Chapter 17
Collaborative Communication and Intelligent Interruption Systems

Nia Peters, Margaret Ugolini, and Gregory Bowers

Abstract Within collaborative environments, humans are not only tasked with interacting with technology, but also with other humans. The interruption management systems literature is dedicated to alleviating the ill-effects of interruptions specifically within single-user, multitasking interactions by proposing temporal presentations of interruptions in the main task that are least disruptive to the entire interaction. There is less work focused on this concept within multi-user, multitasking environments. In this chapter we propose various temporal presentations of information at low cognitive workloads and evaluate how these timings affect human performance. In measuring objective and subjective individual and team metrics within a dual-user, dual-task paradigm, performance is optimized for low cognitive workload interruption timings compared to high cognitive ones. This work contributes to the overall body of literature by proposing temporal presentations of information within multi-user, multitasking interactions that circumvent the disruptiveness of disturbances in these domains.

Keywords Interruption timing · Interruption management · Intelligent interruption systems

17.1 Introduction

As humans continue to multitask and collaborate among semi-autonomous systems, their conversations and tasks are being interrupted more often. These interruptions are a side effect of technical advancements in general and specifically semi-autonomous

N. Peters (✉)
711th Human Performance Wing, Air Force Research Laboratory, 4164 Seventh Street, Wright Patterson AFB, Greene County, OH 45459, USA
e-mail: nia.peters.1@us.af.mil

M. Ugolini · G. Bowers
Ball Aerospace & Technologies, 2875 Presidential Dr, Fairborn, OH 45324, USA
e-mail: margaret.ugolini.ctr@us.af.mil

G. Bowers
e-mail: gregory.bowers.4.ctr@us.af.mil

technologies which have self-governing capabilities, but also engage with users to achieve task goals. Intelligent software technology can escalate the Human-Computer Interaction (HCI) problem of interruptions by inducing negative effects on human cognition, productivity, affect state, and task performance (Adamczyk & Bailey, 2004) by inundating users with too much information at inconvenient times that do not consider the user's current state. User-interruptions have been studied within the medical domain (e.g., Grundgeiger & Sanderson, 2009), military domain (e.g., Goyal & Fussell, 2017) and (Hodgetts, Tremblay, Vallières, & Vachon, 2015), and commercial domain (e.g., Pradhan, Qiu, Parate, & Kim, 2017, Horvitz, 2001, and Prajapati, Yamada, Unehara, & Suzuki, 2016) to inform intelligent notification systems. The ubiquitous nature of interruptions makes alleviating the ill-effects of this phenomenon a significant area of exploration within human-computer interaction.

Interruption science focuses on how interruptions affect human performance as well as interventions to ameliorate the disruptions caused by them. Although there are many factors that account for the disruptiveness of interruptions, their timing relative to the main task is particularly influential. Research from Gould, Brumby, and Cox (2013), Iqbal and Bailey (2005, 2006), Katidioti, Borst, and Taatgen (2014), and Monk, Boehm-Davis, Mason, and Trafton (2004) suggest disseminating interruptions at times of lower cognitive workloads or at (sub)task boundaries in order to alleviate their disruptiveness.

Since previous research within single-user, multitasking interactions suggests that interruptions within the main task should be sent at periods of lower cognitive workloads and at sub(task) boundaries, we aim to explore whether similar effects are present within multi-user, multitasking interactions. This work is motivated by the limitation of theories and studies dedicated to interruptions in multi-user, multitasking domains such as air traffic control, unmanned aerial systems (UAV) operations, and emergency personnel exercises.

To contextualize this, imagine UAV and ground troop operators collaborating over push-to-talk to identify a target when looking at it from two different perspectives (e.g., the UAV operator has an aerial perspective and the ground troop operator has a first-person perspective). Simultaneously, both collaborators must attend to information in their immediate environment (e.g., UAV operator must monitor changing UAV states). An *interruption* within these interactions can be defined as an unanticipated request for task switching from a person, an object, or an event while collaborating and multitasking. The challenges of single-user interruptions extend to these more complex domains. This extension makes alleviating their ill-effects critical and also more challenging since factors beyond the needs of a single individual who is multitasking have to be extended to multiple users who are also multitasking. The outcomes from this research will not only inform follow-up studies to better understand these relationships, but also motivate the development of theoretical frameworks in this space.

