. = negotiated; Med. = mediated; Sch. = scheduled.**



accepting increased risk of type II error is a decrease in risk of type I error, or α errors (finding differences when there are none).

The decision to use nonparametric tests avoids potential confusions about the validity of parametric analyses. For example, it may be argued that the data do not have consistency of variance between conditions, because the different experiment conditions did not give participants equivalent kinds of control over all kinds of multitask performance.

5.1. Test of Hypothesis

When comparing alternative UI design solutions the bottom line is, “Does it matter, and if it does, which solution is the best?” The Friedman test was selected for testing the hypothesis, that is, the nonparametric Friedman two-way analysis of variance (ANOVA) by ranks test with correction for ties (denoted by F), with methods for post hoc comparisons (Siegel & Castellan, 1988). This test is appropriate for analysis of ordinal or interval data taken from three or more related samples. The Friedman test is calculated on within-subject ranks. Data for each participant are converted to ranks that denote on which condition they scored highest and on which condition they scored next highest, and so on. This analysis technique avoids the irrelevant influence caused by differences in participants’ overall game playing skill level, and therefore negates any possible biasing effect of outliers on the results. An alpha level of 0.05 is used to make decisions of significance.

The data from each pair of treatment trials were summed into single scores for this analysis. It is assumed that any effects related to differences between Trial 1 and Trial 2 for each treatment will not affect the results of a combined analysis. An evaluation of this assumption is presented in a later section.

This within-subject ranking scheme isolates the relevant within-subjects variance. It is effective because it removes much of the irrelevant variance, and is the reason that the Friedman test is immune to the concerns about the quality of variance needed for comparable parametric test. Figure 10 shows an example of the within-subject transformation of data for the observations from one participant.

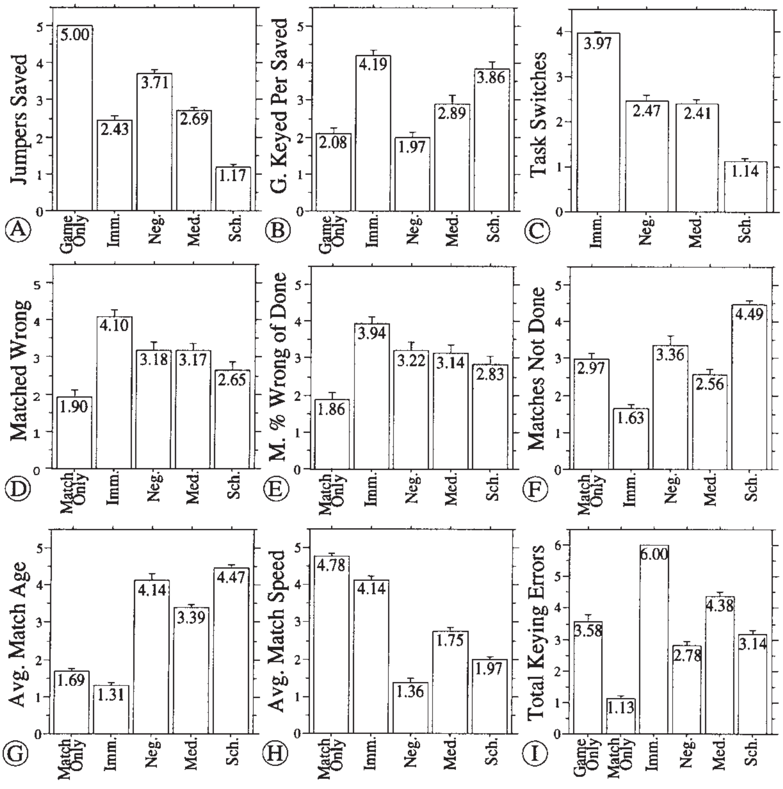
Within-subjects ranking of the data is not only useful for statistical analysis, but it is a valuable method for graphical communication of within-subjects effects observed in repeated measures experiments. This graphing technique helps make plain the answers to the question, “Do the differences between treatments matter and, if ‘yes,’ then on which experimental condition did participants do best and on which did they do worst?” This within-subjects ranking scheme clarifies the appearance of the graphed data somewhat, and facilitates visual judgments of significance between the individual conditions by exposing what’s important.

Figure 11 contains bar charts showing within-subjects ranks of the same data displayed by Figure 9. These are the ranks of the nine performance measures split by method of coordination of interruptions. The means and error bars (one standard error) are included for each graphic. Note that, when the distortive effects of irrelevant variance and outliers are removed, the relative ordering of bars can be different from that for the raw data. See examples of differences in relative orderings in Graphs D, F, and G between Figure 11 and Figure 9.

Figure 10. **Within-subject ranking transformation of experimental data (jumpers saved) from one participant.**

	Game Only		Match Only		Immediate		Negotiated		Mediated		Scheduled	
Trial	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**
Raw Data	39	41	—	—	21	25	34	31	22	24	17	22
Trials Combined	80		—		46		65		46		39	
Ranks	5		—		2.5		4		2.5		1	

Figure 11. **Within-subjects ranking of raw performance data from experimental trials (same raw data as Figure 9). Performance ranked separately for each participant with 1 =** ***lowest ranking*** **(values split for ties). Bar charts show the mean with error bars for one standard error. Game Only = base case: game only with no interruptions; Match Only = base case: matching task only with no game; Imm. = immediate; Neg. = negotiated; Med. = mediated; Sch. = scheduled.**



5.2. Overall Effects of Interruption

There must be an overall effect of interruption—otherwise a discussion of the differential effects of alternative methods for coordinating interruptions would not make sense. Two base case treatment conditions were included in this experiment so that this assertion could be validated before testing the main hypothesis. Figure 12 summarizes the results of the Friedman test to determine whether there is any significant difference between the relevant conditions for each measure of performance (the four treatment conditions and the relevant base case[s]). For comparison, F_r must be greater than 9.49 for its p value to test above the chosen α level of 0.05 ($F_r > 13.28$ for $\alpha = 0.01$; and $F_r > 18.46$ for $\alpha = 0.001$).

These results validate the basic assertion that being interrupted affects people's behavior. The significance of these results permits post hoc analyses. Figure 13 summarizes the results of a comparison of individual conditions with the appropriate base cases using the Friedman test's post hoc analysis methods (Siegel & Castellan, 1988, pp. 181–183). Each cell reports the results of significance tests with $\alpha = 0.05$. Figure 11 can be used to determine the direction of significant differences between pairs of experimental conditions.

5.3. Effects of Different Interruption Coordination Methods

Do the different methods of coordinating interruption affect people differently? The Friedman test is used to try this hypothesis. Figure 14 summarizes the results of the Friedman test to determine whether there is any significant difference between the four experimental conditions for each measure of performance (base cases are not included). For comparison, F_r must be greater than 7.82 for its p value to test above the chosen α level of 0.05 ($F_r > 11.34$ for $\alpha = 0.01$; and $F_r > 16.27$ for $\alpha = 0.001$).

The data from all nine performance measures support the main hypothesis with statistical significance and permit H_o to be rejected. It is concluded that separate implementations of the four primary design solutions for coordinating user interruption will result in UIs that differentially affect users' performance on interrupt-laden, computer-based multitasks. These differences indicate the potential importance of UI design for human interruption in HCI.

These significant results permit post hoc analyses (Siegel & Castellan, 1988, pp. 180–181). Figure 15 summarizes the results of comparisons between the four experimental conditions using the Friedman test's post hoc analysis methods. Each cell reports the results of significance tests with $\alpha = 0.05$. Figure 11 can be used to determine the direction of significant pairs.

Figure 12. **Comparison to base cases.**

Performance Measure	Base Case	F_r	p	$p <$ •
Jumpers saved	Game only	120.410	<.0001	Yes
G. keyed per saved[a]	Game only	58.711	<.0001	Yes
Task switches	[no appropriate base case]			
Matched wrong	Match only	39.627	<.0001	Yes
% M. wrong of done[b]	Match only	32.911	<.0001	Yes
Matches not done	Match only	65.960	<.0001	Yes
Average match age	Match only	117.956	<.0001	Yes
Average match speed	Match only	118.978	<.0001	Yes
Total keying errors	Both	137.393	<.0001	Yes
	Game only	92.569	<.0001	Yes
	Match only	123.036	<.0001	Yes

[a]Number of key presses per jumper saved on game task. [b]Percent of matches done wrong of those attempted.

Figure 13. **Post hoc comparison to base cases.**

Performance Measure	Base Case	Base and Immediate	Base and Negotiated	Base and Mediated	Base and Scheduled
Jumpers saved	Game only	Yes	Yes	Yes	Yes
G. keyed per saved[a]	Game only	Yes	No	No	Yes
Task switches	[No appropriate base case]				
Matched wrong	Match only	Yes	Yes	Yes	No
% M. wrong of done[b]	Match only	Yes	Yes	Yes	Yes
Matches not done	Match only	[c]			
Average match age	Match only	No	Yes	Yes	Yes
Average match speed	Match only	No	Yes	Yes	Yes
Total keying errors	Game only	Yes	No	No	No
	Match only	Yes	Yes	Yes	Yes

[a]Number of key presses per jumper saved on game task. [b]Percent of matches done wrong of those attempted. [c]The findings for this row are not being reported because there is some indication that the base case for the matches not done metric has a slight problem. The data for the experimental cases are clean. However, one extra matches not done count may have been added in error to some of the base case data counts for subjects. This potential slight contamination of this one base case metric does not affect any of the other findings, but we are withholding this row as a conservative attempt to maintain the overall high degree of validity of all reported results.

Figure 14. **Analysis of experimental conditions.**

Performance Measure	F_r	p	$p < \alpha$
Jumpers saved	72.263	<.0001	Yes
G. keyed per saved[a]	43.000	<.0001	Yes
Task switches	87.000	<.0001	Yes
Matched wrong	17.599	.0005	Yes
% M. wrong of done[b]	10.267	.0164	Yes
Matches not done	53.034	<.0001	Yes
Average match age	78.100	<.0001	Yes
Average match speed	78.733	<.0001	Yes
Total keying errors	84.092	<.0001	Yes

[a]Number of key presses per jumper saved on game task. [b]Percent of matches done wrong of those attempted.

Figure 15. **Post hoc analysis of main effect.**

Performance Measure	Immediate and Negotiated	Immediate and Mediated	Immediate and Scheduled	Negotiated and Mediated	Negotiated and Scheduled	Mediated and Scheduled
Jumpers saved	Yes	No	Yes	Yes	Yes	Yes
G. keyed per saved[a]	Yes	Yes	No	No	Yes	No
Task switches	Yes	Yes	Yes	No	Yes	Yes
Matched wrong	Yes	Yes	Yes	No	No	No
% M. wrong of done[b]	No	No	Yes	No	No	No
Matches not done	Yes	No	Yes	No	No	Yes
Average match age	Yes	Yes	Yes	No	No	Yes
Average match speed	Yes	Yes	Yes	Yes	No	No
Total keying errors	Yes	Yes	Yes	Yes	No	Yes

[a]Number of key presses per jumper saved on game task. [b]Percent of matches done wrong of those attempted.

5.4. Summary of Other Effects

The results showed a large interruption effect for the experimental multitask, and a strong differential causal relationship between UI solutions for coordinating human interruption—immediate, negotiated, mediated, and scheduled—and user performance. This positive empirical evidence leads our creation of tentative UI design guidelines for human interruption in HCI (see Section 6). The results have sufficient internal and external validity to support limited guidelines, especially for tasks that match well with the class of multitasks abstracted for the experiment.

There are other important results of this experiment. These also have influenced the guidelines proposed in Section 6. These other results include observations of the existence, or nonexistence, of the following:

1. General practice effects.
2. Differential carry-over effects.
3. Individual differences.
4. Correlations between subjective values and objective performance values.
5. Correlations between subjective rankings and objective performance rankings.
6. UI manipulation errors.

These other observed effects are summarized here. The full reports of these analyses are contained in Appendix B. The validity of the guidelines depends on these findings and so the details must be included. We recognize, however, that some readers would prefer to reference them separately.

Experimental Effects of Repeated Measures (See Appendix B1)

The repeated measures design used in this experiment did not have a confounding influence on the results. The Wilcoxon Signed Ranks test with correction for ties was used to test for two kinds of general practice effects—trial sequence and treatment sequence. Seven of the nine objective performance measures showed no trial sequence effects. The two with significant effects, average match speed and total keying errors, were inspected and the differences were very small. It is postulated that learning improved average match speed a trivial amount from Trial 1 to Trial 2; and fatigue degraded total keying errors a trivial amount from Trial 1 to Trial 2. These are statistically significant findings, but the size of these effects is too small to have confounded the overall results. There was no effect of treatment sequence whatever.

The Kruskal–Wallis one-way ANOVA by ranks with correction for ties was used to test for differential carryover effects. Seven of the nine measures showed no effects. The two with significant effects, jumpers saved and task switches, showed an interesting elevation in performance level for order Groups 1 and 4. It is postulated that there is a primacy effect, and that the treatment condition that participants saw first influenced the strategies that they used throughout the experiment. If participants had explicit control of the game task on their first encounter with the multitask (the game-only base-case or the negotiation-based UI solution), they saved more jumpers overall and did more task switching where possible. These effects, however, resulted in generally elevated levels across all treatment conditions for participant Groups 1 and 4 and did not introduce confounding influence on comparisons between conditions.

Individual Differences (See Appendix B2)

A 17-question entrance questionnaire (see Appendix A1) was used to measure individual characteristics of participants. The questions included biographical queries (4 of the 17) about sex, age, dominant hand, and years of education. The questionnaire also asked 13 questions of participants' self-perception of their level of skill and experience relevant to performing computer-based multitasks, general capability for handling interruptions, and UI manipulation proficiency. These results show large individual diversity in the group of participants who participated in this experiment.

The Mann–Whitney U test with correction for ties was used to test for effects of sex and effects of dominant "handedness" on the nine objective measures of human performance. There were no significant effects whatever. The total lack of any effect of sex was surprising given the literature on the participant (West, 1982; Zimmerman & West, 1975).

A correlation analysis revealed no strong correlations between participants' answers to the entrance questionnaire and their performance on the experimental multitask. A *correlation matrix* was calculated to compare all 945 pairwise combinations of 15 entrance questionnaire topics (all questions except sex and handedness) and 63 kinds of performance on the experimental multitask (the results of the six treatment conditions [with trials combined] for all nine objective performance metrics, plus totals for each excluding base cases).

