Investigating the Cognitive Load – Productivity Tradeoff in Multitasking

by

Maximilian Alexander Chis

Submitted to the Graduate Faculty of

the School of Computing and Information in partial fulfillment

of the requirements for the degree of

Master of Science in Information Science

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This thesis was presented

by

Maximilian Alexander Chis

It was defended on

December 7th 2022

and approved by

Michael Lewis, PhD, Department of Informatics and Networked Systems

Na Du, PhD, Department of Informatics and Networked Systems

Christian Lebiere, PhD, Psychology Department, Carnegie Mellon University

Investigating the Cognitive Load – Productivity Tradeoff in Multitasking

Maximilian Alexander Chis, M.S.

University of Pittsburgh, 2023

Although multitasking is considered solely to negatively impact performance, the majority of analyses of multitasking have been doing in experimental settings where the tasks analyzed are of a non-interdependent nature. I argue that multitasking as it occurs in the workforce is a highly complex task which balances multiple competing needs of logistical efficiency, cognitive load, and minimizaton of idle time. I examine this model of multitasking in the context of study 3 of the Artificial Social Intelligence for Successful Teams (ASIST) program, funded by the Defense Advanced Research Project Agency (DARPA). Though the study was not designed with multitasking in mind and thus has a number of confounds, enough evidence exists to suggest a number of probable reasons for the persistence of multitasking in modern life and the multiple "task axes" on which multitasking can occur.

Table of Contents

List of Tables

List of Figures

Preface

I would like to thank Professor Michael Lewis, Professor Christian Lebiere, and Professor Na Du for agreeing to be on my thesis committee and guide me in the proper form and direction of my thesis.

In particular, I wish to thank my advisor, Michael Lewis, for the many hours and countless emails sent back-and-forth as he helped me to iron out the focus and structure of my thesis, as well as assisting me in catching and correcting errors. I deeply appreciate the patience you afforded me, and I hope this thesis is a reflection of that.

I additionally want to thank Professor Christian Lebiere for our conversations regarding the ACT-R model and its utilization in the ASIST dataset, as well as to him and his team at CMU for incorporating the Cognitive Load and Probability of Forgetting modules into the ASIST Metadata.

I would like to thank Noel Chen for assisting me early on in retrieving and processing the ASIST metadata, and for kindly allowing me to bounce my natal ideas about phases off her. I may not have had the time to uncover this otherwise.

I would additionally like to acknowledge Professor Katia Sycara, Dana Hughes, and the rest of the CMU-TA1 team for providing me feedback on my initial phase research, encouraging me to dive deeper into it, and allowing me the time to pick the idea apart further.

I gratefully acknowledge the United States Defense Advanced Research Projects Agency (DARPA), for funding the ASIST program which developed the data I utilized.

I acknowledge and thank all members of the ASIST program. Without their work in developing and running the Study 3 experiments, this thesis would not exist.

On a more personal note, I would like to thank my friends and loved ones, who for privacy's sake I will not name but who know who they are. The foundations of support and care you provided for me helped bring me to the point where I was able and willing to tackle a thesis, which I likely would not have even contemplated a few years prior. In between the 1's and 0's of the data in this Thesis are parts of you as well.

1.0 Introduction

The popular scientific consensus is that multitasking negatively impacts performance [22, 23, 31, 32]. However, a survey of existing research literature shows a more complex situation. Studies demonstrating an unequivocally negative relationship between multitasking and performance exist alongside studies which find multitasking has a positive relationship with performance under qualified conditions (sometimes these two contradictory conclusions exist in different research studies investigating the same conditions, such as emergency rooms) [35, 18]. These different conclusions can be explained in part by how multitasking is evaluated under different definitions of what constitutes multitasking, in different conditions, with the impacts measured using different metrics. In particular, many experimental evaluations of multitasking rely on the performance of two completely unrelated tasks, where the outcome of one has no influence on the other [17, 26, 2], rather than the more interdependent tasks common in modern workplaces, which often cannot be completed in a single sitting [15].

I argue that multitasking is a necessary component of a modern workplace where tasks often cannot be completed by a single individual, and must instead be passed back and forth between multiple parties before completion. In this environment, the individual must take on multiple tasks or else risk wasting time waiting for intermediate processes to be completed by others. This scenario nonetheless suffers the drawbacks of multitasking, but these drawbacks do not outweigh the gains in productivity until enough tasks are managed that the overall burden on working memory increases the rate of error to make taking on additional tasks reduce overall productivity.