17.2   Interruptions in Multi-user Multitasking Interactions

Similar to Peters, Romigh, Bradley, and Raj (2017b), this chapter evaluates human performance as a function of different temporal presentations of interruptions within the main task specifically within multi-user, multitasking interactions. Peters et al. (2017b) explored the manipulation of interruption timings delivered in the main task (fixed, random, and human determined) and assessed the main and interruption task accuracy and completion times. The results suggest human determined interruptions (a proxy for lower cognitive workload interruptions) significantly improved interruption task performance. Additionally, Peters, Romigh, Bradley, and Raj (2017a) found that 53% of human determined interruptions occurred within 2 s of *(sub)task boundaries* defined as a temporal interval after one task is complete, but before another begins.

Within these interactions, humans are not only multitasking, but also collaborating. The proposal of interruptions at a time of low cognitive workload must be considered for more than one person performing multiple tasks. More formally we can think about interrupting at times of lower cognitive workload as the avoidance of task co-occurrence or *dual-tasking* which is an individual performing two tasks simultaneously. As an example, one reason humans in the Peters et al. (2017a) study may be interrupting at (sub)task boundaries is to prevent dual-tasking.

Compared to single-user multitasking interactions, when considering multiple users, the main task consists of two tasks instead of one: users speaking and listening. There are two implications of dual-tasking in collaborative tasks: (1) if the interruption is intended for the speaker, the speaker must speak to their teammate while listening to the message or stop speaking and listen to the message, and (2) if the interruption message is for the listener, they must now attend to two steams of information. We propose low cognitive workload interruption timings that mitigate dual-tasking in multi-user multitasking interactions. We formally define interruption timings at low cognitive workload as those that minimize the probability of dual-tasking or avoid sending messages when users are either speaking or listening.

17.2.1   Low Cognitive Interruption Timings

Motivated by Adamczyk and Bailey (2004) and Peters et al. (2017a), SUBTASK and KEYWORD are interruption timings that send interruptions after detecting the end of a (sub)task. The SUBTASK timing detects the end of a SUBTASK prior to interrupting and KEYWORD timing detects affirmation cues predictive of (sub)task boundaries. A (sub)task boundary is defined as a temporal interval after a (sub)task is complete, but before another begins. Shivakumar, Bositty, Peters, and Pei (2020) found a lexical category of keywords and phrases called affirmation cues (i.e., got it, copy that, OK I'm done) that are predictive of the occurrence of (sub) task boundary or transition from one (sub)task to the next. We posit that if users are not currently

performing an ongoing task, they are not speaking or listening to content related to that task, and allowing a timing that detects transitions between (sub)tasks can minimize dual-tasking.

SILENCE and PUSH TO TALK OFF (PTT OFF) are interruption timings that send interruptions after detecting the end of a conversational turn. These are novel interruption proposals that could result in an interesting trade space. On the one hand for interactions where (sub)tasks are long and provide less opportunities to interrupt, these strategies may provide opportunities to interrupt by analyzing the task at a lower granularity. Conversely, if conversational turn-taking is moving too fast or the length of the interruption message exceeds the available temporal opportunity to interrupt, dual-tasking may occur. There is no clear indication of the implications for these timings, but because they monitor the ongoing task prior to sending interruptions, we believe they may provide opportunities to minimize dual-tasking.

HUMAN are interruption timings sent by a third human listening to the ongoing task and making decisions on when to interrupt that are least disruptive to the overall interaction. The variability in human decisions and the reality that not all human interruption decisions will optimize overall task efficiency must be acknowledged. For our purposes, since previous literature suggest that more than half of human interruptions occurred at task boundaries (Peters et al. 2017b), this gives us some indication that humans are using strategies to minimize dual-tasking.