There are three possible explanations for the missing statistical relationships between self-reported abilities and observed performance on the experimental multitask: (a) none of the skill topics measured in the entrance questionnaire are relevant predictors of people's performance on the multitask used in this experiment; (b) the entrance questionnaire constructured for this experiment was not a good measure of participants' self-perceptions; or (c) participants are not able to accurately report their true levels of experience, skill, and vulnerability. We assert that the third explanation—participants are not able to report their own skills relative to interruptions—is the most probable.

Subjective Effects (See Appendix B3)

A 21-question exit questionnaire (see Appendix A2) was used to measure four kinds of subjective values: (a) participants' overall anxiety and motivation (questions 18, 19); (b) participants' opinions about specific methods for coordinating interruption (questions 30–36); (c) participants' relative rankings of the four different coordination methods on various dimensions (questions 20–29); and (d) perceived complexity of the multitask relative to the four experimental conditions (questions 37, 38). The medians say something about the overall effects, and can be used to make generalizations. However, many of the measures show large variances, and this reveals strong disagreements between participants about those subjective topics.

Question types 1 and 2 (questions 18, 19, 30–36) elicited a single-valued judgment from 0 (*least*) to 10 (*most*). These were subjective judgments of level of anxiety, motivation, and utility of specific UI coordination solutions. A correlation analysis between these nine measures and participants' actual performance revealed no substantial correlation whatever.

A correlation matrix was calculated that compared all 567 pairwise combinations of these nine subjective exit questionnaire topics, and the 63 kinds of performance on the experimental multitask (the nine objective performance metrics with six treatment conditions each [trials combined], plus totals for each excluding base cases). There were no correlation coefficients greater than .5, and only three greater than .4. Most pairs showed virtually no correlation.

Some relationships may have been reasonably expected, and their conspicuous absence is worth noting. If the exit questionnaire collected valid and reliable data, then there must be no relation between people's subjective judgments of their internal states and opinions of UI solutions and their actual performance levels.

The Friedman two-way ANOVA by ranks test with correction for ties was used to determine whether there were significant between-subjects agreement in rankings (question type 3; questions 20–29) of the four UI coordination solutions—immediate, negotiated, mediated, and scheduled. Five of the 10 rankings showed significant agreement across participants: preference; game errors made; feeling of interruption; predict interruptions; and complexity of game at start of interruption.

These findings can be summarized as follows:

1. Participants said that they preferred the negotiated solution over both the immediate solution and the scheduled solution (question 20).
2. Participants said that they made relatively fewer mistakes on the game task in the negotiated solution (question 23).
3. Participants reported feeling less interrupted in the negotiated condition than they did in either the immediate or scheduled conditions (question 25).
4. Participants said they could predict the onset of interruptions better on the negotiated and scheduled solutions than they could on the immediate or mediated solutions (question 27).
5. Participants said that task switches started at more convenient places (lower complexity of game task) in the negotiated solution than they did in either the immediate or scheduled conditions (question 28).

The results of the other five ranking questions did not show agreement among participants. This indicates a strong individual differences effect.

An analysis was performed to calculate the correlations between participants' subjective rankings of the four different coordination methods (questions 20–29) and their objective within-participants rankings of differing performance on the corresponding four experimental conditions. The Kendall rank-order correlation coefficient with correction for ties was used to calculate the degree of agreement between the subjective and objective ranks. It was calculated for all 90 combinations of the 10 subjective ranks (questions 20–29) and the 9 objective performance measures.

Statistical tests support six general assertions:

1. People prefer those UIs that allow them to be more effective, efficient, and precise on the continuous task and process interruptions quickly.
2. People are better at reporting relative ranks of different UI designs than they are at reporting absolute values for isolated opinions about individual UIs.
3. UI designs that cause people to feel highly interrupted hinder their effectiveness, efficiency, and precision on the continuous task and suppress their ability to process interruptions quickly.
4. UI designs that increase people's feeling of distractedness impede their effectiveness on the continuous task and impede their ability to process interruptions accurately and quickly.
5. UI designs that increase the predictability of interruptions enable people to process interruption tasks more quickly and make fewer total keying errors than interface designs that do not. However, increased predictability also resulted in poor performance in completeness and timeliness on the intermittent task.

6. UI designs that allow interruptions to be presented at the lull points of the continuous task enable people to be more effective and efficient on the continuous task and to process interruptions quickly.

UI Manipulation Errors (See Appendix B4)

UI manipulation errors are unintentional keying mistakes that participants made during moments of confusion about how to make the computer do what they want. The experimental platform was constructed so that these errors could not directly affect the outcome of task activity. Participants could, however, intentionally do the wrong thing, and these kinds of errors are discussed in other sections.

Participants made a total of 9942 UI manipulation errors, or keying errors, on the experimental trials. That is an average of about 23 keying errors on each of the 432 experimental trials (36 Participants × 6 Conditions × 2 Experimental trials each). These errors are useful indicators of participants' level of confusion and wasted effort as they performed the multitask.

The total keying errors metric is a sum of five kinds of keying errors: redundant moves on game task, game keys during matching task, match keys when not matching, illegal negotiation attempts, and unused keys. Most keying errors were as follows: redundant moves on game task (72.85%) and game keys during matching task (22.70%).

The Friedman test was used to determine whether UI coordination solution had a causal effect on the frequency of the two most common kinds of keying errors. Redundant moves on game task and game keys during matching task showed similar results. The immediate solution for coordinating interruptions caused people to make the most keying errors and the negotiated and scheduled solutions caused people to make the fewest keying errors. We assert that the negotiated and scheduled solutions allowed people to stay in control of the UI better than the immediate and mediated solutions.

6. GUIDELINES FOR UI DESIGN

To maximize the validity of the guidelines proposed here, all guidelines are supported by statistically significant findings from this experiment. They represent the most concise summary of the empirically validated results. The results of this experiment reveal that there is no one "best" method for coordinating interruptions for all kinds of human performance. There are instead, tradeoffs.

The results of the objective metrics used in this experiment support two basic generalizations relative to the experimental multitask. First, people perform very well when they can negotiate for the onset of interruptions; however, giving people this kind of control also means that they may not handle interruptions in a timely way. Second, when people are forced to handle interruptions immediately, they get the interruption tasks done promptly but make more mistakes and are less effective overall. These two finding support a general UI design guideline that the negotiation-based solution is the best overall design solution except for problems with which small differences in timeliness of beginning the interruption task are critical, and then the immediacy-based solution is best.

The results presented in this article have good internal validity because they were obtained through controlled experimentation. The results also have some external validity because the experimental multitask was carefully contrived to have many similarities with a class of common real world multitasks. However, the external validity of these results cannot be strongly argued for two reasons: (a) This investigation was not a field study, and (b) the experimental multitask does not represent *every* kind of interrupt-laden multitask. The multitask, for example, was composed of independent tasks. It is possible to argue that the results would have been different if these tasks had interdependencies.

Any general design guidelines must be based on results that have both internal and external validity. However, since general guidelines are needed but not available, this article proposes some *tentative* design guidelines based on the results of this experiment. The overall generalizability of these guidelines is debatable. However, they serve as a useful summary of the significant findings from this research and may be useful if implemented with caution.

6.1. Overall Best and Worst

Tentative guidelines are proposed for maximizing specific types of human performance. These guidelines suggest optimal design solution(s) for the *average.* All recommendations of "best" or "worst" solution are supported with statistically significant observations. In cases where no statistical significance was found between different solutions, they were each included in the figure. For example, an entry of *Negotiated/Mediated* in the "Best" column means that either negotiated or mediated would be equally good solutions for ensuring success of the design goal associated with that row in the figure.

These guidelines may be useful for designing the default behavior of interactive systems and should allow *average* users to perform "well." Figure 16, however, does not include findings of significant individual differences in coordinating the interruption of people. The next section provides guidelines that are relative to individual users.

Figure 16. **Overall best and worst: Tentative design guidelines.**

Design Goal	Best	Worst
Accuracy on continuous task	Negotiated	Scheduled
Efficiency on continuous task	Negotiated/Mediated	Immediate/Scheduled
Fewest task switches	Scheduled	Immediate
Accuracy on intermittent task	Not Immediate	Immediate
Completeness on intermittent task	Immediate/Mediated	Scheduled/Negotiated
Promptness on intermittent task	Immediate	Scheduled/Negotiated
Efficiency on intermittent task	Negotiated/Scheduled	Immediate
Keying accuracy	Negotiated/Scheduled	Immediate
User preference	Negotiated/Mediated	Immediate/Scheduled
User perception of their own accuracy on continuous task	Not Immediate	Not Negotiated
User perception of least interruptive	Negotiated/Mediated	Immediate/Scheduled
User perception of most predictable	Scheduled/Negotiated	Immediate/Mediated
User perception of complexity of continuous task when interrupted	Negotiated/Mediated	Immediate/Scheduled

6.2. Relative Best and Worst

Some design knowledge is only meaningful in relation to specific information about individual users. UIs that can adjust to the needs of each individual user can result in better human performance than alternative "one-size-fits-all" designs.

There are significant correlations between subjective experience and actual performance. These relationships are within individual participants, and reflect how their perceptions and performance levels are tied together. This information could be used to individually customize UI designs. For example, there were large differences in user preference between the four solutions. Figure 17 shows that *whichever* solution a user said they preferred most was also the same solution that helped them get their best performance in the specified performance categories. If a system requirement puts a priority on these kinds of performance, then allowing users to choose the coordination method they prefer should be the best solution.

Figure 17 identifies significant correlations between people's subjective experience with the four design alternatives and their actual performance. These relationships can be used as guidelines for making design solutions relative to individual users.

This experiment found that not all kinds of user-specific information would be useful for guiding UI designs (at least for the kinds of performance that were measured). This can be important information because it means

Figure 17. **Relative best and worst: Tentative design guidelines.**

Individual's Subjective Value	Individual's Performance Level
Best preferred	Best effectiveness and efficiency on the continuous task, best efficiency on intermittent task; and best overall keying accuracy.
Best ease of use	Best accuracy on intermittent task.
Worst interruptive	Worst effectiveness and efficiency on the continuous task, worst efficiency on intermittent task, and worst overall keying accuracy.
Worst distractive	Worst effectiveness on the continuous task, and worst accuracy and efficiency on intermittent task.
Best predictability of interruptions	Best efficiency on intermittent task, and best overall keying accuracy, however, also worst completeness and timeliness on intermittent task.
Best timing of onset of interruptions to occur when continuous task is not difficult	Best effectiveness and efficiency on the continuous task, and best efficiency on the intermittent task.

that some differences between users can be ignored during design. This "probably-OK-to-ignore" list includes the following:

- Sex.
- Education.
- Handedness.
- All individual self-judgments of interruption-relevant experience and ability examined in this experiment (see Figure 11).
- All single-scale self-measures of subjective experience examined in this experiment (see Figure B–6 in Appendix B2). (The only useful subjective metrics were some of those that asked participants to rank the alternative solutions.)
- Relative perceived stressfulness.

Individualized solutions are optimal but may not always be possible. If a critical system cannot be made individually adjustable, it may be necessary to restrict the set of human operators to only those people that "fit" with the given system design.

7. DISCUSSION AND FUTURE WORK

Some interesting insights can be made about human interruption based on extensive first-hand observations of participants' behaviors in this experiment. These observations are not used to make any statistical claims about human interruption but are used to shed light on potentially fruitful avenues of future research.

There were important differences observed in how participants performed tasks. Participants' behaviors can be used to infer individual strategies for performing the experimental multitask. These inferred strategies are insights into how people's individual perceptions of themselves and the computer-based multitask differentially affect their behavior. (Note that every effort was made in this experiment to ensure that all participants clearly understood a common set of specific instructions. The differences in strategies, therefore, reflect what different people thought was the most appropriate use of their distinct individual abilities for the multitask.)

7.1. Ability to Mentally Simulate the Game Task

Participants appeared to have different abilities for multitasking. It seemed that some participants were able to run a simulation of the game task in their minds while they performed the matching tasks. This allowed them to know exactly what they would see when the game task resumed (except, of course, for the new jumpers that started while the game was hidden). While performing the matching task, these participants knew where the falling jumpers were, even though they could not see them. When the game task resumed, these participants would successfully take up the game task in progress almost immediately. They remembered where they had left the stretcher-bearers (left, middle, or right) and they would move them quickly to catch a jumping character that had been falling while out of view.

Other participants seemed to not have, or be willing to use, this ability. When the game task was hidden, it was gone—"out of sight, out of mind." These participants apparently had no sense that specific jumping characters were falling to destruction when the game task was hidden. While performing matching tasks, they exhibited no signs of frustration over having missed particular jumping characters. When the game task returned, these participants were completely lost for a short time, while they tried to recover the state of the game task from scratch. They had to look to see where they had left the stretcher-bearers and study the pattern of jumping characters to make a plan about what to do. Some of these participants said in the exit questionnaire (question 37) that they believed there were fewer total jumping characters on trials of the scheduled condition than on the other conditions. This inaccurate perception probably results from these participants not being able to simulate the dynamically changing rate of jumpers continuing to jump while the game task was hidden.

This difference in participants' ability to mentally simulate the game task when it was hidden affected participants' strategy for task switching on the negotiated interruption condition. Participants who could mentally simulate the game learned to be able to predict how long it would take them to perform one matching task. In the negotiated condition their strategy for task switching could be described as "merge the tasks." They would attempt to predict a space in the game task where they could fit in a matching task, and they would preplan the moves they would need to recover the game after being away, and execute. These participants learned to predict when the game was in a situation where they could switch out and do a matching task, and then switch back into the game task without missing any critical movements of the stretcher-bearers. They would position the stretcher-bearers strategically so they would be in the correct position for the next necessary bounce. Then, when resuming the game after performing a matching task, they would already have decided where they needed to move the stretcher-bearers next.

Participants who could not mentally simulate the game used a different strategy on the negotiation condition. Their strategy could be described as "put out fires." These participants would attempt to save as many jumpers as possible when the game task was heavy. While they worked on the game task the queue for the waiting matching tasks would grow quite large. Then whenever the game task became relatively light these participants would switch to the matching task and perform all the queued matches. The game task was light when they last saw it, and they were completely unaware of the number of jumpers that could have started while they were away. They often did not even use the basic strategy of positioning the stretcher-bearers at the left position before switching to the matching task. This strategy would have ensured that all new jumpers would make at least the first bounce while the participant was away doing the matching task.