I investigated multitasking within the specific context of a three-subtask configuration, where upon completion of the first subtask, a party must give results of the first subtask to a second party who must complete the second task and then returns those results to the original party for the completion of a third subtask. The critical component of this context is that information of the results of the first subtask must be retained within working memory and utilized within the third task, and the third task cannot be performed immediately after the first task – there is a period where the first party must perform some other action or else do nothing at all.

I examined how the degree of interleaving of these three-subtask task instances impacts overall performance and rate of error. I performed this examination through analysis of experimental trial results as well as ACT-R computational metrics taken during those trials which are derived from the ACT-R cognitive architecture [3].

I investigated this by utilizing a subset of the data accrued from Study 3 of DARPA's Artificial Social Intelligence for Successful Teams (ASIST) program. Research has previously been published based on the ASIST program, examining modifications to the program during the COVID-19 pandemic [19], discussing its potential for near-future artificial social intelligence [6], and developing neural heuristics for route optimization [34].

I then discuss the implications of my findings in the overarching context of how multitasking relates to performance.

2.0 Literature Review

Different studies have come to different conclusions about multitasking, with some unequivocally viewing it as degrading performance, and others with more nuanced perspective. The following literature review examines

- The typically observed costs to multitasking
- Discussions of the prevalence of multitasking
- The limited scope of multitasking research
- And how observational studies produce more variable conclusions about multitasking, compared to experimental studies.

2.1 Multitasking Costs

Plentiful research exists on the cost of multitasking. These costs are typically described as either

- Switch costs, where response time is longer and error rate higher immediately after switching tasks
- Mixing costs, where performance recovers partially after the initial drop in performance but remains degraded if task-switching continues [22].

Numerous models for these costs have been proposed. Among them (in no way comprehensive) are the following

- That multitasking involves two executive processing stages, one goal shifting stage where goals are inserted and deleted from working memory as needed, and one rule activation stage where the previous task's rules are "unloaded" from working memory, and the new task rules are "loaded"[29].
- That two or more simultaneous tasks compete for the same cognitive resources in such a way that a "bottleneck" results leading to diminished performance [23].

- That cognition is “threaded,” where some tasks can be performed in parallel but others, as in the bottleneck theory above, must be competed one at a time, leading to delays [31]
- That these costs result from cognitive resources being devoted to “loading” and “unloading” information from working memory [32] .

Some of these costs have been found to persist whether the task switch was voluntary or involuntary [5].

2.2 Why is Multitasking So Prevalent?

A major contributor to modern interest in multitasking relates to the rise of personal computer and mobile phone technology enabling media multitasking [27]. Beuckels et al., in a literature review of media multitasking, identify an explosive growth in the number of media multitasking publications occurring in the 2011-2019 period [7]. However, my focus is not on media multitasking, which involves the simultaneous use of multiple media sources. Rather, my focus is on workplace multitasking, which involves frequent switching between multiple tasks which, though occurring in close chronological proximity with one another, nonetheless occur (generally) one at a time, though they may be switched between at high frequency.

The prevalence of multitasking in modern work is in part due to the unique needs of modern work. “In Multitask Learning and the Reorganization of Work: From Tayloristic to Holistic Organization,” Lindbeck et al. identify the following contributors to a change in modern work

- Computerized information and communication systems and technology, enabling greater inter-employee communication and information-sharing, as well as more rapid feedback from customers
- A resulting “decentralization of decision making within firms,” where individual employees now must exercise more individual discretion in task management, and less often

defer to managers.

- An increase in teamwork and job rotation
- Increase in "human capital," with a more highly educated and trained workforce with both a deeper knowledge of skills as well as knowledge of a wider variety of skills
- Relatedly, increase in worker preference for jobs of a more complex nature, permitting the exercise of that wider variety of skills [20].

In "'Constant, Constant Multi-tasking Craziness': Managing Multiple Working Spheres," Gonzalez et al. describe field observations of information workers and observe that such workers were often tasked with multiple concurrent projects, and often had to not only work on these projects but also respond to requests for information or assistance from other employees on those or related projects. Gonzalez et al. argue that modern workers thus organize their work into "working spheres," or a higher-level work unit composed of thematically related tasks, varying in their urgency and importance to the individual. These working spheres themselves had to be organized and managed via "metawork," such as in scheduling meetings and organizing communications about work [15].

2.3 Limited Scope of Multitasking Research

The applicability of multitasking research is hampered by a general focus in experimental conditions on independent tasks of a simplistic nature, which do not correlate to the complex circumstances in which multitasking arises in modern life. Burgess, P.W. notes that many everyday multitasking situations tend to contain the following attributes:

1. Numerous tasks: several discrete and different tasks must be completed
2. One task at a time: due to physical or cognitive constraints, it is not possible to perform one task at a time
3. Interleaving required: Performance on these tasks must be dovetailed; the most time-effective course of action is not to finish one task before moving to another, but to switch between them as appropriate.