17.2.2 High Cognitive Interruption Timings

Motivated by Peters et al. (2017a), the RANDOM FEW and RANDOM MANY are interruption timings that send interruptions at random times in the interaction. RANDOM FEW interruptions are sent less frequently than RANDOM MANY. Both have the potential to increase the probability of dual-tasking because they do not monitor where teammates are in their interaction and can easily interrupt while people are speaking or listening. RANDOM FEW may be less detrimental than RANDOM MANY because it is sending fewer interruptions, inherently minimizing dual-tasking compared to RANDOM MANY. Also motivated by Peters et al. (2017a) FIXED interruption timings are sent at fixed timed intervals. Similar to RANDOM MANY and RANDOM FEW, FIXED interruptions have the potential to increase dual-tasking with little consideration of the ongoing task.

17.3   Methods

Within a dual-user, dual-task scenario, we aim to compare individual and team performance between high cognitive load interruption timings and low cognitive load interruption timings to a baseline condition and evaluate the effect these timings have on human performance metrics. We hypothesize that the single (main) task baseline

condition will provide optimal performance, high cognitive workload interruption times will degrade performance, and low cognitive workload interruption times will be the same as baseline.

We explore the following research questions:

Research Question 1: Is there a difference in *main task team performance* between interruption times at high cognitive workload, low cognitive workload, and the main task baseline condition?

H_{01}: There is no difference in main task team performance between interruption times at high cognitive workload, low cognitive workload, and the main task baseline condition.

H_1: There is a difference in main task team performance between interruption times at high cognitive workload, low cognitive workload, and the main task baseline condition.

Research Question 2: Is there a difference in *individual subjective measures* between interruption times at high cognitive workload, low cognitive workload, and the main task baseline condition?

H_{02}: There is no difference in individual subjective measures between interruption times at high cognitive workload, low cognitive workload, and the primary task single-task baseline conditions.

H_2: There is a difference in subjective metrics between interruption times at high cognitive workload, low cognitive workload, and the primary task single-task baseline conditions.

Research Question 3: Is there a difference in *individual interruption task performance* between interruption times at high cognitive workload and low cognitive workload?

H_{03}: There is no difference in individual interruption task performance between interruption times at high cognitive workload and low cognitive workload.

H_3: There is a difference in individual interruption task performance between interruption times at high cognitive workload and low cognitive workload.

17.3.1 Data Collection

To explore the aforementioned research questions, we simulate a simple multi-user, multitasking interaction; a dual-user, dual-task scenario. The main task simulates a coordination task where teammates must ground their knowledge of a scene from two different perspectives. The secondary task is the interruption task that simulates people having to monitor information independent of the collaborative task.

In our experiment, the main task is a collaborative *Spot the Difference* task and the interruption task is a *UAV keeping-track* task. Users were tasked with performing this dual-task within a 15-min time limit. Participants were instructed to go through a series of *Spot the Difference* tasks and answer as many UAV queries as possible within the allotted time. Additionally, participants were instructed to prioritize both tasks equally. The subjects were from ages 20 to 35, four females and six males. From these 10 participants, we constructed 10 teams, with each participant serving on exactly two teams.

Spot the Difference The main task is a collaborative, computer-system implementation of the *Spot the Difference* task illustrated in Fig. 17.1.

(a) User I - Spot the Difference   (b) User II - Spot the Difference

Fig. 17.1   GUI for the spot the difference task

Two users speak over a push-to-talk interface to identify differences in their pictures. When users identified a difference in their respective pictures and both of the users clicked on that difference, if correct a visual of the difference appeared. Users were also given an indication of how many differences they found in a picture via a scoreboard.