The "put out fires" strategy could also be called the "once seen must do" strategy. Participants who adopted this strategy seemed to have decided that once they started putting out a fire they were going to keep working on it until it was out; then they would move on to the next fire. These participants had a very hard time leaving the game task when there were any jumpers still visible. But as soon as there were zero or one visible jumpers they would perceive the fire as out, and they could then move on to the waiting matching tasks. It was also difficult for these participants to leave queued matching tasks. The Negotiated solution allowed participants to leave the matching task and return to the game. However, as soon as a participant would do one matching task on the queue the next one was immediately visible. These participants, therefore, felt responsibility for it once they had seen it and had to stay with the matching task until they completely emptied the queue.

One future research idea would be to measure participants' ability to mentally simulate a dynamically changing situation and correlate this with their performance on the experimental multitask. This might be a useful indicator of which design solutions might be most appropriate for individuals with different degrees of ability to mentally simulate dynamic tasks when interrupted.

7.2. Ability to Focus on Stated Goal

It was observed that participants had strong differences in their feelings of responsibility for performing tasks. It was carefully explained to participants that success on the game task meant saving jumpers, not just bouncing them. However, a few participants were not able to focus on maximizing the number saved. They instead insisted on adopting strategies that maximized extending the lives of all jumpers. This often meant that they were unwilling to sacrifice a new jumper and miss it at the first bounce in order to successfully make the third (and final) bounce for another jumper. These participants seemed to "want to please everybody." The end result was that they ended up saving fewer jumpers than other participants who were willing to "sacrifice the few to save the many."

These participants allowed themselves to focus on a task objective that was not the understood priority. Intentionally sacrificing a jumper to save another is an uncomfortable decision. Some participants based their multitask strategy on their own feeling of performance comfort instead of the stated priority. The existence of this kind of performance bias could be a critical topic for system design. A future research project could be conducted to determine ways of focusing participants on the stated task performance priority. Also, a measure could be constructed to identify individuals who allow their biases to easily affect their performance.

7.3. Subjective Preference of Design Solution

It was observed that there were sizable numbers of participants that said they preferred each of the different solutions. About half of the 36 participants (17 out of 36) said they preferred the negotiated interruption solution (see Section 6.6). This generalization, however, does not show that the other half of the participants picked other solutions as their first choices: Six chose immediate, seven chose mediated, and five chose scheduled (there was one missing data point for a participant that did not make a valid preference choice ranking). Nine participants had negotiation as their second choice, 4 participants had it as third, and 5 participants had it as last. This means that there was a full one fourth of the participants that did not like the negotiated solution. An interface design that is meant to please everyone must, therefore, give people choices about how they want their interruptions managed.

7.4. Effect of a "Context of Interruption"

Participants' behavior on the game task changed when interruptions were added. Members of one of the Latin squares order groups saw the game only condition for their first practice trial. It was observed that during the game only conditions these 6 participants took their time making game actions. It seemed that they were following a "plan for tomorrow" strategy. They seemed to use all possible time in considering every possibility, and then wait to make game actions until needed so as to achieve the most efficient and effective plan. These participants changed their behavior, however, when they encountered their first interrupt-laden experimental condition. Participants changed their strategy to "live for today, for tomorrow we'll be interrupted." They began making actions in anticipation of where the stretcher-bearers would need to be next. As soon as a bounce was made, participants would have already decided where the stretcher-bearers would be needed next and move them there immediately. This strategy made more sense with the possibility of interruption because moving the stretcher-bearers in anticipation meant that if an interruption happened the next needed bounce could still take place out of sight, and participants could potentially get back to the game before having to make their next move. This "act now" strategy was prudent in the face of interruptions, but it came at the cost of subtlety in planning on the game task. Future research would be useful to determine these subtle effects caused by the possibility of interruption.

7.5. Perception of Being in Command of the Machine

It was observed that some participants learned a more extensive set of typing actions for performing the game task than just "move left" and "move right." Some participants invented a double strike action for moving stretcher-bearers the two spaces between the left and right extremes. This double strike was delivered as a single action to move the stretcher-bearers two places without any intention of pausing in the middle location. Other participants, however, did not learn this convention, and would make two deliberate move actions even when it was their clear intention to move two spaces.

Also, the participants who did not learn the double strike action, seemed to make more "redundant move on game task" keying errors than other participants. For example, when they wanted to quickly move the stretcher-bearers two spaces from the left-most position to the right-most position, they would hit the Move Right button several times. The game task responded quickly to move commands, and the stretcher-bearers would visibly be at the right-most location after the participant's first two key presses. The participants, however, kept pressing the Move Right key until they could glance down to the bottom of the window and verify the position of the stretcher-bearers with their foveal vision. It seemed that they did not trust that their key presses were being obeyed and needed to consciously verify that action was taken. It appeared that these participants did not perceive their game key actions as directly commanding–controlling the game. Instead, they perceived their actions as only entreaties to the computer and that somehow the computer could not be relied on to directly execute their commands. Future research is needed to verify this observation about differences in perceived degree of direct control over the computer. Design solutions may be discovered that would afford perception of direct control when warranted.

7.6. Strategies for Resuming the Game Task After Interruption

Participants sometimes made many unnecessary key presses when they tried to resume the game task after finishing a matching task. Some participants were relatively unsuccessful at recovering the game task quickly. They seemed to execute a "panic until full recovery" strategy. They might have reasoned, "I see that I need to make several well-timed movements right now to save all the jumpers, I do not have time to plan the timing, so I'm just going to make a bunch of actions and hope that it all works out by chance." Usually this strategy resulted in missing all the current jumpers, and then a few seconds later participants would recover their game-playing ability all at once. Other participants were more successful. They seemed to execute a "recover one at a time" strategy. They might have reasoned, "I see that I do not now have good enough situation awareness of the game to save all the current jumpers; therefore, I will focus my actions on saving that one jumper there, and I will purposely miss all the rest until I get reoriented." Then one by one they would retake responsibility for the falling jumpers, until they again could handle the big picture.

Future research is needed to validate this observation. A potential measure for this individual difference could have important implications for UI design and for the process of selecting individuals for certain highly critical multitasks.

7.7. Self-Monitoring—Reflection During Action

It was observed that participants were acutely aware whenever they made matching errors, indicated by exclamations like "Whoops!" or "Damn!" It seemed that they were continuously running a meta-level process of reviewing and reflecting on their own task performance, as if they were providing themselves feedback on how well they were doing. After completing each matching task it would go away, and then participants would cognitively question whether they had succeeded or not. When they made a matching error they seemed to say to themselves, “Whoops, I made an error; I’ve got to do better.” (Note that the experimental multitask provides no feedback whatever about whether a matching task is completed correctly or incorrectly.)

This self-awareness, however, was not in effect for keying errors. These UI manipulation errors did not directly affect the task, and were therefore not internalized by participants. When they made a keying error and their intended action was not carried out by the computer, they seemed to say to themselves, “What’s wrong now? Oh the stupid computer didn’t understand me.”

The topic of reflection during action could be an important research domain for future work. This meta-level cognition seems to happen automatically with little perceived effort. The negotiated interruption solution has obvious potential for tapping into this innate ability.

7.8. Perception of Interruption Frequency

In general, participants did not have an accurate idea of how many times they had been interrupted. In one of the pilot tests, the number of jumpers was held constant across eight trials of the immediate condition, but the numbers of matching tasks for each 4.5-min trial was changed over a range from 38 matches for the first trial to 87 matches for the last trial. This was done to determine a reasonable frequency of interruptions for the multitask. It was discovered in an exit interview with the participant of this pilot test that they were completely unaware that the number of matching tasks had been varied among trials. It was also observed in the exit questionnaire of the actual experiment that many participants believed there were different numbers of matching tasks presented in the trials of the different experimental conditions. Some participants explained that they were unaware of how many Matches they had done because they did not perceive the multitask as a whole, but instead saw the matching task as an annoying thing that happened intermittently. A paraphrased quote is, “Matches had to be handled effectively, but the goal was just to get them out of my face.” This perspective is similar to a metaphor of swatting mosquitoes while trying to work in the garden. Each mosquito has to be hit just right, but since that is not perceived as part of the real work, the total number killed is not remembered.

There is some evidence in the literature that people’s preference of coordination strategy can be dependent on the frequency of interruptions. People are not consciously aware of the number of interruptions they encounter; however they may gather and use information about interruption frequency at a preattentive level to decide which coordination strategy is most appropriate. Zijlstra and Roe (1999) found that, when there were few telephone interruptions, people in an office environment adopted an immediate coordination solution, but, where there were frequent interruptions, they adopted a negotiated solution. A future study should be conducted with varying numbers of interruptions to validate that the guidelines presented here generalize to low-frequency interruption environments.

7.9. Discussion of Workload Metric

An unexpected positive feedback loop was observed and was caused by the particular workload metric used in the mediated solution for this experiment. This feedback loop did not significantly affect the results of this experiment, but its potential for mischief in implementations of the mediated solution makes it worth discussing here. The difficulty of the game was determined by how many jumpers were in the air at a time. The mediator was designed to save interruptions in a queue when the game task was difficult, and then dump them on people when the game became less difficult. The problem came because, while participants were doing the matching tasks, all the new jumpers on the game task were not visible and were falling to destruction; then, of course, when people returned to the game task, there would be no jumpers in the air. The mediator would conclude that this was a good time to interrupt people, because the difficulty metric said the workload was low, and then would pass through the next few interruptions without pause until participants could manage to get a couple of jumpers in the air. One way to avoid this feedback loop would be to have the workload metric dependent on the observed jumping rate for the game task instead of dependent on the degree of participants' success in bouncing jumpers.

Future research is needed to gain design wisdom about how to implement the mediated solution. For example, should the mediator be based on metrics of current human workload or on metrics of current task demands? It was observed that these two kinds of metrics are not the same, because people may have inaccurate perceptions about current task demands.

7.10. Perceived Responsibility

Interruptions were annoying for all participants. However, it was observed that different participants had different perspectives on what interruptions meant to them and therefore why interruptions were annoying. Some participants internalized responsibility for the experimental multitask. For them, interruptions were bothersome because interruptions interfered with their ability to successfully perform the multitask. These participants "cared" about the fate of the jumping characters in the game task. When these participants would save a jumper or miss a jumper they would have an emotional reaction, and their motivation seemed to stem from pursuit of positive internal reactions and avoidance of negative internal reactions. They also seemed to care about performing the matching task correctly. Their sense of accomplishment for this effort seemed to be based on the degree of success achieved in performing the multitask. Other participants externalized responsibility for the experimental multitask. For these participants, interruptions were bothersome because they increased the required mental effort for performing the multitask. These participants were not necessarily less motivated than the participants who internalized the multitask; their motivation was just based on something else. These participants cared about their own integrity. They perceived a need to fulfill a commitment. Their sense of accomplishment for this effort seemed to be based not on the degree of success on performing the multitask, but on how successful they were at giving 100% of their effort.

Future research needs to be done to validate this observed individual difference in perceived responsibility. The results might be used to create a useful prediction of how people's performance will change over time. It is reasonable to expect that the effectiveness of different personnel management strategies would vary between people with different perspectives on their responsibility for accomplishing tasks.

7.11. Combining Design Solutions

Three participants volunteered the idea that a combination of interruption coordination methods would be more successful than any of the four basic methods in isolation. They each suggested a combination of the negotiated and mediated solutions. The mediated, they said, would be best as the default, but the UI should allow them to override the mediator at any time and switch to a negotiated solution. It would be useful to conduct a future research project to examine the utility of combining the four primary solutions for coordinating user interruption.

There is some evidence in the literature that a combination of interruption coordination methods may be best solution for some tasks. Cook, Corbridge, Morgan, and Turpin (1999) propose dynamic function allocation (DFA) for managing automation in Naval Command and Control systems. They, however, have found that giving the human user explicit control over DFA scheduling has some significant advantages over the alternative solution of putting the automation in control of DFA. This "explicit DFA" is a combination of the negotiated solution and the mediated solution. Users are in control of negotiating with the automation for the services provided by the mediator. It was found that people prefer to explicitly control the mediation services provided by the computer system.

7.12. Other Future Work

There are other important design questions that need to be answered to increase the general usefulness of the design guidelines proposed in the following section. These future topics include multitasks with subtasks of differing priority; multitasks where the separate tasks are dependent or interdependent; interruption tasks with varying or unpredictable lengths; multitask subtasks with more or varying level of memory requirement of users; and multitask subtasks with more or varying level of cognitive processing requirement of users.

8. CONCLUSIONS

This article presents the first empirical comparison of four primary UI design solutions to the problem of coordinating the interruption of people in HCI. This topic is an important factor for UI design of systems that use a delegation/supervision style of HCI. These systems cause user interruptions, and interrupting people can degrade their performance and cause them to make serious mistakes. The results of this experiment support a set of UI design guidelines for enabling people to process interruptions most successfully. Each guideline is relative to a different performance metric and summarizes a statistically significant finding about the comparative utility of the four alternative design solutions for causing best or worst performance. The overall result is that the negotiation-based solution is best except for cases where small differences in timeliness of beginning the interruption task are critical, and then the immediacy-based solution is best.

NOTES

Background. This article contains revised and/or broadened pieces of two previous publications: a paper presented at INTERACT '99 (McFarlane, 1999) and the author's doctoral dissertation (McFarlane, 1998). The majority of the results and discussion presented in this article, however, are unique to this document. This research was conducted as part of the Naval Research Laboratory's HAIL Project (Human Alerting and Interruption Logistics). The HAIL project homepage is available at http://www.aic.nrl.navy.mil/hail/.

Acknowledgments. The manuscript was submitted while the author was working at the Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence. Portions of this work were performed as part of a doctoral dissertation at GWU under the direction of John Sibert. Editing contributions for this article were made by James Ballas, Astrid Schmidt-Nielsen, Susan McFarlane, Derek Brock, Justin McCune, William Lawless, Kara Latorella, and Jeff Richmond.

Support. Major funding for this work was provided by James Ballas at the Naval Research Laboratory. Additional funding was provided by Michael Shneier at the Office of Naval Research, and by John Sibert at George Washington University, and by Gerard Mayer at Lockheed Martin Advanced Technology Laboratory.

Authors' Present Addresses. Daniel McFarlane, Lockheed Martin Advanced Technology Laboratories, 1 Federal St., A&E–3W, Camden, NJ, 08102, USA. E-mail: mcfarlane@acm.org Project Homepage: http://www.aic.nrl.navy.mil/hail/.