4. Delayed intentions: The time for a switch or return to a task is not signaled directly by the situation [9].

Burgess additionally states that, while there have been advances in "understanding many situations which have some relevant to aspects of multitasking (e.g., dual- or multiple-task paradigms, task switching, etc.), more complex situations... have been rarely studied within an experimental psychology or cognitive neuroscience framework" [9].

Additionally, multitasking experiments tend to examine performance on multiple tasks where information obtained from one has no bearing on the other. In "The Cost of a Voluntary Task Switch", for example, participants were presented a series of single digits and had to judge either whether the digits were odd or even or whether the digit was larger or smaller than five and were instructed to switch between these tasks equally [5]. Determining whether a digit is odd or even provides little to no relevance for determining whether the digit was larger or smaller than five, and vice-versa.

Similarly, Thomas Buser and Noemi Peter aimed to conduct an experiment with aim to "investigate the type of multitasking which occurs in a modern work environment where employees switch between several demanding and ongoing tasks" [10]. However, the tasks they selected to model this were word search and Sudoku. While an argument can be made that such cognitively demanding tasks can put one "in the mindset" for other cognitive tasks of a similar nature, on a concrete level the participant obtains no information from Sudoku that can be utilized in word search, and vice versa.

In experimental settings, research discussing the benefits of multitasking is also limited in scope. Common discussed benefits of multitasking include relieving boredom [26], or creativity on subsequent tasks [17], the latter suggesting that multitasking increases creativity by "providing activating resources that then stimulate cognitive flexibility ... the ability with which individuals can attend to divergent perspectives". Perception of multitasking may also impact performance – in one study, telling participants that listening to and transcribing a video constituted multitasking (as opposed to being two elements of the same single task) subsequently increased productivity [30]. Adler et al. suggest that the U-shaped increases in performance they found are related to the Yerkes-Dodson law, where an optimal amount of arousal leads to superior performance compared to arousal either too low or too high. [2].

In the above cases, the discussed benefits of multitasking involve either

- The mental state of the participant during the multitasking activity
- Their performance on a subsequent task not involving multitasking

And again, in three of the four examples above, all tasks were independent:

- One task pair in Kapadia et al.'s publication involved suggesting new ways to fund students' organizations and replying to e-mail messages about their work schedules
- One task pair in Ralph et al. involved solving a 2-back task while an unrelated video was playing.
- One task pair in Adler et al. involved completing a sudoku puzzle and filling in the missing number in a series.

The exception was Srna et al., which chose a scenario that was deliberately ambiguous as to whether it was two tasks performed simultaneously, or one task. That scenario points to an alternative, rarely-examined form of multitasking – where the two tasks are not independent, but instead interdependent subtasks of a single overarching task. While the two subtasks in Srna et al.'s study are so tightly coupled that whether performing the two tasks constitutes multitasking is an open question, other interdependent subtasks, to be discussed shortly, have less ambiguity in this regard.

2.4 Variable Multitasking-Related Conclusions in Observational Studies

Research suggesting benefits of multitasking to the tasks being multitasked themselves are more often found in observational studies and surveys of real-life data. In "Information, Technology and Information Worker Productivity," for example, Aral et al. utilize a combination of survey, interview, and accounting data to suggest that "multitasking is associated with more project output, but with diminishing marginal returns." [4]. In "Does Multitasking Improve Performance", Diwas Singh KC collected "patient flow and clinical data from an East Coast metropolitan hospital's emergency department" and found a U-shaped relationship between multitasking and productivity, with too little and too much multitasking yielding reduced productivity compared to a moderate amount of multitasking [18].

In these observational studies, benefits of multitasking are discussed yet less frequently cite other studies compared to their costs. Nonetheless, Aral et al. and KC discuss the following possible benefits to multitasking

- Increased worker utilization reducing nonproductive idle time [4, 18]
- Utilizing information and knowledge from one task in other tasks [4]

Observational studies do not uniformly suggest benefits to multitasking. In another observational study of emergency room physicians Westbrook et al. found that "Interruptions, multitasking and poor sleep were associated with significantly increased rates of prescribing errors among emergency physicians" [35].

Multitasking effects on performance depend in part also on metric used, with Adler et. al finding the previously aforementioned U-shaped curve between multitasking and productivity but also that accuracy is inversely correlated with multitasking [2].

The discrepancy between the results of experimental and observational data suggest that some real-world scenarios contain components that enables multitasking to have more positive effect, and that these components are rarely studied. Indeed, Diwas states that "there are few studies that have examined the operational effects of multitasking using transactional data from a field study, involving activities over longer periods of time" [18].