UAV Keeping Track The interruption task was a *Keeping Track of Unmanned Aerial Vehicle (UAV) States* task inspired by Venturincv (1997) where each subject was asked to keep track of three different pieces of information about changing UAV states: name, attribute, and attribute value. An example is:

Raven-3 (UAV name) Fuel (UAV attribute) is 50% (UAV attribute value)

There were 5 UAV names, 5 UAV attributes, and 5 attribute values giving a total of 125 randomly selected changing UAV states that could be sent as interruptions. Once a UAV state was sent, the next interruption prompted the user to repeat what they heard: "Repeat the Previous Statement." An example of the interruption sequence presented to the users (regardless of the interruption timing condition) follows:

Interrupt 1 for User 1: Raven-3 Fuel is 50%
Interrupt 1 for User 2: Raptor-25 Play is Parallel Sweep
Interrupt 2 for User 1: Repeat Previous Statement
Interrupt 2 for User 2: Repeat Previous Statement

This task was completed individually so participants could not hear the interruptions meant for their teammate.

Interruption timings are inherently a function of the interaction. For instance, if the push-to-talk button was pressed in quick succession due to a conversation with rapid turn-taking, the interruptions in a condition associated with pressing the push-to-talk button would also occur in very rapid succession. This situation would make it incredibly difficult to compare this condition to other conditions with more temporally spaced interruptions. To avoid this undesirable co-occurrence, interruptions were only available to be sent once every 15 s.

Users received a synthetic audio stimulus and a persistent text of the interruption message that was present for the same length of time as the audio. Interruption messages were presented in a pop-up window that partially obscured the main task window. Users verbally articulated their response to the query "Repeat Previous Statement" and their response was scored by the experimenter on which pieces of information they answered correctly. The pop-up window was closed when the user responded and pressed the OK button to close the window (Fig. 17.1b).

17.3.2 Conditions

We used a within-team design with the following 9 conditions (1 control; 8 manipulations):

– **MAIN CONTROL**: Spot the Difference Task only. This is the baseline condition.
– **RANDOM FEW**: Dual-task with randomly timed interruptions occurring at longer temporal delays between 0–45 s.
– **RANDOM MANY**: Dual-task with random interruptions occurring at shorter temporal delays between 0–15 s.
– **FIXED**: Dual-task with interruptions sent every 15 s.
– **PUSH TO TALK OFF (PTT OFF)**: Dual-task with interruptions triggered after push-to-talk was released.
– **SILENCE**: Dual-task with interruptions sent when audio energy was below -70 dBFS for 1 s.
– **KEYWORD**: Dual-task with interruptions sent after a keyword spotter detects a predefined set of keywords from affirmation cues (Shivakumar et al., 2020).
– **SUBTASK**: Dual-task with interruptions sent after both users click a difference.
– **HUMAN**: Dual-task with a third human participant listening in and making interruption decisions.

The presentation order of conditions was counterbalanced across teams. Participants were not told which condition they were running. All participants served on a team as well as serving as a human interrupter at least once. The potential interrupter participant was present for at least the beginning of every session, regardless of condition type, to ensure that the "Human" condition set-up procedures were not noticeably different from the other conditions.

Team Performance Measures We used metrics motivated by the single-user, multitasking interruption literature. Since this design is a dual-user, dual-task paradigm, some were more appropriate at the team level and others at the individual level.

The following team-performance measures for the main Spot the Difference task were evaluated:

– **Average Main Task Time of Completion (min)**: Total time for completed pictures divided by the number of completed pictures.

– **Number of Differences Found**: A count of the number of differences found in the Spot the Difference Task within 15 min.
– **Average Time to Find a Difference (s)**: The average time elapsed between finding one difference and the next difference.
– **Average click delay(s)**: The average time elapsed between one participant clicking a difference and their partner also clicking that difference to confirm.

Individual Performance Measures Since the interruption task was an individual task and the partner does not participate in this task, we extracted individual performance measures from the UAV Keeping Track task.

– **Interruption Score**: Number of queries answered with all three attributes correct divided by the total number of queries sent to an individual.
– **Partial Credit Interruption Score**: Number of correct attributes reported divided by the total number of attributes requested (3 per query). For example, if subjects correctly report 2 attributes, they receive a score of 2/3 (66.66%) for that query.
– **Response Duration for Correct Query Response**: Duration of vocal response when answering correctly.
– **Response Time for Correct Query Response**: Time to click the push-to-talk to respond to a query when the response was correct.
– **Percentage of Unanswered Queries**: Number of queries that were unanswered divided by the total number of interruptions sent.