HCI Editorial Record. First manuscript from Daniel McFarlane received on July 6, 1999. Revision received August 7, 2000. The *HCI* editors decided to split the submission into two manuscripst. First version of this manuscript received October 29, 2001. Revision received November 16, 2001. Accepted by Ruven Brooks. Final manuscript received December 18, 2001. — *Editor*

REFERENCES

Aronson, E. (1995). *The social animal.* New York: W. H. Freeman.

Ballas, J. A., Heitmeyer, C. L., & Pérez, M. A. (1992a). *Direct manipulation and intermittent automation in advanced cockpits* (NRL Formal Report NRL/FR/5534—92–9375). Washington, DC: The Naval Research Laboratory.

Ballas, J. A., Heitmeyer, C. L., & Pérez, M. A. (1992b). *Evaluating two aspects of direct manipulation in advanced cockpits.* Paper presented at CHI '92 Conference on Human Factors in Computer Systems, ACM, New York.

Berger, T., Kamoun, A., & Millot, P. (1988). *Real time measurement of workload in discrete multitask situations and extensions to continuous tasks.* International Conference on Human Machine Interaction and Artificial Intelligence in Aeronautics and Space, Toulouse-Blagnac, France.

Braune, R., & Wickens, C. D. (1986). Time sharing revisited: Test of a componential model for the assessment of individual differences. *Ergonomics, 29*, 1399–1414.

Burton, R. R., & Brown, J. S. (1979). An investigation of computer coaching for informal learning activities. *International Journal of Man–Machine Studies, 11*, 5–24.

Cabon, P., Coblentz, A., & Mollard, R. (1990). Interruption of a monotonous activity with complex tasks: Effects of individual differences. *Proceedings of the Human Factors Society 34th Annual Meeting,* Orlando, FL, 912–916.

Cellier, J. M., & Eyrolle, H. (1992). Interference between switched tasks. *Ergonomics, 35*(1), 25–36.

Chapanis, A. (1978). *Interactive communication: A few research answers for a technological explosion.* Text of an invited address given at the 86th Annual Convention of the American Psychological Association, Toronto, Ontario, Canada. (ERIC Document Reproduction Service No. ED168084/CS502432)

Cherry, E. C. (1953). Some experiments on the recognition of speech with one or two ears. *Journal of the Acoustical Society of America, 25*, 975–979.

Chignell, M. H., & Hancock, P. A. (1988). Intelligent interface design. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 969–995). New York: Elsevier.

Clark, H. H. (1996). *Using language.* New York: Cambridge University Press.

Cohen, S. (1980). Aftereffects of stress on human performance and social behavior: A review of research and theory. *Psychological Bulletin, 88*(1), 82–108.

Cook, C., Corbridge, C., Morgan, C., & Turpin, E. (1999). Investigating methods of dynamic function allocation for naval command and control. *Proceedings of People in Control, Bath, UK, IEEE Conference Publication No. 463,* 388–393.

Cooper, R., & Franks, B. (1993). Interruptibility as a constraint on hybrid systems. *Minds & Machines, 3*(1), 73–96.

Cypher, A. (1986). The structure of user's activities. In D. A. Norman & S. W. Draper (Eds.), *User centered system design* (pp. 243–263). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Czerwinski, M. P., Chrisman, S. E., & Rudisill, M. (1991). *Interruptions in multitasking situations: The effects of similarity and warning* (Tech. Rep. JSC–24757). Houston, Texas: National Aviation and Space Administration, Lyndon B. Johnson Space Center.

Czerwinski, M., Chrisman, S. E., & Schumacher, B. (1991). The effects of warnings and display similarities on interruption in multitasking environments. *SIGCHI Bulletin, 23*(4), 38–39.

Czerwinski, M., Cutrell, E., & Horvitz, E. (2000a, December). Instant messaging and interruption: Influence of task type on performance. *Proceedings of OZCHI 2000, Sydney, Australia.*

Czerwinski, M., Cutrell, E., & Horvitz, E. (2000b). Instant messaging: Effects of relevance and time. In S. Turner & P. Turner (Eds.), *People and computers XIV: Proceedings of HCI 2000, Vol. 2, British Computer Society*, 71–76.

Davies, S. P., Findlay, J. M., & Lambert, A. J. (1989). The perception and tracking of state changes in complex systems. In G. Salvendy & M. J. Smith (Eds.), *Designing and using human-computer interfaces and knowledge based systems* (pp. 510–517). Amsterdam: Elsevier.

Field, G. E. (1987). Experimentus interruptus. *ACM SIGCHI Bulletin, 19*(2), 42–46.

Galdes, D. K., & Smith, P. J. (1990). Building an intelligent tutoring system: Some guidelines from a study of human tutors. *Proceedings of the Human Factors Society 34th Annual Meeting,* 1407–1411.

Gaver, W. W. (1989). The sonicfinder: An interface that uses auditory icons. *Human–Computer Interaction, 4*, 67–94.

Gillie, T., & Broadbent, D. E. (1989). What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological Research, 50*(4), 243–250.

Hall, B. L., & Hursch, D. E. (1982). An evaluation of the effects of a time management training program on work efficiency. *Journal of Organizational Behavior Management, 3*(4), 73–96.

Hammer, J. M., & Small, R. L. (1995). An intelligent interface in an associate system. In W. B. Rouse (Ed.), *Human/technology interaction in complex systems* (Vol. 7, pp. 1–44). Greenwich, CT: JAI.

Hess, S. M., & Detweiler, M. C. (1994). Training to reduce the disruptive effects of interruptions. *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting,* 1173–1177.

Husain, M. G. (1987). Immediate and delayed recall of completed-interrupted tasks by high and low anxious subjects. *Manas, 34*(1–2), 67–71.

Ishihara, S. (1996). *Ishihara's tests for colour-deficiency, concise edition.* Tokyo, Japan: Kanehara & Co.

Jensen, A. R., & Rohwer, W. D., Jr. (1966). The Stroop color-word test: A review. *Acta Psychologica, 25*, 36–93.

Jessup, L. M., & Connolly, T. (1993). The effects of interaction frequency on the productivity and satisfaction of automated problem-solving groups. *Proceedings of the Twenty-Sixth Hawaii International Conference on System Sciences,* 142–151.

Joslyn, S., & Hunt, E. (1998). Evaluating individual differences in response to time-pressure situations. *Journal of Experimental Psychology: Applied, 4*(1), 16–43.

Karis, D. (1991). Evaluating transmission quality in mobile telecommunication systems using conversation tests. *Proceedings of the Human Factors Society 35th Annual Meeting.*

Katz, R. (1995). Automatic versus user-controlled methods of briefly interrupting telephone calls. *Human Factors, 37*, 321–334.

Keppel, G. (1991). *Design and analysis: A researcher's handbook.* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction—Why an aid can (and should) go unused. *Human Factors, 35*, 221–242.

Kirschenbaum, S. S., Gray, W. D., Ehret, B. D., & Miller, S. L. (1996). *When using the tool interferes with doing the task.* Short paper published in the conference companion of the ACM CHI '96 Conference Human Factors in Computing Systems, Vancover, British Columbia, Canada.

Kreifeldt, J. G., & McCarthy, M. E. (1981). Interruption as a test of the user-computer interface. *Proceedings of the 17th Annual Conference on Manual Control, JPL Publication,* 81–95.

Latorella, K. A. (1996a). Investigating interruptions: An example from the flight deck. *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting,* Santa Monica, CA, 87–91.

Latorella, K. A. (1996b). *Investigating Interruptions: Implications for flight deck performance.* Doctoral dissertation, State University of New York at Buffalo (also published in 1999 as NASA Technical Memorandum 209707, National Aviation & Space Administration, Washington, DC).

Latorella, K. A. (1998). Effects of modality on interrupted flight deck performance: Implications for data link. *Proceedings of the Human Factors and Ergonomics Society 42th Annual Meeting.*

Lee, W. O. (1992). The effects of skill development and feedback on action slips. *Proceedings of the HCI '92 Conference on People and Computers VII.*

Lieberman, H. (1997). *Autonomous interface agents.* Paper presented at CHI '97 Conference on Human Factors in Computer Systems, ACM, New York.

Lustig, M. W. (1980). Computer analysis of talk-silence patterns in triads. *Communication Quarterly, 28*(4), 2–12.

McCarthy, J. C., & Monk, A. F. (1994). Channels, conversation, cooperation and relevance: All you wanted to know about communication but were afraid to ask. *Collaborative Computing, 1*, 35–60.

McDonald, S., & Stevenson, R. J. (1996). Disorientation in hypertext: The effects of three text structures on navigation performance. *Applied Ergonomics, 27*(1), 61–68.

McFarlane, D. C. (1997). *Interruption of people in human-computer interaction: A general unifying definition of human interruption and taxonomy* (NRL Formal Report NRL/FR/5510—97–9870). Washington, DC: Naval Research Laboratory.

McFarlane, D. C. (1998). *Interruption of people in human-computer interaction*. Unpublished doctoral dissertation, George Washington University, Washington, DC.

McFarlane, D. C. (1999). Coordinating the interruption of people in human-computer interaction. In M. A. Sasse & C. Johnson (Eds.), *Human-Computer Interaction—INTERACT '99* (IFIP TC.13, pp. 295–303). Edinburgh, England: IOS.

McFarlane, D. C., & Latorella, K. A. (2002). The scope and importance of human interruption in HCI design. *Human-Computer Interaction*, *17*, 1–61.

Miyata, Y., & Norman, D. A. (1986). Psychological issues in support of multiple activities. In D. A. Norman & S. W. Draper (Eds.), *User centered system design* (pp. 265–284). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Moray, N., Dessouky, M. I., & Kijowski, B. A. (1991). Strategic behavior, workload, and performance in task scheduling. *Human Factors, 33*, 607–629.

Morrin, K. A., Law, D. J., & Pellegrino, J. W. (1994). Structural modeling of information coordination abilities: An evaluation and extension of the Yee, Hunt, and Pellegrino Model. *Intelligence, 19*, 117–144.

National Transportation Safety Board. (1988). *Aircraft accident report, Northwest Airlines Inc., McDonnell-Douglas DC–9–82, N312RC, Detroit Metropolitan Wayne County Airport, Romulus, Michigan* (NTSB–AAR–88–05). Washington, DC: National Transportation Safety Board.

Nintendo. (1980–1981). *Game & watch: Fire*. Nintendo of America Inc., P.O. Box 97032, Redmond, WA 98073–9732.

Nintendo. (1997). *Game & watch gallery, for Game Boy*. Nintendo of America Inc., P.O. Box 97032, Redmond, WA 98073–9732.

Oberg, B., & Notkin, D. (1992). Error reporting with graduated color. *IEEE Software, 9*(6), 33–38.

Obermayer, R. W., & Nugent, W. A. (2000). *Human-computer interaction for alert warning and attention allocation systems of the multi-modal watchstation*. SPIE 2000, SPIE–The International Society for Optical Engineering, Bellingham, WA.

Osga, G. A. (2000). 21st century workstations—Active partners in accomplishing task goals. *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting,* San Diego, CA.

Raby, M. & Wickens, C. D. (1991). Strategic behaviour in flight workload management. *Proceedings of the Sixth International Symposium on Aviation Psychology,* Columbus, OH, 1130–1135.

Rich, C. (1996). Window Sharing with Collaborative Interface Agents. *SIGCHI Bulletin, 28*(1), 70–78.

Rouncefield, M., Hughes, J. A., Rodden, T., & Viller, S. (1994). Working with "constant interruption:" CSCW and the small office. *Proceedings of the CSCW '94 conference on computer-supported cooperative work* (pp. 275–286). New York: ACM.

Sheridan, T. B. (1987). Supervisory control. In G. Salvendy (Ed.), *Handbook of human factors* (pp. 1243–1268): New York: Wiley.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Smith, S. L., & Mosier, J. N. (1986). *Guidelines for designing user interface software* (Report ESD–TR–86–278). Bedford, MA: MITRE.

Speier, C., Valacich, J. S., & Vessey, I. (1997). The effects of task interruption and information presentation on individual decision making. *Proceedings of the Eighteenth International Conference on Information Systems* (pp. 21–36). New York: Association for Computing Machinery.

Storch, N. A. (1992). *Does the user interface make interruptions disruptive? A study of interface style and form of interruption* (UCRL–JC–108993; report number DE92011295). Springfield, VA: Lawrence Livermore National Laboratory, distributed by the National Technological Information Service.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.

Tulga, M. K., & Sheridan, T. B. (1980). Dynamic decisions and work load in multitask supervisory control. *IEEE Transactions on Systems, Man, & Cybernetics, SMC–10*(5), 217–232.

Van Bergen, A. (1968). *Task interruption.* Amsterdam: North-Holland Publishing Company.

Wagenaar, W. A. (1969). Note on the construction of digram-balanced latin squares. *Psychological Bulletin, 72*, 384–386.

Weiner, B. (1965). Need achievement and the resumption of incompleted tasks. *Journal of Personality and Social Psychology, 1*(2), 165–168.

West, C. (1982). Why can't a woman be more like a man? An interactional note on organizational game-playing for managerial women. *Work and Occupations, 9*(1), 5–29.

Williams, C. L. (1995). Potential conflicts of time sharing the flight management system control display unit with data link communications. *Proceedings of the Eighth International Symposium on Aviation Psychology,* 341–347.

Woods, D. D. (1995). The alarm problem and directed attention in dynamic fault management. *Ergonomics, 38*(11), 2371–2393.

Zijlstra, F. R. H., & Roe, R. A. (1999). Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology, 72*, 163–185.

Zimmerman, D. H., & West, C. (1975). Sex roles, interruptions and silences in conversations. In B. Thorne & N. Henley (Eds.), *Language and sex: Difference and dominance* (pp. 105–129). Rowley, MA: Newbury.

APPENDIX A: QUESTIONNAIRES

A1. Entrance Questionnaire

Note that this appendix only contains the text from the entrance questionnaire and does not contain the original formatting or space for participants' answers to questions.

[The answers you give will be kept strictly confidential. In any report we might publish, we will not include any information that might make it possible to identify you as a participant. (Please do not write your name on this questionnaire.)]