3.0 Argument: The Modern Workplace Requires and Benefits from Multitasking

Modern workplaces are indeed defined by rapidly alternating between many tasks, but part of the reason for this rapid alternating is because the results of completed subtasks are then passed on to other people, who complete their own subtask and then return the result. The original person does not and cannot follow the task through every subtask – some subtasks must be completed by other people with different knowledge, different capabilities, and/or different perspectives. In such situations, the original person can at most only observe the subtask being performed by someone else and cannot make additional contributions until the subtask is completed. In such contexts, the original person would waste their time if they simply waited for the other person to complete the subtask; they are able to accomplish more if they perform a different task while waiting for the first task to complete, even though this means they are now dividing their attention. However, the person then faces a dilemma in terms of how many such tasks they can reliably balance before incurring penalties that outweigh the gains in productivity.

Consider the following three modern workplace scenarios:

1. A waiter, who must balance multiple orders at a time, passing them on to a cook, who then returns the results to the waiter to pass on to the customers. Having one waiter per table is obviously inefficient, but logistical and cognitive considerations occur the more tables a single waiter is burdened with – not only can the waiter reach their physical limits when it comes to how many they can serve at a time, but they also run the risk of making errors the more tables they have to keep track of, such as mixing up orders between tables.
2. A tech support specialist, who often must balance multiple tickets at a time, and where they often have to reach out to other people to resolve some issue, who can themselves take a variable amount of time to perform some task. Again, there may be a limit to how much they can physically accomplish in a frame of time, but there is also a cognitive limit which may make itself known even before reaching that physical threshold.

3. A manager, who must field and delegate multiple tasks to multiple people. One manager per task defeats the purpose of a manager, but too many tasks and a manager may struggle to keep track of them, even if they physically have enough time to respond to and communicate regarding each of the tasks.

These operate on different time frames – a waitress has to task-switch multiple times within a few minutes, while a tech support agent and particularly a manager may be able to task-switch less frequently – but in all of these cases, these workers are in positions where they can increase their overall productivity by managing more tasks, but they eventually reach a cognitive limit, sometimes before they reach a physical limit to the number of tasks they can manage. Multitasking is necessitated by these situations, but the marginal benefit diminishes over time, until eventually managing an additional task hampers performance more than it increases overall productivity.

This is a more complicated dynamic than what multitasking literature typically focuses on. Multitasking literature typically investigates situations where a task can be performed to completion, not one where a task must necessarily be handed off to another party and where a participant can accumulate multiple such tasks at a time. In such situations, task-switching is an inevitable cost, but one that can be outweighed, to a point, by the benefits of managing additional tasks. Furthermore, many of these tasks may be in a state where they require similar task-relevant knowledge or where it is more physically convenient to perform these similar tasks together than to alternate between these and dissimilar tasks (for example, a waiter fielding multiple orders at once, rather than moving between the front and back of the restaurant multiple times), and thus which tasks to switch between becomes a relevant concern as well.

3.1 A-B-C Task Model

I thus propose a task model to illustrate the various effects of multitasking. This model does not apply to all tasks, but does apply, as I argue, to at least some workplace tasks.

In this model, a **task** is defined as a sequence of actions which must be performed in specific order. A task **type** T encompasses all tasks which follow a similar pattern and which require a similar set of knowledge and skills. A task **instance** I_T refers to a specific performance of the task type T which relies on information relevant to that instance but no other instances.

To use the waiter example, a task **type** would be the ordering, preparation, and delivery of meals to a customer. A task **instance** would be the order requisition, preparation, and delivery of meals to a specific customer, with instance-specific information including the location of the customer in the restaurant as well as what food was ordered.

Two parties are involved in the performance of this task: The primary party, P_1, which oversees the beginning and end of the task; and the secondary party, P_2 which performs an intermediate subtask (to be defined shortly) which the primary party cannot.

Tasks in this model are further divided into three *subtasks*:

- **Task A**: The initiating subtask, performed by P_1. This introduces the task instance and involves the obtaining of instance-relevant information, which is then passed on to P_2 for Task B.
- **Task B**: The intermediate subtask, performed by P_2. Using relevant information obtained from P_1 in Task A, P_2 performs a sequence of actions which P_1 cannot participate in. The output of this is then returned to P_1 for Subtask C.
- **Task C**: The concluding subtask, performed by P_1. P_1 uses the output of subtask B to conclude and resolve the task instance.

The nature of these subtasks means that all task instances must be completed in the order A-B-C. P_2 cannot, for example, perform subtask B and then have P_1 perform subtasks A and C. Subtask B is dependent on information obtained by P_1 in subtask A. A cook, for example, cannot prepare meals for customers unless they know what meals they are preparing for customers.