Due to data processing errors, we did not report response time and duration for the Incorrect Query Responses.

Finally, after each run, we gave users the NASA-TLX survey developed by Hart and Staveland (1988) to extract subjective measures. This survey measures Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Participants rated their impression of the runs rating these factors from 1–10 (Low–High). The questions on the survey are:

– **Mental Demand**: How mentally demanding was the task?
– **Physical Demand**: How physically demanding was the task?
– **Temporal Demand**: How hurried or rushed was the pace of the task?
– **Performance**: How successful were you in accomplishing what you were asked to do?
– **Effort**: How hard did you have to work to accomplish your level of performance?
– **Frustration**: How insecure, discouraged, irritated, stressed, and annoyed were you?

For the Silence condition, there is a data point missing from Team 1, so this condition has 18 data points compared to the other conditions with 20 data points.

17.4   Results and Discussion

A one-way analysis of variance (ANOVA) at the 0.05 level was used for our analyses. We hypothesized in measuring team, individual, and subjective performance, the *baseline* condition will be optimal, performance will be degraded in the *high cognitive* conditions, and unchanged from baseline performance in the *low cognitive* conditions. The baseline condition is the MAIN CONTROL; the high cognitive load conditions are RANDOM MANY, RANDOM FEW, AND FIXED; and the low cognitive load conditions are SILENCE, PTT OFF, SUBTASK, KEYWORD, and HUMAN.

17.4.1   Team Performance Analyses

The analyses below will allow us to answer Research Question 1. Although the results are not significant, we do want to report trends that suggest some of the low cognitive conditions being similar to or exceeding baseline conditions. Conversely the high cognitive conditions more often degraded baseline performance.

For the dependent variable **Average Main Task Time of Completion (min)**, the ANOVA was *not significant*, $F(8,81) = 1.129$, $p = 0.353$. Compared to baseline, the worst condition was RANDOM FEW where users took an average 2.3 min longer to complete the main task. Compared to baseline, the best condition was FIXED where on average users took 1.3 s less time to complete the main task. Here, a high cognitive load condition RANDOM FEW degraded baseline performance and a high cognitive load condition FIXED exceeded baseline performance.

For the dependent variable **Number of differences found** the ANOVA was *not significant*, $F(8,81) = 0.758$, $p = 0.640$. Compared to baseline, the worst performance condition was RANDOM MANY where on average users found 6.2 fewer differences. Compared to baseline, the best performing condition was SUBTASK where users found 1.9 more differences. Here, a high cognitive load condition RANDOM MANY degraded baseline performance and a low cognitive load SUBTASK condition exceeded baseline performance.

For the dependent variable **Average time to find a difference(s)**, the ANOVA was *not significant*, $F(8,81) = 1.048$, $p = 0.408$. Compared to the baseline, the worst performance condition was RANDOM FEW where on average users took 11.8 s longer to find differences. Compared to baseline, the best performing condition was SUBTASK where on average users took 5.6 s less to find differences. Again a high cognitive load condition RANDOM FEW degraded baseline performance and a low cognitive load SUBTASK condition exceeded baseline performance.

For the dependent variable **Average Click Difference(s)** or the time difference between when the first person identified a difference and then the other person spotted that difference, the ANOVA was *not significant*, $F(8,81) = 0.520$, $p = 0.838$. Compared to baseline, the worst performance condition was the HUMAN condi-

tion where users took on average 0.5 s longer to click after the first partner found a difference. Compared to baseline, the best performance condition was RANDOM FEW where on average users took only 0.2 s longer to click after their partner finds a difference. Here both low and high cognitive load conditions degraded baseline performance, but a high cognitive load condition RANDOM FEW degraded it to a lower extent than HUMAN, a low cognitive condition.