1. Your sex? (a) male (b) female
2. Your age?
3. Your dominant hand? (a) right (b) left
4. High school graduate? (a) yes (b) no

4a. If "yes," how many years of education since high school? 0 1 2 3 4 5 6 7 8 9 >9

[On the following questions, please make a mark on the answer line to indicate where your answers fall between the two extremes of possible answers. These questions ask for your informal judgments of your own experiences, abilities, and preferences. Therefore, your answers can not be "correct" or "incorrect."

We realize that you may not have experience answering opinion questions on paper. Some people find it difficult because of this unnatural context. It may help to try to imagine how you would answer each question if it were asked in a more natural context. For example, suppose you are having lunch with some friends and acquaintances and one of them asks …]

5. How much computer experience do you have (i.e., amount of time spent working on computers)? [none – considerable]
6. How skilled are you with computers (i.e., proficiency with computer tasks)? [no skill – expert]
7. How much video game experience do you have (i.e., amount of time spent playing video games)? [none – considerable]
8. How skilled are you with video games (i.e., proficiency with video games)? [no skill – expert]
9. How skilled are you at juggling (i.e., proficiency juggling physical objects)? [no skill – expert]
10. How much typing experience do you have (i.e., amount of time spent typing)? [none – considerable]
11. How skilled are you at typing (i.e., proficiency typing)? [no skill – expert]
12. How skilled are you at touch-typing (i.e., proficiency typing without looking at keyboard)? [no skill – expert]
13. How much experience do you have performing more than one task at a time (i.e., amount of time spent performing multiple tasks at the same time by switching back and forth between different tasks)? [none – considerable]
14. How skilled are you at performing more than one task at a time (i.e., proficiency)? [no skill – expert]

15. To what degree do interruptions affect you (i.e., to what degree do interruptions negatively affect your ability to perform tasks)? [none – considerable]
16. To what degree do distractions affect you (i.e., to what degree do distractions negatively affect your ability to perform tasks)? [none – considerable]
17. How much do you try to avoid distractions and interruptions when working (i.e., amount of effort and planning you normally expend to avoid distractions and interruptions when you must get things done)? [none – considerable]

A2. Exit Questionnaire

Note that this appendix only contains the text from the exit questionnaire and does not contain the original formatting or space for participants' answers to questions. Also note that participants were given only Greet letters for the names of the four experimental conditions. For reference in this article, these letters stand for the following: ψ = *scheduled condition;* ξ = *immediate condition;* δ = *negotiated condition; and* λ = *mediated condition.*

[The answers you give will be kept strictly confidential. In any report we might publish, we will not include any information that might make it possible to identify you as a participant. (Please do not write your name on this questionnaire.)]

[On the following questions please make a mark on the answer line to indicate where your answers fall between the two extremes of possible answers. These questions ask for your informal judgments of your own experiences, abilities, and preferences. Therefore, your answers cannot be "correct" or "incorrect."]

18. How much anxiety did you feel during this experiment? [no anxiety – considerable]
19. How motivated did you feel while performing the experimental trials? [not motivated – extremely motivated]

[The following questions ask about your perceptions and opinions of the different conditions of the experiment. Please refer to the written instructions as a reminder of the identities of the different UI designs denoted with the Greek letters ψ ξ δ λ. In questions that ask for a ranking, no ties please.]

20. Please rank the conditions ψ ξ δ λ by how well you liked or preferred them as 1 2 3 and 4 (1 = most liked, 4 = least liked).

21. Rank the conditions ψ ξ δ λ by how easily they allowed you to perform the dualtask as 1 2 3 and 4 (1 = most easy, 4 = least easy).
22. Rank the conditions ψ ξ δ λ by how many errors you made on the matching task as 1 2 3 and 4 (1 = least errors, 4 = most errors).
23. Rank the conditions ψ ξ δ λ by how many errors you made on the game task as 1 2 3 and 4 (1 = least errors, 4 = most errors).
24. Rank the conditions ψ ξ δ λ by how much stress you felt while performing the computer dualtask as 1 2 3 and 4 (1 = least stress, 4 = most stress).
25. Rank the conditions ψ ξ δ λ by how interrupted you felt while performing the computer dualtask as 1 2 3 and 4 (1 = least interrupted, 4 = most interrupted).
26. Rank the conditions ψ ξ δ λ by how distracted you felt while performing the computer dualtask as 1 2 3 and 4 (1 = least distracted, 4 = most distracted).
27. Rank the conditions ψ ξ δ λ by how well you were able to predict the time interval between interruptions (i.e., how long it would be until you would stop performing the game task and begin performing a matching task) as 1 2 3 and 4 (1 = most predictable interruptions, 4 = least predictable interruptions).
28. Rank the conditions ψ ξ δ λ by how busy with the game task you were likely to be when interrupted (i.e., how busy with the game task you were likely to be when you had to stop performing the game task and begin performing a matching task) as 1 2 3 and 4 (1 = least busy, 4 = most busy).
29. Rank the conditions ψ ξ δ λ by how complex the game task was likely to be when you had to resume playing the game after being interrupted (i.e., how complex the game task was likely to be after you finished performing the matching task(s) and begin to perform the game task again) as 1 2 3 and 4 (1 = least complex, 4 = most complex).
30. In condition ψ, while performing the game task, how well were you able to anticipate the next 25-sec cycle of interruptions (i.e., the next switch to the queued matching tasks)? [no anticipation – considerable]
31. How much did you like the direct control over when to process interruptions provided by condition δ? [none – considerable]
32. Was the direct control over when to process interruptions provided by condition δ useful for performing the computer dualtask? [not useful – useful]
33. In condition δ, how much extra work was it to have to deliberately switch the matching task on and off? [no extra work – considerable]
34. In condition δ, how distracting were the flashes of the pager that announced the occurrences of matching tasks? [not distracting – extremely distracting]

35. In condition δ, it was possible for a trial to end without you having attempted all of the announced matching tasks. How many of the total number of matching tasks did you complete before the trial ended? [none completed – all completed]
36. In condition λ, how well was the computer able to judge the difficulty of the game task (i.e., how well did the computer schedule the presentation of the matching tasks so that you performed the matching tasks only when the game task was less demanding)? [not well – very well]
37. Did you notice that the game task was less complex under any of the conditions ψ ξ δ λ (i.e., did some conditions have fewer total jumping diplomats)? (a) yes (b) no
37a. If "yes," please describe.
38. Did you notice that the matching task was less complex under any of the conditions ψ ξ δ λ (i.e., did some conditions have fewer total matching tasks)? (a) yes (b) no
38a. If "yes," please describe.

[Blank space is provided below for any comments you have. (Please refer to particular experimental conditions by their Greek letters. Please refer to particular questionnaire questions by their numbers.)]

APPENDIX B. DETAILED RESULTS

B1. Experimental Effects of Repeated Measures

It was assumed that the structures imposed by the design of the experiment did not differentially affect participants' behavior across the treatment conditions (the different methods of coordinating interruption). A validation of these assumptions adds to the credibility of the observed main effect.

No experimental effects were expected. However, there were three main experimental design structures that could have affected participants' performance. The first two, trial sequence and treatment sequence, could potentially cause general practice effects. The third structure, Latin squares order grouping, could cause differential carryover effects.

General Practice Effects

General practice effects are changes in participants' performance caused by increasing exposure to the experimental context. These effects are caused by the processes of learning, fatigue, or boredom. Although learning, fatigue, and boredom are important topics, they are not relevant to the main hypothesis of this article. The experimental design has been contrived, therefore, to avoid these influences. General practice effects can be difficult to avoid because they require opposing controls. The practice period for the experiment had to be made long enough so that participants would have finished most of their learning of the multitask before they began the experimental trials, and the total length of the experiment had to be made short enough so that participants would not succumb to fatigue or boredom.

The following methods were included as attempts to control the potentially confounding general practice effects: 1 hr of practice on the same multitask used in the experiment; detailed written instructions and practice on all six of the different treatment conditions; difficulty ramping-up scheme for practice trial pairs; experimental conditions split into two 4.5-min trials each; experimental multitask modeled after an engaging video game; and uniqueness of multitask trials guaranteed with a constrained randomization scheduling scheme.

Trial sequence effects are seen when participants' behavior is significantly different between Trial 1 and Trial 2 irrespective of other comparisons. Figure B–1 shows a summary of an analysis of the data for trial sequence effects. This analysis was of the experimental data (not the practice data) of the four experimental conditions (not the two base cases). The Wilcoxon signed ranks test with correction for ties (denoted by W_i) was selected as an appropriate test (Siegel & Castellan, 1988). The Wilcoxon signed ranks test is useful for testing within-subjects effects like the Friedman test, but for the two-sample case. For comparison, W_i must be less than –1.96 for its p value to test above the chosen α level of 0.05 ($W_i < -2.58$ for $\alpha = 0.01$; and $W_i < -3.29$ for $\alpha = 0.001$).

Two of the nine performance metrics show unexpectedly significant trial sequence effects—avg. match speed and total keying errors. Figure B–2 shows that, although these effects are significant, they are relatively small and do not confound the main effect. Matching speeds are slightly less on Trial 2 than on Trial 1. This may be due to some learning still going on within the experimental conditions. Total keying errors are slightly more on Trial 2 than on Trial 1. This may be caused by an increase in fatigue or boredom on the second trial.

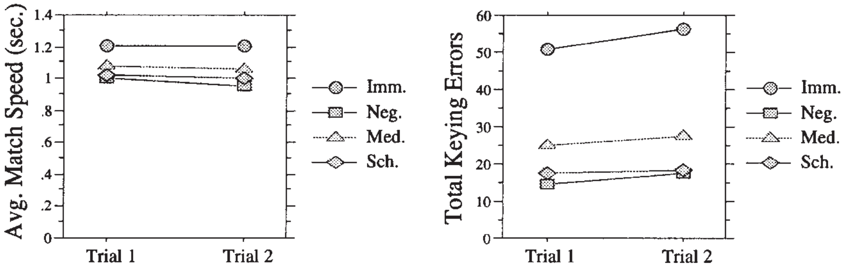
Treatment sequence effects are seen when participants' behavior changes significantly as a result of increasing exposure to the experimental environment over time irrespective of other influences. This experiment had 12 experimental treatments in sequence from first to last. Figure B–3 shows a summary of an analysis of the data for treatment sequence effects. This analysis was of the experimental data (not the practice data) of the four experimental conditions (not the two base cases). The effects of trial sequences were removed by summing the values from individual trials. The Friedman test was selected as an appropriate test. For comparison, F_r must be greater than 11.07 for its p value to test above the chosen α level of 0.05 ($F_r > 15.09$ for $\alpha = 0.01$; and $F_r > 20.52$ for $\alpha = 0.001$).

Figure B–1. **Effects of trial sequence.**

Performance Measure	W_{i}	p	$p < \alpha$
Jumpers saved	–1.255	.2095	No
G. keyed per saved[a]	–1.147	.2514	No
Task switches	–1.820	.0688	No
Matched wrong	–1.828	.0675	No
% M. wrong of done[b]	–1.854	.0638	No
Matches not done	–1.145	.2522	No
Average match age	–0.031	.9749	No
Average match speed	–3.095	.0020	Yes
Total keying errors	–3.266	.0011	Yes

[a] Number of key presses per jumper saved on game task. [b] Percent of matches done wrong of those attempted.

Figure B–2. **Graphical depiction of the two unexpectedly significant trial sequence effects—avg. matching speed and total keying errors. These graphs show the data for the experimental data (not the practice data) of the four experimental conditions (not the two base cases). Data are grouped by trial and split by experimental condition. Line charts show means. The game task had 59 subtasks (individual jumping diplomats). The matching task had 80 subtasks (individual matches). The total time for each trial was 4.5 min (270 sec). Imm. = immediate interruption; Neg. = negotiated interruption; Med. = mediated interruption; Sch. = scheduled interruption.**



In summary, (a) there were only two significant trial sequence effects for the nine measures of performance, and these two effects were relatively small and inconsequential; and (b) none of the nine performance metrics showed significant treatment sequence effects. The design of this experiment, therefore, effectively controlled learning, fatigue, and boredom, and successfully avoided the potentially confounding influences of general practice effects.

Differential Carryover Effects

Differential carryover effects are changes in participants' performance caused by interference between experimental conditions. The risk of invok-

Figure B–3. **Effects of treatment sequence.**

Performance Measure	F_r	p	$p < \alpha$
Jumpers saved	0.879	.9717	No
G. keyed per saved[a]	1.258	.9392	No
Task switches	0.147	.9996	No
Matched wrong	1.270	.9380	No
% M. wrong of done[b]	1.307	.9342	No
Matches not done	1.617	.8992	No
Average match age	0.343	.9968	No
Average match speed	0.556	.9899	No
Total keying errors	0.356	.9965	No

[a] Number of key presses per jumper saved on game task. [b] Percent of matches done wrong of those attempted.

ing these effects is always the negative consequence of using a repeated-measures design for experiments. There is a danger that the process of repeating variations of treatment conditions can result in the effects of one treatment condition persisting beyond its imposed completion boundary and affecting participants' performance on subsequent trials. Controls must be put in place to try to counteract this effect. These controls should ensure that participants are fresh for each new experimental trial and not still unencumbered by lingering effects of the previous experimental tasks.

The following methods were included as attempts to control the potentially confounding differential carryover effects: a digram-balanced Latin squares counterbalanced grouping scheme, 25-sec minimum rest periods imposed between all trials, a graphically neutral mask used to block the display during rest periods, consistent on-screen reminders of multitask instructions displayed before each trial, and detailed written instructions that were always available to participants for review.

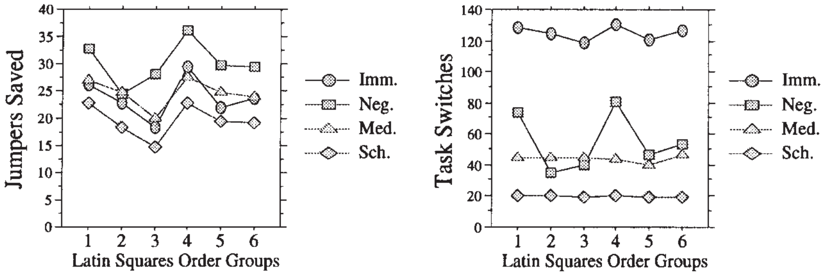
Treatment condition order groupings have potentially confounding influence when participants' behavior is significantly different between the different Latin squares order groupings irrespective of other comparisons. Figure B–4 shows a summary of an analysis of the data for condition order effects. This analysis was of the experimental data (not the practice data) of the four experimental conditions (not the two base cases). The Kruskal–Wallis one-way ANOVA by ranks with correction for ties (denoted by *KW*) was selected as an appropriate test (Siegel & Castellan, 1988). The Kruskal–Wallis test is useful for testing between-subjects effects for three or more sample cases. For comparison, *KW* must be greater than 11.07 for its p value to test above the chosen α level of 0.05 ($KW > 15.09$ for $\alpha = 0.01$; and $KW > 20.52$ for $\alpha = 0.001$).