Furthermore, P_1 and P_2 have skills and resources which allow them to only complete their designated subtasks, or at least incentivizes them to focus on their designated subtasks rather than subtasks which they have less expertise or resources for – A waiter, for example, is not as skilled in cooking (if at all) as the cook. And the cook cannot service customers and maintain the kitchen at the same time, and may be less skilled at servicing customers than the waiter.

Multiple task instances can be active at a time for a given task type: I_{T1}, I_{T2}, I_{T3} and so on.

Furthermore, these instances can be in a different state of completion, such that:

- I_{T1} can be in state A-ready, requiring subtask A to be completed next
- I_{T2} can be in state C-ready, requiring subtask C to be completed next
- I_{T3} can be in state B-ready, requiring subtask B to be completed next
- and so on.

For any subtask of any task instance, two kinds of information are required:

- Subtask-relevant information, S, information on how to perform that subtask, shared across all instances of that subtask.
- Instance-relevant information, I, information relevant to that task instance, shared across every subtask within that instance

Thus, to complete Instance 1 of task type T, subtask A, subtask-relevant information S_{TA} is required, and instance-relevant information, I_{T1}, is required. The total information required for this subtask is thus (S_{TA}, I_{T1}). To complete Task 2, subtask B, in contrast, would require (S_{TB}, I_{T2}).

P_2's role is static as far as these three tasks are concerned. P_2 can focus exclusively on subtask B, and its concern with each of the task instances persists only for as long as it requires to complete subtask B for that instance. P_2 thus always focuses on subtask-relevant information S_{TB}, and the only thing that changes is the Instance relevant information. P_2 can thus proceed from (S_{TB}, I_{T1}) to (S_{TB}, I_{T2}) to (S_{TB}, I_{T3}), and so on.

P_1's role is more dynamic. At any point in time, P_1 can either complete another instance of subtask A, or complete another instance of subtask C. In performing an instance of subtask A, P_1 accrues instance-relevant information that will need to be used in subtask C for that instance, after which it can be discarded without issue. However, because it takes time for P_2 to perform subtask B, P_1 cannot immediately utilize and discard this information after completing subtask A. P_1 could theoretically wait for subtask B to be completed for this instance but depending on the time required for subtask B to be completed, as well as the relative priority of other goals, this may not be practical. To reduce idle time, P_1 can thus either perform another instance of subtask A, accruing additional and separate instance-relevant information in memory; or perform subtask C for a different instance whose subtask B has been completed and is thus now C-ready, utilizing-and-discarding the instance-relevant information for that instance.

Either choice incurs a cognitive cost. If P_1 chooses to not perform another instance of subtask A and instead switch to performing an instance of subtask C, P_1 must unload subtask-relevant information S_A and load subtask-relevant information S_{TB} (P_1 may also need to change aspects of their physical environment, such as their location or the tools they use, but we will ignore that for now). Performance decreases following such a task-switch are expected [5]. Additionally, P_1 must "reload" instance-relevant information for this separate instance of the task, which persists in working memory but in a partially decayed state. However, if P_1 performs subtask A for another instance, they don't have to unload S_{TA} and load different subtask-relevant information, and thus avert that cost, but they still must load and unload instance-relevant information, and more importantly load additional instance-relevant information, which must now exist in working memory alongside the instance-relevant information for other instances, increasing the overall amount of information they need to keep track of. Increased working memory load increases risk of decreased performance and other errors [14].

What choice P_1 ought to make – whether to continue perform subtask A for a new instance, or switch to subtask C for an existing instance – likely depends on a multitude of factors, including the cost of inter vs. intra-subtask switching, the current working memory load of P_1, whether the environment currently incentivizes starting new instances of the task or resolving existing instances, and so on. In either case, a certain amount of multitasking is required because P_1 cannot realize any task instance to completion at a single time without sacrificing overall productivity. P_1 must task switch – the question is what sort of task switch P_1 should do.

P_1's actions can thus be defined according to whether they align with one of two strategies:

- *Instance-focused*, which emphasizes minimizing the amount of task instances active at a time by rapidly switching between subtasks A and C. The most extreme version of this strategy performs subtask C as soon as a task instance is C-ready, and only performs subtask A when no C-ready instances exist.
- *Subtask-focused*, which emphasizes minimizing the amount of subtask switching and thus has a non-minimized number of active task instances. The most extreme version of this strategy performs subtask A until there are no more instances which are A-ready, and only then switches to subtask C, performing this subtask until there are no more instances which are C-ready.