17.4.2 Individual Subjective Analyses

The below analyses will allow us to answer Research Question 2. Here we will not only report conditions that significantly degrade or exceed baseline performance, but even if a condition is not significantly different from baseline, we will report the extent to which it is different.

For the dependent variable **Mental Demand**, the ANOVA was *significant*, $F(8,169) = 2.230$, $p = 0.028$. A post-hoc Tukey analysis illustrated a mean difference of 4.95 between the RANDOM MANY and MAIN CONTROL conditions, $p_{tukey} = 0.012$ indicating that the high cognitive load RANDOM MANY condition was significantly more mentally demanding than the baseline condition.

For the dependent variable **Physical Demand**, the ANOVA was *not significant*, $F(8,169) = 0.609$, $p = 0.769$. This result is intuitive since there was no expectation for physical demand based on the nature of the task.

For the dependent variable **Temporal Demand**, the ANOVA was *significant*, $F(8,169) = 2.779$, $p = 0.007$. A post-hoc Tukey analysis illustrated a mean difference of 4.95 between the RANDOM MANY and MAIN CONTROL conditions, $p_{tukey} = 0.005$, indicating that the high cognitive load RANDOM MANY condition was significantly more temporally demanding than the baseline condition.

For the dependent variable **Performance**, the ANOVA was *significant*, $F(8,169) = 3.5865$, $p < 0.001$. A post-hoc Tukey analysis illustrated a mean difference of 4.2 between the FIXED and MAIN CONTROL conditions, $p_{tukey} = 0.030$ and mean difference of 6.65 between the RANDOM MANY and MAIN CONTROL conditions, $p_{tukey} < 0.001$. These results indicate users perceived their performance was significantly worse in the two high cognitive load FIXED and RANDOM MANY conditions compared to baseline.

For the dependent variable **Effort**, the ANOVA was *not significant* at the 0.05 level, $F(8,169) = 1.707$, $p = 0.1$. A post-hoc Tukey analysis illustrated a mean difference of 4.55 between the RANDOM MANY and MAIN CONTROL conditions, $p_{tukey} = 0.057$, indicating that on average subjects were scoring their effort on the RANDOM MANY condition task 4.55 points higher than the baseline. These results indicate users expended more effort on the high cognitive load RANDOM MANY condition compared to baseline.

For the dependent variable **Frustration**, the ANOVA was *not significant* at the 0.05 level, $F(8,169) = 1.94$, $p = 0.057$. A post-hoc Tukey analysis illustrated a mean difference of 4.65 between the FIXED and MAIN CONTROL conditions, $p_{tukey} =$

0.070. Additionally there was a mean difference of 4.45 between the RANDOM MANY and MAIN CONTROL conditions, $p_{tukey} = 0.098$. These results indicate that users were more frustrated in the two high cognitive load FIXED and RANDOM MANY conditions compared to baseline.

17.4.3   Individual Interruption Task Measures

The below analyses will allow us to answer Research Question 3. Although none of the results are significant, we aim to illustrate the extent to which the performance metrics of select low cognitive load conditions are different from high cognitive load conditions.

For the dependent variable **Interruption Score**, the ANOVA was *not significant*, $F(7,152) = 0.74$, $p = 0.740$. Across all conditions the **Interruption Score** was $\mu = 68.8\%, \sigma = 24.5\%$. The highest score was from the low cognitive load condition HUMAN ($\mu = 72.4\%, \sigma = 22.2\%$), and the lowest was from the high cognitive load condition RANDOM MANY ($\mu = 59.8\%, \sigma = 25.5\%$) with a 12.5% difference between the two.

For the dependent variable **Partial Credit Interruption Score**, the ANOVA was *not significant* at the 0.05 level, $F(7,152) = 0.511, p = 0.825$. Across all conditions the **Partial Credit Interruption Score** was $\mu = 79.9\%, \sigma = 20.6\%$. The highest score was from the low cognitive load condition SILENCE ($\mu = 82.7\%, \sigma = 16.6\%$), and the lowest was from the high cognitive load condition RANDOM MANY ($\mu = 72.6\%, \sigma = 23.8\%$) with a 10.1% difference between the two.