Figure B–4. **Effects of Latin squares counterbalance ordering.**

Performance Measure	KW[a]	p	$p < \alpha$
Jumpers saved	15.552	.0082	Yes
G. keyed per saved[b]	9.526	.0899	No
Task switches	13.780	.0171	Yes
Matched wrong	8.265	.1422	No
% M. wrong of done[c]	8.054	.1533	No
Matches not done	8.852	.1151	No
Average match age	4.456	.4857	No
Average match speed	10.757	.0564	No
Total keying errors	7.015	.2196	No

[a] Kruskal–Wallis one-way analysis of variance by ranks with correction for ties. [b] Number of key presses per jumper saved on game task. [c] Percent of matches done wrong of those attempted.

Figure B–5. **Graphical depiction of the two unexpectedly significant Latin squares counterbalance order grouping effects—jumpers saved and task switches. These graphs show the data for the experimental data (not the practice data) of the four experimental conditions (not the two base cases). Data are grouped by ordering group and split by experimental condition. Line charts show means. The game task had 59 subtasks (individual jumping diplomats). The matching task had 80 subtasks (individual matches). The total time for each trial was 4.5 min (270 sec). Imm. = immediate interruption; Neg. = negotiated interruption; Med. = mediated interruption; Sch. = scheduled interruption.**



Two of the nine performance metrics show unexpectedly significant condition order grouping effects—jumpers saved and task switches. Figure B–5 shows that although these effects are significant, they do not confound the main effect. Participants in order groups 1 and 4 save more jumpers on average and have more task switching on average in the negotiated treatment condition than participants in the other groups.

These unexpected results mean that the order in which participants were presented the treatments affected their overall performance on saving jumpers. This interaction effect between order and condition defies those methods put in place to control differential carryover effects. Something important and strange must have affected participants. It appears that participants in Groups 1 and 4 learned different and more successful game-playing strategies than participants in other groups. On average, they switch more often and save more jumpers.

This order grouping effect may be interesting, but it is not part of the main hypothesis of this experiment. This analysis must determine whether the discovered order effect exerts a confounding influence on the main effect. The line charts in Figure B–5 show how participants' performance varied across order groupings. The Jumper Saved graph shows that, although participants' performance varies between groups, the pattern of scores across the different treatment conditions remains constant. Groups 1 and 4 just did better overall than other groups. The relative performance across treatment conditions, however, did not vary between order groupings. Therefore, this differential carryover effect is not a confound of the main effect.

The Task Switches graph in Figure B–5 shows that differences between order groupings differs only in the negotiated interruption condition. The Latin squares order grouping caused participants in Groups 1 and 4 to do more task switching on the negotiated interruption condition than participants in the other groups. There is some crossover between the negotiated condition and the mediated condition. This brings suspicion on the main effect results for task switches between these two treatment conditions. Figure 15 (post hoc analysis of main effect), however, reports that no significant difference was found here, so this is a nonissue.

These differential carryover effects are not confounding the main effect, but why are they there at all? The results of the Kruskal–Wallis tests reported in Figure B–4 (effects of Latin squares counterbalance ordering) rank order the six order groupings. The groups ranking (from highest to lowest) for jumpers saved is 4, 1, 6, 5, 2, 3, and for task switches is 1, 4, 6, 5, 2, 3—almost identical.

The "counterbalanced treatment order" figure (Figure 2) shows that the treatment conditions were order balanced for the different groups. Therefore, the most obvious explanation for a significant order effect is that the particular treatment condition that participants saw first differentially affected their process of constructing strategies for performing the dualtask. The powerful effect of first treatments is common and recognized in psychology as the "primacy effect" (Aronson, 1995).

Participants between groups saw different treatment conditions first. Using the 4, 1, 6, 5, 2, 3 rank for jumpers saved, participants from the six groups saw the following treatment conditions first: (Group 4) negotiated interruption, (Group 1) base case—game only, (Group 6) scheduled interruption, (Group 5) mediated interruption, (Group 2) base case—matching only, and (Group 3) immediate interruption.

It seems that participants formed rigid task strategies based on whatever treatment they saw first. A reasonable explanation for this order effect is degree of perceived control. If participants felt that they were in control of when to handle interruptions, then they formed more successful strategies than if they felt that they had no control. Participants who saw the negotiated solution first performed best, and participants who saw the immediate solution first performed worst.

The 1 hr of practice given to all participants was intended to negate any primacy effect from order of treatment conditions. However, it appears that participants formed rigid task strategies that differ depending on which treatment condition they experienced in their first practice trial. The primacy effect on these task strategies is not negated by 1 hr of practice. It is asserted that this order effect is the result of a stubborn primacy effect regarding participants' perception of degree of control implemented in the first treatment condition they encountered, and that this does not pose a confound to the main effect.

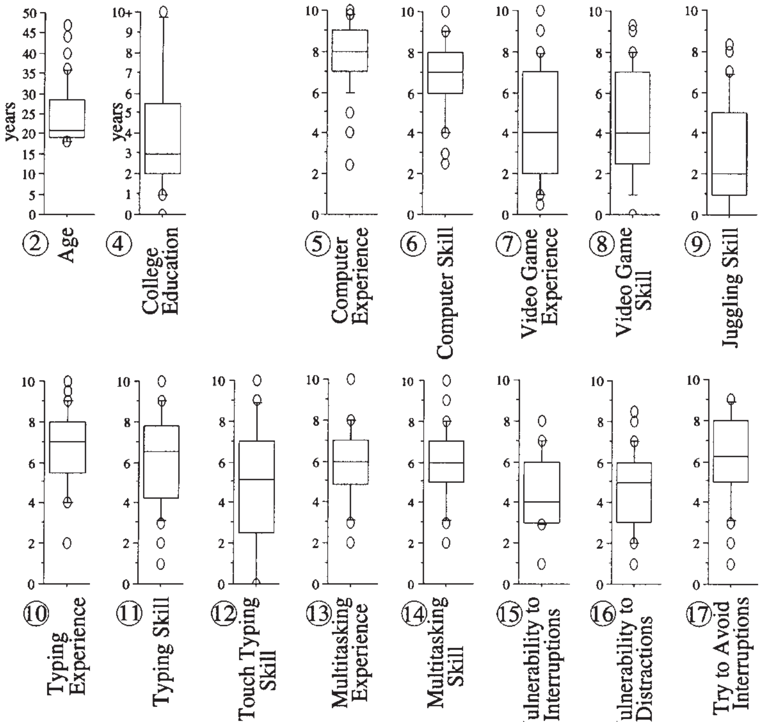
B2. Individual Differences

Each participant completed an entrance questionnaire (17 questions; see Appendix A1) before they were introduced to the experimental multitask. These questions were designed to measure biographical and self-perception information. Participants were asked to self-report their perceptions of their own experience and abilities on tasks that are potentially relevant for handling interruption during human–computer interaction. This questionnaire was administered to participants before they were told anything about the nature of the experiment. This constraint was an attempt to avoid any possible confounding influence on their responses.

Figure B–6 shows summaries of the results of the entrance questionnaire except for question 1, about sex, and question 3, about handedness. Participant volunteerism was constrained so that there were an equal number of male (18) and female participants (18). Two of the 36 participants (one male and one female) reported being left-handed and were subsequently allowed to perform the experimental multitask with their left hands. All questions of subjective self-perceptions (questions 5 through 17) were recorded by having participants mark a position on a free form number line from 0 (*least*) to 10 (*most*). Question number 6, for example, was not an objective measure about number of years of computer training, but a subjective judgment of how "skillful" participants considered themselves at performing computer-based tasks. See Appendix A for the actual questionnaire text.

These results show the diversity in the group of participants who participated in this experiment. A measure of race would have been interesting, but it was not included because race was irrelevant to the main hypothesis and because measuring race is emotionally charged and complicated. It was ob-

Figure B–6. **Measures of central tendency of participants' reports on the entrance questionnaire. The boxes in the box plots contain the center 50% of scores with the centerline at the median. The outer brackets enclose 80% of values, and extreme 10% outliers displayed as single points above and below. The box plots are numbered with their question numbers on the questionnaire. Graphics 2 and 4 show biographical information and are reported in units of years. Graphs 5 through 17 show self-perception information and are reported in subjective units from 0 (*none or no skill*) to 10 (*considerable or expert*).**



served indirectly, however, that the participant group was very racially diverse and probably had significant representatives of all of the major racial groups from the Washington, DC, area.

Sex differences for interruption have been proposed in the literature (West, 1982; Zimmerman & West, 1975). Participants' performance in this experiment, however, did not support general claims of sex differences for interruption events. Figure B–7 shows that there were no effects of sex on participants' different kinds of overall performance. This analysis was of the experimental data (not the practice data) on sums of the four experimental conditions (not the two base cases). The Mann–Whitney U test with correction for ties (de-

Figure B–7. **Effects of sex.**

Performance Measure	***MW*[a]**	***p***	***p* < α**
Jumpers saved	−1.092	.2749	No
G. keyed per saved[b]	−0.981	.3267	No
Task switches	−0.854	.3928	No
Matched wrong	−0.348	.7277	No
% M. wrong of done[c]	−0.316	.7517	No
Matches not done	−0.507	.6122	No
Average match age	−0.063	.9495	No
Average match speed	−0.443	.6578	No
Total keying errors	−0.854	.3929	No

[a] Mann–Whitney U test with correction for ties. [b] Number of key presses per jumper saved on game task. [c] Percent of matches done wrong of those attempted.

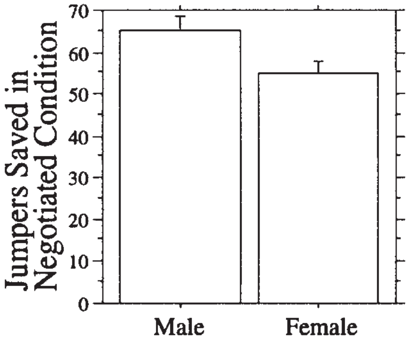
noted by *MW*) was selected as an appropriate test (Siegel & Castellan, 1988). The Mann–Whitney U test is useful for testing between-subjects effects for two cases. For comparison, *MW* must be less than −1.96 for its *p* value to test above the chosen α level of .05 ($MW < -2.58$ for $\alpha = .01$; and $MW < -3.29$ for $\alpha = .001$).

None of the nine performance metrics show significant main effects for sex. A similar analysis, also using the Mann–Whitney U test, showed no significant effects of handedness.

As a small aside, a post hoc analysis on sex was conducted using the Mann–Whitney U test. The results are not statistically meaningful because no main effect was found. However, the literature prediced strong sex differences that were not found, so it may be useful to grasp at straws on this one topic. This analysis looked for effects of sex on the 45 combinations of the nine performance measures split by condition. These are interactions between sex and condition for each of the nine measures. Only one of these 45 different kinds of performance implied an effect of sex—the jumper saved performance on the negotiated interruption treatment condition. The Mann–Whitney U test for this point resulted in $MW = -2.026$, $p = .0427$. Figure B–8 implies (nonstatistically) that male participants saved more jumpers on the game task under the negotiated interruption condition than female participants did. These results only serve to hammer home that this experiment found virtually no effects of sex whatever.

A correlation analysis revealed that there were no strong correlations between participants' answers to the entrance questionnaire and their performance on the experimental multitask. A correlation matrix was calculated to compare all 945 pairwise combinations of 15 entrance questionnaire topics (all questions except sex and handedness) and 63 kinds of performance on the

Figure B–8. **Bar charts of sex effect for the jumper saved performance on the negotiated interruption treatment condition. The bar charts show the mean with error bars that depict one standard error. Scores are for summed trials.**



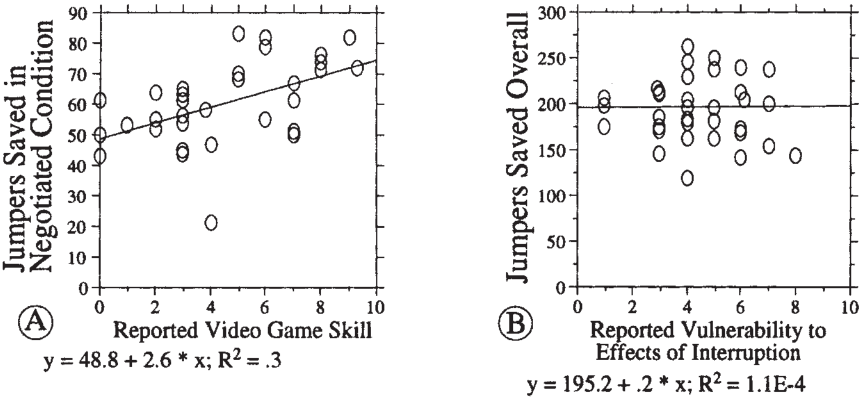
experimental multitask (the results of the six treatment conditions [with trials combined] for all 9 performance metrics, plus totals for each excluding base cases).

The results show none or weak correlations between participants' self-reports of their experience and skill and their performance on the multitask used in this experiment. Most pairs showed virtually no correlation. There were only two pairs out of the 945 combinations that had correlation coefficients greater than .5. These two correlations were between participants' reported levels of video game experience and skill and their jumper saved performance on the negotiated interruption treatment condition. The pair involving game skill is shown in Figure B–9A.

Instead of trying to interpret meaning into the sparse and weak correlations that were found, it is useful to look at the places where there may have been correlations but where none were found. The following examples are a very few of the many paired comparisons that had correlation coefficients smaller than .1 (virtually no correlation). They are examples of relationships that are rational and may have been reasonably expected, but that are conspicuously missing in the data:

1. Self-reported level of multitasking skill (question 14) and overall jumpers saved performance on the multitask. It seems reasonable to have expected that people with a high level of multitasking skill should have been able to do better on the game task than those with low skill.
2. Self-reported level of vulnerability to the effects of interruption (question 15) and overall matches not done performance on the multitask (see Figure B–9B). Participants that were more susceptible to the nega-

Figure B–9. **Scattergrams of correlations between participants' self-reported qualities and their subsequent performance on the experimental multitask. Graph A shows one of the two correlations in the correlation matrix of 945 pairs with a correlation coefficient greater than .5. The *y*-axis is scored as sums of the two trials for each of the 36 participants on the negotiated interruption treatment condition. Graph B shows a more typical result of the correlation analysis between self-reported ability and observed performance—no correlation whatever. The *y*-axis is scored as sums of the four treatment conditions (no base cases) for each participant.**



tive effects of interruption should have been less able to finish all matching tasks than those with more resistance.