The degree to which one of these strategies is favored over the other can be determined via the number of *phases*, which are here defined as a sequence of performances of the same subtask, across multiple task instances, without interruption by performance of another subtask. An instance-focused strategy thus has a high number of phases, and a subtask-focused strategy has a low number of phases.

Phases alone, however, may not be able to fully capture strategy; the number of phases will naturally increase as more instances are completed, regardless of strategy. I thus introduce an additional derived metric based on the number of phases per completed instance termed the *phase-completion ratio*. An subtask-focused strategy, which minimizes the amount of subtask-switching, will thus have a low phase-completion ratio. In contrast, an instance-focused strategy, which more frequently switches between subtasks, will have a higher phase-completion ratio.

Table 1: Table of strategies and associated markers

Instance-Focused	**Subtask-Focused**
More phases	Fewer phases
Higher phase-completion ratio	Lower phase-completion ratio

This formulation is more complicated than the usual discussed multitasking because it proposes two kinds of information: subtask-relevant information S, and instance-relevant information I. I would argue, however, that this formulation is closer to common workplace environments than the more commonly studied kind of multitasking.

Table 2: Classification of tasks and subtasks of three roles according to A-B-C task paradigm

Example	Task T	Subtask A	Subtask B	Subtask C
Waiter	Complete order	Retrieve order from customer, give to cook	Cook prepares order, gives results to waiter	Waiter returns food to customer
Tech Support Agent	Complete ticket	Retrieve details of ticket, pass on to relevant specialists	Specialist resolves issue, returns results to agent	Agent resolves ticket
Manager	Delegate other tasks	Retrieves details about task to delegate, assigns worker to complete	Worker completes delegated task and reports back to manager	Manager reviews results, communicates to stakeholders

To return to the previous three examples, I categorized the examples according to the above schema. In each instance, there is both subtask-relevant information (for example, for a waiter, how to retrieve orders) as well as instance-relevant information (for a waiter, the details of the order and who it belonged to). Multiple such tasks can be engaged with at a time, but every additional task involves more information which must be tracked, with multiple instances of similar information risking interference.

4.0 Research Goal

I investigated performance on tasks whose configuration was comparable to the configuration described above.

The first subtask A involves identifying the spatial location of objects in an environment and giving information on the location of this object for a second party P_2. P_2's subtask B involves preparing these objects for removal by P_1. Subtask C involves P_1 returning to the locations previously identified and removing them.

Table 3: Applying A-B-C task paradigm to experiment design

Subtask A	P_1 identifies the spatial location of an object, provides information on this object location to P_2
Subtask B	P_2 goes to the object and prepares it for removal by P_1
Subtask C	P_1 returns to the object and removes it.

Instance-relevant information I is thus the location of the object and is what P_1 gives to P_2 in Task A, what P_2 requires in Task B to find the object, and what P_1 requires in task C to return to the object for removal.

While P_1 is waiting for P_2 to complete subtask B_1 in task I_{T1}, they can begin work on another task instance I_{T2}, containing the same subtasks. However, by performing A_2 in I_{T2}, they accrue additional instance-relevant information I_2, which now exists in working memory alongside I_1. Not only can this information degrade, but the similarity of the two sets of information may lead to interference [8]. For example, P_2 might complete B_1, and P_1, intending to go to the object identified by I_1, now prepared for subtask C_1, might instead accidentally go to the object identified by I_2, which has not been prepared.

I hypothesized that

- Each performance of subtask A increases the amount of instance-relevant information in memory

- Each performance of subtask C decreases the amount of instance-relevant information in memory
- More frequent switching between subtasks A and C, by managing the overall amount of instance-relevant information in memory, will be associated with better performance, while less-frequent switching (i.e., by following up performance of subtask A with another instance of subtask A, or C with another instance of C) will be associated with poorer performance.

Of primary interest was whether the degree to which the participant alternates between these tasks had an impact on the following:

1. Overall performance, defined as the total number of objects successfully removed.
2. Error rate, defined as how many objects identified in Task A are also successfully removed in Task C.
3. "Probability of Forgetting," measured according to the ACT-R model of Probability of Forgetting [36]. In this case, the probability that an object in L will not be removed after previously being identified and placed in L.
4. "Cognitive Load," a modification of the above "Probability of Forgetting" model to instead express working memory in terms of number of chunks. In this case, those chunks would correspond to the total number of objects in both L and L'.

These points of focus were examined to answer the following questions:

1. Does the amount of task switching negatively impact overall performance in terms of speed of performance, as is expected by previous literature on task switching [22]?
2. Does the amount of task switching negatively impact accuracy, here defined as the ratio of the number of objects retrieved in subtask C over the number of objects identified in subtask A, as is expected by previous literature on task switching?
3. Does the amount of task switching result in an increase or decrease in cognitive resources, as measured by cognitive load?