For the dependent variable **Avg Response Duration for Correct Query Response(s)**, the ANOVA was *not significant*, $F(7,152) = 0.76$, $p = 0.622$. Across all conditions, the **Avg Response Duration for Correct Query Response(s)** was $\mu = 3.18, \sigma = 0.64$. The shortest duration was from the low cognitive load condition SUBTASK ($\mu = 2.95, \sigma = 0.63$), and the longest duration from the high cognitive load condition RANDOM MANY ($\mu = 3.355, \sigma = .43$) with a 0.4 difference between the two.

For the dependent variable **Percentage of Non Responses**, the ANOVA was *not significant* at the 0.05 level, $F(7,152) = 0.41$. Across all conditions, the **Percentage of Non Responses** was $\mu = 9\%, \sigma = 17.6\%$. The lowest percentage was from the low cognitive load condition KEYWORD ($\mu = 5.3\%, \sigma = 11.9\%$), and the largest percentage was from the high cognitive load condition RANDOM MANY ($\mu = 14\%, \sigma = 23.2\%$) with an 8% difference between the two.

For the dependent variables, **Correct Interruption Response Time**, the ANOVA was *not significant* at the 0.05 level, $F(7,152) = 1.511, p = 0.167$. Across all conditions, the **Correct Interruption Response Time** was $\mu = 2.45, \sigma = 0.69$. The shortest duration was from a low cognitive load condition SILENCE ($\mu = 2.19, \sigma = 0.41$) and the longest duration from a high cognitive load condition RANDOM MANY ($\mu = 2.68, \sigma = .67$) with a 0.49 difference between the two.

17.5 Discussion

For Research Question 1, we can accept the null hypothesis. Although none of the results were significant. Other than the dependent variables **Average Main Task Completion** and **Average Click Difference**, there was a common trend in low cognitive load conditions exceeding or being comparable to baseline performance and high cognitive load conditions degrading baseline performance. This finding is a promising result because it gives some indication that low cognitive load interruptions will not induce the negative effects of interruption timings we have seen in the previous literature (e.g., Adamczyk & Bailey, 2004). Additionally, we found that for dependent variables such as **Number of differences found** and **Average time to find a difference**, the low cognitive load condition SUBTASK actually exceeded baseline condition performance. It is possible that interruption tasks with low cognitive load interruption timings actually increase motivation to allocate more effort to the primary task when interruptions were not present.

Variability in team differences and a sample size of only 10 teams makes it difficult to draw any strong conclusions in relating the interruption timings to main task performance. An expansion of this study and carefully minimizing team-performance variability in the primary task will better allow us to make stronger inferences from results in similar paradigms.

For Research Question 2, we can partially reject the null hypothesis specifically for the dependent variables **Mental Demand** and **Temporal Demand**; where the high cognitive load RANDOM MANY condition was significantly different from the baseline; and for the **Performance** variable, where two high cognitive load conditions RANDOM MANY and FIXED were significantly different from the baseline. These results corroborate similar results from Adamczyk and Bailey (2004) and illustrate how random interruptions negatively influence affect states or the emotional component of completing these tasks. As we hypothesized, across all subjective metrics, none of the low cognitive load conditions were significantly different from the baseline. Finally there was a trend of low cognitive load conditions such as SUBTASK, KEYWORD, HUMAN, and SILENCE and the high cognitive load condition RANDOM FEW having subjective metrics comparable to baseline. The most interesting part of this result is that one of the high cognitive load conditions (RANDOM FEW) was similar to the baseline based on subjective rating. This finding could give some indication that this condition is more comparable to low cognitive load interruption timings especially when measuring subjectivity.

For Research Question 3, we can accept the null hypothesis. Although across all the dependent variables none were significant, we did find a pattern of best performance coming from low cognitive load conditions and the worst coming from a high cognitive load condition (mainly RANDOM MANY). This corroborates findings from Peters et al. (2017a) which indicated random interruption timings significantly degraded interruption task performance. The present study extends this work by evaluating more interruption task metrics to capture the implication of interruption timings on an interruption task.