3. Self-reported level of video game experience (question 7) and overall matched wrong performance on the multitask. It could be expected that participants with a lot of experience with the fast changing action of video games would be more able to quickly transition between tasks and therefore make fewer matching errors than those with less experience.
4. Self-reported level of typing skill (question 11) and overall total keying errors performance on the multitask. Should not good typists make fewer keyboarding errors than poor typists?
5. Self-reported level of computer skill, and overall delay in handling matching tasks (avg. match age) performance on the multitask. It seems reasonable to expect that participants with good computer skills should be better at processing computer-based interruptions in a timely way than those with poor computer skills.

Not only were these five relationships not significant, but they were totally absent. There are three possible explanations for the missing statistical relationships between self-reported abilities and observed performance on the experimental multitask: (a) none of the skill topics measured in the entrance questionnaire are relevant predictors of peoples' performance on the multitask used in this experiment; (b) the entrance questionnaire constructed for this experiment was not a good measure of participants' self-perceptions; or (c) participants are not able to accurately report their true levels of experience, skill, and vulnerability.

B3. Subjective Effects

Each participant completed an exit questionnaire (21 questions; see Appendix A2) immediately after they finished all trials of the experimental multitask. These questions were designed to measure three kinds of subjective values: (a) participants' overall anxiety and motivation (questions 18 and 19); (b) participants' opinions about specific methods for coordinating interruption (question 30 was about the scheduled solution; questions 31–35 were about the negotiated solution, and question 36 was about the mediated solution); and (c) participants' relative rankings of the four different coordination methods on various dimensions (questions 20–29).

Participants were asked to report their subjective opinions. This questionnaire was administered to participants before they were debriefed about the purpose of the experiment. Participants' were also not told anything about their actual performance scores. These constraints were an attempt to avoid any possible confounding influence on their subjective responses.

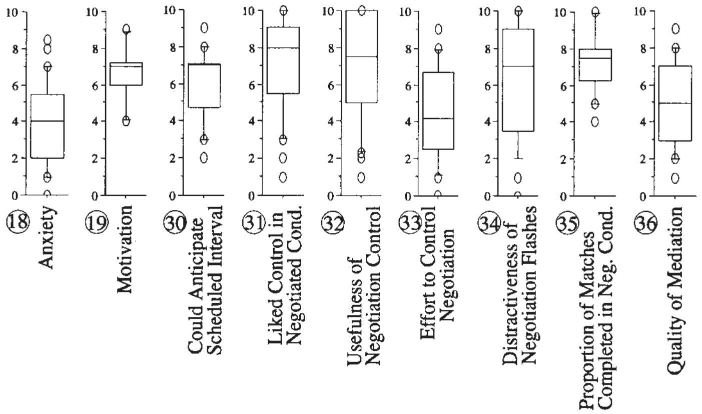
Figure B–10 shows summaries of the results of the exit questionnaire for those questions that asked for a 1-valued answer (questions 18, 19, and 30–36). Participants answered these questions by marking a position on a free form number line from 0 (*least*) to 10 (*most*). See Appendix A2 for the actual questionnaire.

The medians say something about the overall effects, and can be used to make generalizations. However, many of the measures show large variances, and this reveals strong disagreements between participants about those subjective topics.

Figure B–11 shows the average within-subjects rank results of the exit questionnaire for questions 20–29. They show, on average, what method for coordinating interruption participants ranked lowest (a rank of 1), and what they ranked highest (a rank of 4). Note that the relative orderings of the bars are meaningful because each participant gave relative rankings (1, 2, 3, and 4) of all four coordination solutions for each question. The means and error bars (one standard error) are included for each graphic.

Are the differences displayed in Figure B–11 statistically significant? Figure B–12 summarizes the results of an analysis using the Friedman test. For comparison, F_r must be greater than 7.82 for its p value to test above the chosen α level of 0.05 ($F_r > 11.34$ for $\alpha = 0.01$; and $F_r > 16.27$ for $\alpha = 0.001$).

Figure B–10. **Measures of central tendency of participants' reports on part of the exit questionnaire. The boxes in the box plots contain the center 50% of scores with the center line at the median. The outer brackets enclose 80% of values, and extreme 10% outliers displayed as single points above and below. The box plots are numbered with their question numbers on the questionnaire, and scores are reported in subjective units from 0 (*least*) to 10 (*most*).**



The data from 5 of these 10 subjective measures show significant differences. It is concluded that people show consistent opinions about the relative ranking of the different methods for coordinating interruptions for these metrics. These significant results permit post hoc analyses (Siegel & Castellan, 1988, pp. 180–181). Figure B–13 summarizes these post hoc comparisons for the five metrics with significant results in Figure B–12 Each cell reports the results of a significance test with $\alpha = 0.05$. Figure B–11 can be used to determine the direction of significant pairs.

The results show that participants mostly agree on five subjective topics about the relative differences between the four primary methods for coordinating interruptions. These results confirm some intuitive notions about the existence of differences between the four solutions.

People perceive themselves as self-aware. It is therefore plausible that people should report being consciously aware of dynamic changes in the demands on their attention while they perform continuous tasks. Their perception of themselves as self-aware should lead them to claim a degree of sensitivity to their own cognitive processes regardless of whether they actually are.

The negotiated method for coordinating interruption taps directly into this idea of self-awareness of workload. It gives people direct control over when to handle interruptions, and this only works if people can access their internal awareness of their own interruptibility. However, since people perceive them-

Figure B–11. **The average relative ranking of participants' self-reports to questions 20–29 of the exit questionnaire. Responses were ranked from 1 (*least*) to 4 (*most*). Note that questions 20, 21, and 27 asked participants to rank backwards—from 1 (*most*) to 4 (*least*)—but the resulting data were inverted in this analysis for consistency and are represented above in the standard—1 (*least*) to 4 (*most*)—scale. The bar charts show the mean with error bars that depict one standard error. Imm. = immediate; Neg. = negotiated; Med. = mediated; Sch. = scheduled.**

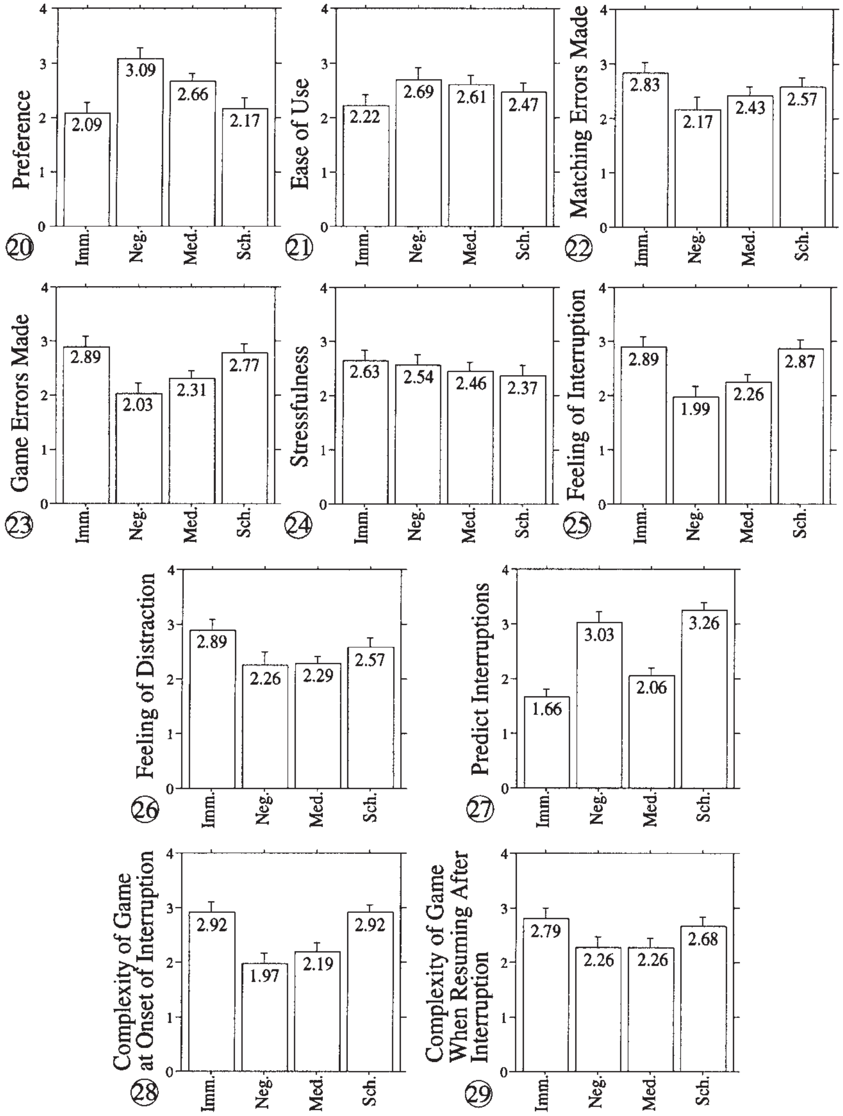


Figure B–12. **Subjective effects by treatment conditions.**

Subjective Measure	F_r	p	$p < \alpha$
Preference (20)	13.594	.0035	Yes
Ease of use (21)	2.767	.4290	No
Matching errors made (22)	4.749	.1912	No
Game errors made (23)	10.063	.0180	Yes
Stressfulness (24)	9.9771	.8563	No
Feeling of interruption (25)	12.851	.0050	Yes
Feeling of distraction (26)	5.434	.1426	No
Predict interruptions (27)	36.943	<.0001	Yes
Complexity of game at start of interruption (28)	15.533	.0014	Yes
Complexity of game at end of interruption (29)	4.659	.1986	No

Figure B–13. **Post hoc analysis of subjective effects.**

Subjective Measure	Immediate and Negotiated	Immediate and Mediated	Immediate and Scheduled	Negotiated and Mediated	Negotiated and Scheduled	Mediated and Scheduled
Preference (20)	Yes	No	No	No	Yes	No
Game errors made (23)	Yes	No	No	No	No	No
Feeling of interruption (25)	Yes	No	No	No	Yes	No
Predict interruptions (27)	Yes	No	Yes	Yes	No	Yes
Complexity of game at start of interruption (28)	Yes	No	No	No	Yes	No

selves as self-aware, they should say they prefer the negotiated solution. The subjective results in Figure B–13 confirm that people agree with this conjecture:

1. Participants said that they preferred the negotiated solution over both the immediate solution and the scheduled solution (question 20).
2. Participants said that they made relatively fewer mistakes on the game task in the negotiated solution (question 23).
3. Participants reported feeling less interrupted in the negotiated condition than they did in either the immediate or scheduled conditions (question 25).
4. Participants said they could predict the onset of interruptions better on the negotiated and scheduled solutions than they could on the immediate or mediated solutions (question 27).

5. Participants said that task switches started at more convenient places (lower complexity of game task) in the negotiated solution than they did in either the immediate or scheduled conditions (question 28).

These agreements between participants on the relative ranking of different solutions are very useful for forming generalizable UI design guidelines (see Section 6). Generalizations, however, are not the only useful result of this analysis. There are some topics, like "stressfulness," where participants had practically no agreement on the relative rankings. Note that this does not mean that participants ranked all four coordination methods as causing a medium amount of stress. Instead, it means that there was a huge disagreement about which solution participants reported as most stressful and which was least stressful.

There are two important classes of disagreement observed in the results of the exit questionnaire: (a) the large variances shown in several of the box plots in Figure B–10 and (b) the nonsignificant rankings shown in Figure B–12 rankings. These disagreements indicate that there are strong individual differences in subjective assessments of the four primary design solutions. Good UI solutions to the interruption problem must, therefore, include some mechanism for individualizing UIs.

Correlation Between Subjective Reports of Single Values and Objective Performance

There were nine questions on the exit questionnaire about participants' state and their opinions of specific coordination solutions (questions 18, 19, and 30–36). A correlation analysis between these nine measures and participants' actual performance revealed no substantial correlation whatever.

A correlation matrix was calculated that compared all 567 pairwise combinations of these nine subjective exit questionnaire topics, and the 63 kinds of performance on the experimental multitask (the nine objective performance metrics with six treatment conditions each [trials combined], plus totals for each excluding base cases). There were no correlation coefficients greater than .5, and only three greater than .4. Most pairs showed virtually no correlation.

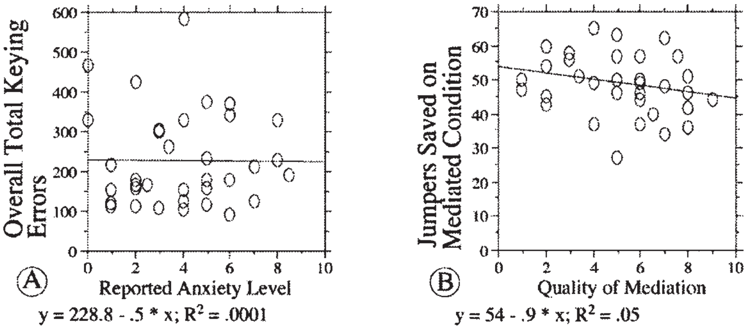
Some relationships may have been reasonably expected, and their conspicuous absence is worth noting. The following are representative examples of pairs of subjective reports and objective performance scores with very little or no correlation (the actual correlation coefficient statistics are reported as *Ccf* = <number>):

1. Level of anxiety (question 18), and overall total keying errors. It seems plausible that people with a high level of anxiety would make more to-

tal keying errors than those with lower anxiety (*CCf*= –.01). Figure B–14A shows that there was no relationship.