5.0 Methodology

Data was obtained from performance of trials run by the DARPA Artificial Social Intelligence for Successful Teams (ASIST) program, a program originally designed to study "agents that demonstrate a machine ToM [Theory of Mind] and the ability to participate in an effective team by representing and helping to maintain shared models" [13, 11]. Data analyzed was from Study 3 of the program.

Study 3 consisted of a series of experimental trials meant to simulate a search-and-rescue mission within a Minecraft environment. In these trials, players performed in teams of three and were divided into three roles with distinct as well as overlapping tasks.

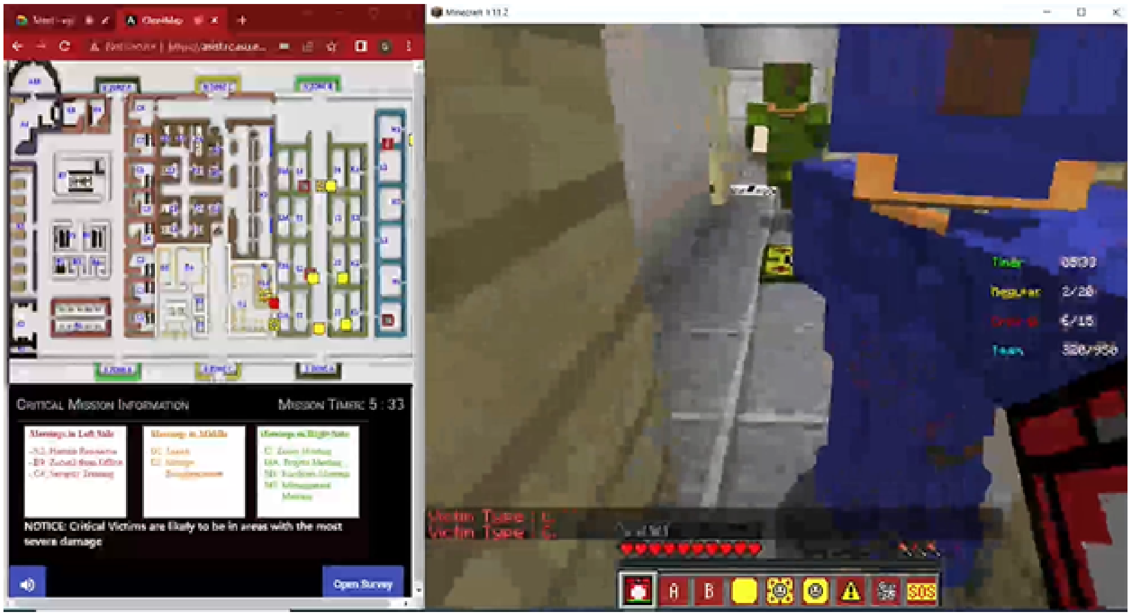
Table 4: Attributes of 3 Roles in Study 3

Role	Movement Speed	Unique Abilities	Primary Purpose
Transporter	Fastest	Detect "signals" indicating presence of victims in rooms	Discover and evacuate victims
Medic	Moderate	Triage wounded victims	Triage discovered victims
Engineer	Slowest	Clear rubble, detect traps	Clear obstacles, assist medic in triaging discovered victims

All players could pick up victims, and both the Engineer and Transporter could assist the medic in triaging critical victims.

Players were placed in an environment simulating a collapsed building complex, and tasked with discovering, triaging, and evacuating victims. Victims were either regular victims, who needed only be triaged by the Medic and were worth 10 points upon evacuation; or critical victims, who needed to be triaged by the Medic and another player (typically the Engineer) and were worth 50 points upon evacuation.

The player interface included both a Minecraft window and a "Client Map", which



NOTICE: Critical Victims are likely to be in areas with the most severe damage

Figure 1: The interface of a player (from Trial 631, Participant E689). Includes the Minecraft interface and Client Map.

included information on rooms which might contain victims, as well as a "Dynamic Map", which included a layout of the complex and was updated in real-time with the position of the player (but not their team members), as well as "marker blocks", objects which could be placed on the ground to communicate information, such as the presence of victims, between players.

The length of missions were 17 minutes in length, with the first 2 minutes devoted to discussion and planning.

Participants engaged in a 10-minute screening session to check their capability to play Minecraft. They then filled out an intake survey whose submission enabled them to sign up for the experiment. Teams of three qualified participants were then scheduled to participate in a 2-hour experiment that involved searching for victims and rescuing them in a Minecraft environment. During that experiment, participants received training videos that introduced the rules of the game and provided hands-on experience with the environment prior to two 17-minute missions. Additional surveys were administered after the training video as well as after each mission. Participants were paid 35$ in Amazon gift cards for their full participation.