17.6   Conclusion

Motivated by the previous literature, we proposed several low cognitive load interruption timings, and then evaluated individual and team performance and subjective measures to gauge how disruptive these proposed timings were in multi-user, multitasking interactions. Our results showed not only that for the most part, lower cognitive load interruption timings degrade baseline performance to a lesser extent than high cognitive load interruption timings, but also in some instances, low cognitive load interruptions may even exceed baseline performance.

Limitations of the study include, but are not limited to, the performance variability within the main task making it difficult to make strong inferences about how interruption timings may degrade main task performance. Additionally, with only 10 teams, there is an opportunity to expand the sample size and increase the power of our study. In future work, we aim to address both of these constraints.

The outcomes from this research will not only inform follow-up studies to better understand the relationship between interruption timings and human performance, but also motivate the development of theories and algorithmic solutions to developing interruption management systems that temporally predict times to disseminate information that are least disruptive to the overall exchange.

References

Adamczyk, P. D., & Bailey, B. P. (2004). If not now, when? the effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 271–278).

Gould, S. J., Brumby, D. P., & Cox, A. L. (2013). What does it mean for an interruption to be relevant? an investigation of relevance as a memory effect. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 57, pp. 149–153). Los Angeles, CA: SAGE Publications Sage CA.

Goyal, N., & Fussell, S. R. (2017). Intelligent interruption management using electro dermal activity based physiological sensor for collaborative sensemaking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*(3), 1–21.

Grundgeiger, T., & Sanderson, P. (2009). Interruptions in healthcare: theoretical views. *International Journal of Medical Informatics, 78*(5), 293–307.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in Psychology* (Vol. 52, pp. 139–183). Elsevier.

Hodgetts, H. M., Tremblay, S., Vallières, B. R., & Vachon, F. (2015). Decision support and vulnerability to interruption in a dynamic multitasking environment. *International Journal of Human-Computer Studies, 79*, 106–117.

Horvitz E. C. M. C. E. (2001). Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *Human-Computer Interaction: INTERACT* (Vol. 1, p. 263)

Iqbal, S. T., & Bailey, B. P. (2005). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems* (pp. 1489–1492)

Iqbal, S. T., & Bailey, B. P. (2006). Leveraging characteristics of task structure to predict the cost of interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 741–750)

Katidioti, I., Borst, J. P., & Taatgen, N. A. (2014). What happens when we switch tasks: Pupil dilation in multitasking. *Journal of Experimental Psychology: Applied, 20*(4), 380.

Monk, C. A., Boehm-Davis, D. A., Mason, G., & Trafton, J. G. (2004). Recovering from interruptions: Implications for driver distraction research. *Human Factors, 46*(4), 650–663.

Peters, N., Romigh, G., Bradley, G., & Raj, B. (2017a). A comparative analysis of human-mediated and system-mediated interruptions for multi-user, multitasking interactions. In *International Conference on Applied Human Factors and Ergonomics* (pp. 339–347). Springer.

Peters, N., Romigh, G., Bradley, G., & Raj, B. (2017b) When to interrupt: A comparative analysis of interruption timings within collaborative communication tasks. In *Advances in Human Factors and System Interactions* (pp. 177–187). Springer.

Pradhan, S., Qiu, L., Parate, A., & Kim, K. H. (2017). Understanding and managing notifications. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications* (pp. 1–9). IEEE.

Prajapati, S., Yamada, K., Unehara, M., & Suzuki, I. (2016). Interruption-information management framework for chat interface. In *2016 Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS)* (pp. 631–636). IEEE.

Shivakumar, A., Bositty, A., Peters, N. S., & Pei, Y. (2020). Real-time interruption management system for efficient distributed collaboration in multi-tasking environments. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1), 1–23.

Venturincv, M. (1997). Interference and information organization in keeping track of continually changing information. *Human Factors, 39*(4), 532–539.