2. Level of motivation (question 19) and overall matches not done. A negative relationship seems reasonable. People with more motivation should work harder to have fewer matches not done at the end of trials than others with lower motivation (*CCf*= –.012).
3. Ability to anticipate interval in scheduled condition (question 30) and total keying errors on the scheduled condition. People who can anticipate the transition from game task to matching task on the scheduled condition should make fewer keying errors (*Ccf*= –.266).
4. Amount liked control over task switching in negotiated condition (question 31) and jumpers saved in the negotiated condition. It seems reasonable that people who saved more jumpers in the game task would like the control afforded in the negotiated condition more than those who saved fewer jumpers (*CCf*= .249).
5. Usefulness of direct control over task switching in negotiated condition (question 32) and jumpers saved in the negotiated condition. People who saved more jumpers in the game task should say the negotiation condition is more useful than other people who saved fewer jumpers (*CCf*= .146).
6. Level of extra effort involved in controlling task switching in negotiation condition relative to other conditions (question 33) and number of task switches actually made in the negotiation condition. It seems plausible that people who reported that the task switching was a lot of extra work would have made fewer switches than others who said it was less work (*CCf*= –.174).
7. Distractiveness of announce flashes in negotiation condition (question 34) and jumpers saved in the negotiation condition. People who said that the flash was more distracting should have saved fewer jumpers on the game task than those who were less distracted (*CCf*= –.079).
8. Proportion of matches completed in negotiation condition (question 35) and actual number of matches not done in the negotiated condition. It seems that there should be a clear negative relationship. People should know approximately how many matches they left undone at the end of trials of the negotiation condition, because they saw the announcement flashes (*CCf*= –.293).
9. Accuracy of workload metric in mediated condition (question 36) and jumpers saved on mediated condition. People who say that the mediator accurately predicted good interruption times in the mediation condition should have saved more jumpers on the game task mediation condition than others who said the mediator was less accurate (*CCf*=

Figure B–14. **Scattergrams of correlations between participants' subjective reports on the exit questionnaire of their subjective status and opinions (*x*-axis) and their actual performance on the experimental multitask (*y*-axis). Graph A shows that there is no relationship between reported level of anxiety and participants' overall total keying errors performance (data from base cases not included). Graph B shows that there is virtually no relationship between subjective assessment of the quality of mediation in the mediated condition and participants' actual performance in number of jumpers saved on the mediated condition.**



–.223). Figure B–14B shows only an illusory relationship, and in the wrong direction.

No noteworthy correlations were found. If the exit questionnaire collected valid and reliable data, then there must be no relation between people's internal states and opinions and their actual performance levels. This would mean that people do not form their opinions about UIs based on how well those interfaces help them perform computer-based multitasks. People must base their opinions on something else.

Correlation Between Subjective Rankings and Objective Performance Rankings

There were 10 questions on the exit questionnaire that measured participants' relative rankings of the four different methods for coordinating interruptions (questions 20–29). Was there any relationship between subjective ranks and actual performance? An analysis was performed to calculate the correlations between participants' subjective rankings of the four different coordination methods and their objective rankings of differing performance on the corresponding four experimental conditions.

The Kendall rank-order correlation coefficient with correction for ties was used to calculate the degree of agreement between the subjective and objective ranks of the four methods for coordinating interruptions. The Kendall correlation coefficient is a measure of association similar to the standard correlation analysis, but specifically designed for analyzing ranks. The resulting Kendall Tau with correction for ties (denoted by K_e-T) was calculated for all 90 combinations of the 10 subjective ranks and the 9 performance measures. Figure B–15 shows the results with an indication of whether the resulting K_e-T passes a significance test of $\alpha = 0.05$. The combinations that pass the significance test support the claim that there does exist a relationship between subjective ranking and objective performance.

Many of the significant combinations from Figure B–15 can be visualized by comparing the patterns of average ranks shown in Figure B–11 (subjective ranks from the exit questionnaire) with those shown in Figure 11 (ranks from objective performance metrics). For example, preference ranks (question 20) of the four experimental conditions graphed in Figure B–11 (question 20) has a very similar profile to the jumpers saved ranks graphed in Figure 11A. Kendall Tau's with negative values indicate negative relationships.

The abstract character of the multitask and the tightly controlled design of this experiment limit the generalizability of the results. However, this investigation does have some useful degree of external validity, especially for the class of multitasks that was used as a model for the experimental multitask. The relationships revealed in Figure B–15 between subjective and objective measures can be used to make some informal, but very useful interpretations:

1. People prefer those UIs that allow them to be more effective, efficient, and precise on the continuous task and process interruptions quickly. See the results of the combinations of *preference* (20) with the following: jumpers saved (A); G. key presses per saved (B); avg. match speed (H); and total keying errors (I). The fact that people's preference was not related to efficiency or accuracy on the matching task suggests that people have a natural bias against interruption tasks.
2. People are better at reporting relative ranks of different UI designs than they are at reporting absolute values for isolated opinions about individual UIs. There was significant agreement between subjective matching errors (22) and objective matched wrong (D), and between game errors (23) and jumpers saved (A). Example 8 in the previous subsection (Correlation Between Subjective Reports of Single Values and Objective Performance), however, showed very little agreement between participants in their reported performance and their actual performance on matches not done on the negotiation condition. Participants may not be able to accurately say how well they did on one particular experimental condition, but they are fairly good at ranking the different interfaces according to which allowed them to do best and worst.

Figure B–15. **Correlation analysis between subjective and objective ranks.**

	(A) Jumpers Saved	(B) G. Key Presses Per Saved	(C) Task Switches	(D) Matched Wrong	(E) M. Wrong of Done	(F) Matches Not Done	(G) Average Match Age	(H) Average Match Speed	(I) Total Keying Errors
Preference (20)	0.237 Yes	–0.178 Yes	0.030 No	–0.018 No	–0.007 No	0.025 No	0.063 No	–0.162 Yes	–0.139 Yes
Ease (21)	0.058 No	–0.101 No	–0.064 No	–0.146 Yes	–0.142 Yes	–0.003 No	0.051 No	–0.087 No	–0.064 No
Matching Errors (22)	–0.154 Yes	0.167 Yes	0.062 No	0.192 Yes	0.220 Yes	–0.138 Yes	–0.074 No	0.169 Yes	0.178 Yes
Game Errors (23)	–0.264 Yes	0.191 Yes	0.021 No	0.089 No	0.086 No	–0.064 No	–0.063 No	0.144 Yes	0.177 Yes
Stress (24)	0.009 No	0.061 No	0.054 No	0.060 No	0.055 No	–0.058 No	–0.006 No	0.012 No	0.056 No
Interruptive (25)	–0.231 Yes	0.187 Yes	–0.041 No	–0.013 No	–0.010 No	–0.011 No	–0.066 No	0.137 Yes	0.145 Yes
Distractive (26)	–0.159 Yes	0.063 No	–0.005 No	0.100 No	0.137 Yes	0.028 No	–0.012 No	0.173 Yes	0.095 No
Anticipate (27)	–0.025 No	–0.068 No	–0.378 Yes	–0.079 No	–0.092 No	0.186 Yes	0.365 Yes	–0.456 Yes	–0.387 Yes
Difficulty Before (28)	–0.226 Yes	0.235 Yes	0.008 No	0.057 No	0.027 No	–0.087 No	–0.104 No	0.160 Yes	0.110 No
Difficulty After (29)	–0.090 No	0.083 No	–0.013 No	–0.010 No	–0.021 No	–0.015 No	–0.052 No	0.099 No	0.060 No

3. UI designs that cause people to feel highly interrupted hinder their effectiveness, efficiency, and precision on the continuous task and suppress their ability to process interruptions quickly. See the results of the combinations of *distractive* (26) with the following: jumpers saved (A); G. key presses per saved (B); avg. match speed (H); and total keying errors (I).
4. UI designs that increase people's feeling of distractedness impede their effectiveness on the continuous task and impede their ability to process interruptions accurately and quickly. See the results of the combinations of *distractive* (26) with the following: jumpers saved (A); M. wrong of done (E); and avg. match speed (H).
5. UI designs that increase the predictability of interruptions enable people to process interruption tasks more quickly and make fewer total keying errors than interface designs that do not. See the results of the combination of *anticipate* (27) with the following: avg. match speed (H); and anticipate (27) and total keying errors (I). However, increased predictability also resulted in poor performance in completeness and timeliness on the intermittent task. See the results of combining anticipate (27) with matches not done (F) and average match age (G).
6. UI designs that allow interruptions to be presented at the lull points of the continuous task enable people to be more effective and efficient on the continuous task and to process interruptions quickly. See the results of the combinations of *difficulty before* (28) with the following: jumpers saved (A); G. key presses per saved (B); and avg. match speed (H).

B4. UI Manipulation Errors

Participants made a total of 9,942 keying errors on the experimental trials. That is an average of 276 keying errors per participant, or about 23 keying errors on each of the 432 experimental trials (36 Participants × 6 Conditions × 2 Experimental trials each). These errors are not task errors like the wrong choices people made on the matching task. Instead, keying errors are UI manipulation errors. These kinds of errors happen when participants became disoriented about how to make the computer do what they wanted—like when participants accidentally used the game control keys for when attempting to make a matching task selection. Kirschenbaum, Gray, Ehret, and Miller (1996) also make this distinction between task errors and tool errors (interface usage errors). People make task errors when they succeed in attempting to do the wrong thing, and people make tool errors when they fail in attempting to do the right thing.

Keying errors can be useful indicators of participants' level of confusion and wasted effort as they performed the multitask. These errors are times when participants lost conscious control over their interactions with the computer. The experiment showed that the different UI solutions for coordinating interruptions caused people to make different amounts of keying errors. Some solutions were, therefore, better than others at allowing people to maintain conscious control over their interaction with the computer. A detailed analysis of keying errors was performed to try to discover the causes of these observed differences in control.

The total keying errors metric is a sum of five kinds of keying errors: redundant moves on game task, game keys during matching task, match keys when not matching, illegal negotiation attempts, and unused keys. An analysis revealed that 95.55% of the total keying errors come from only two of the five kinds: redundant moves on game task (72.85%) and game keys during matching task (22.70%). Figure B–16 shows the total keying errors and the breakdown into the five separate kinds of keying errors. Note that Figure 33A appeared before as Figure 9I (Section 5).

A further analysis of redundant moves on game task and game keys during matching task is useful in interpreting the relationship between keying errors and differences in the UI solution for coordinating interruptions. A statistical analysis using the Friedman test shows that there is a significant main effect of the different interruption coordination solutions on the frequencies of these two kinds of keying errors (see Figures B–17 and B–18). Tests of significance are made with an $\alpha = 0.05$. Figure B–16 can be used to determine the direction of significant pairs from the post hoc analyses (see Figures B–19 and B–20).

Differences in the method for coordinating interruptions caused the relative differences in keying error rates. Each coordinating solution created a different interaction context that affected peoples' levels of confusion and interaction efficiency during task switching. The two kinds of keying errors examined here, redundant moves on game task, and game keys during matching task, showed similar results. The immediate solution for coordinating interruptions caused people to make the most keying errors and the negotiated and scheduled solutions caused people to make the fewest keying errors.

The negotiated and scheduled solutions allowed people to stay in control of the UI better than the immediate and mediated solutions. The conditions for this experiment were contrived so that keying errors would not directly affect people's task performance. The input channels for the two tasks of the multitask were isolated by using totally separate groups of keys for each task. However, in some real world systems, it may not be feasible to isolate the input channels and the UI might require people to perform multiple different tasks with the same input devices. In this case, keying errors become a critical problem because UI usage errors can cause serious task errors.

Figure B–16. **The keying errors for the experimental trials. The bar charts show the mean with error bars that depict one standard error. Note that Graph A is the sum of Graphs B, C, D, E, and F. Game Only = the game only no interruption base case; Match Only = the matching-task-only no interruption base case; Imm. = immediate interruption; Neg. = negotiated interruption; Med. = mediated interruption; Sch. = scheduled interruption.**

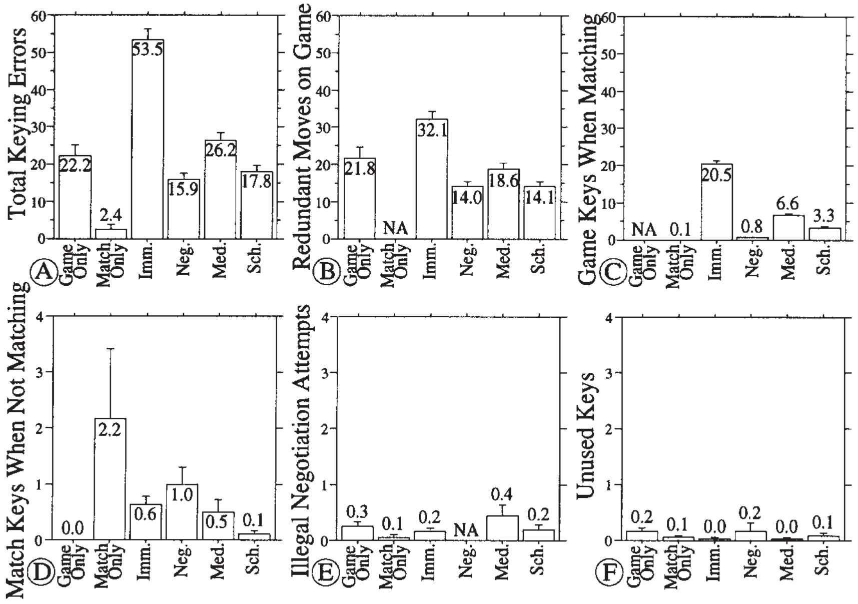


Figure B–17. **Comparison of base cases to experimental conditions.**

Performance Measure	Base Case	F_r	p	$p < \alpha$
Redundant moves on game task	Game only	83.271	<.0001	Yes
Game keys during matching task	Match only	133.530	<.0001	Yes

Figure B–18. **Post hoc comparison to base cases.**

Performance Measure	Base Case	Base and Immediate	Base and Negotiated	Base and Mediated	Base and Scheduled
Redundant moves on game task	Game only	Yes	Yes	No	Yes
Game keys during matching task	Match only	Yes	No	Yes	Yes

Figure B–19. **Comparison of experimental conditions (no base cases).**

Performance Measure	F_r	p	$p < \alpha$
Redundant moves on game task	76.711	<.0001	Yes
Game keys during matching task	97.475	<.0001	Yes

Figure B–20. **Post hoc analysis of main effect.**

Performance Measure	Immediate and Negotiated	Immediate and Mediated	Immediate and Scheduled	Negotiated and Mediated	Negotiated and Scheduled	Mediated and Scheduled
Redundant moves on game task	Yes	Yes	Yes	Yes	No	Yes
Game keys during matching task	Yes	Yes	Yes	Yes	Yes	No