Participants were randomly assigned to one of 8 advisor conditions, either one of the six ASI agents, to a human advisor, or to no advisor. Aside from the No Advisor condition to which 15 teams were assigned, 14 teams were assigned to each of the advisor conditions, with each team performing two trials.

168 out of 226 trials were done under the guidance of an Artificial Social Intelligence (ASI). The ASI's were the following:

- Assisting Teamwork via Learning and Advising Systems (ATLAS), from Carnegie Mellon University (CMU)
- Prescient, Socially Intelligent (PSI)-Coach, from Charles River Analytics (CRA)
- Rita, From Dynamic Object Language Labs, Inc. (DOLL)
- SIFT ASIST:ant, from Smart Information Flow Technologies (SIFT)
- Theory of Mind-based Cognitive Architecture for Teams (ToMCAT) from The University of Arizona (UAZ)
- Agents with Theory of Mind for Intelligent Collaboration (ATOMIC), from The University of Southern California (USC)

These agents utilized distinct architectures and offered different interventions to players.

5.1 Participants

ASIST Study 3 recruited 542 participants, required to be physically located in the US, who participated in 232 trials. Participants had an average age of 24. 67% identified as Male, 22% as female, 3% as nonbinary, and 6% as NA/Prefer not to respond.

5.2 Materials

All participants engaged with the experiment remotely, using a combination of the Parsec remote desktop software to connect to the computer running Minecraft, Google chat for talking to the experimenter and other participants, and a web browser for displaying the client map.

All participants received “briefing” materials that addressed the goal of their missions, described attributes of victim types, how to use virtual tools (a hammer for clearing rubble, a Medic kit for triaging victims, and a stretcher for evacuating victims), job aids including a scoreboard and dynamic map (to be described more shortly), and how to interact with the Minecraft environment.

5.3 Data Collection

Trials were performed between April and July 2022.

Data from trials was collected and presented in the form of metadata files, from which relevant information I compiled and organized. Messages included information such as the current location of players, what actions they were performing, what objects appeared in their field of view, and the time during the mission that these events occurred.

5.4 Data Analysis

5.4.1 Phase Definition

A “Phase” is defined here as “a sequence of the same actions performed consecutively without interruption by an action relevant to another Phase.” For the Transporter, I divided their behavior into two Phases – Evacuation (E), and Discovery (D). Discovery Phases were marked by the consecutive performance of one of two actions:

1. A signal indicating the presence of victims in a nearby room, which had previously not been discovered via signal
2. A victim not previously seen entering the Transporter's Field-of-View.

Either or both of these actions could be performed.

Evacuation Phases, by contrast, were marked by the evacuation of players to their respective evacuation zones.

Thus, as an example, a Transporter who sees a new victim would begin a Discovery Phase, and would remain in that Phase for as long as the Transporter continued to find new victims, either via signals or FoV. As soon as the Transporter evacuated a victim to their proper evacuation zone, the Transporter would enter an Evacuation Phase. If the Transporter then discovers a new victim via a signal, they conclude the Evacuation Phase and enter a new, separate Discovery Phase. At this point, the Transporter has three Phases – two Discovery, and one Evacuation.

5.4.2 Phase Count and Phase-Evacuation Ratio

While some analyses were performed using raw Phase count, I more often utilized a ratio of the number of Phases to number of evacuations – what I referred to generically as the Phase-Completion Ratio but here is referred to as the Phase-Evacuation Ratio – as my primary means of analyzing Phases. Using Phase count alone introduced a confound where a team which evacuated more players would almost certainly have more Phases, simply because players tended to evacuate only some players in a Phase at a time. The Phase-Evacuation Ratio controls for this by accounting for how many evacuations were performed, on average, in each Phase. A higher ratio would imply fewer evacuations per Phase, and a lower ratio would imply more. In the case of a substantial discrepancy between the two measurements, I include both side by side. Otherwise, I show only Phase-Evacuation Ratio.

One issue that still persists is that not all victims are equal in worth to teams – critical victims, which require the Medic and one other player to triage, are worth 50 points, while regular victims, requiring only the Medic, are worth 10.

I ultimately decided not to divide evacuations by whether they were critical victims or regular victims, on the assumption that, because the Transporter would be focused on discovering and evacuating victims and not assisting the Medic in triaging critical victims, the type of victim was unlikely to have an impact on the Transporter's behavior.

5.4.3 Cognitive Load and Probability of Forgetting

The ASIST Testbed additionally calculated several secondary measures, among them cognitive load and probability of forgetting, based on the ACT-R cognitive architecture.

ACT-R was first developed in 1993 by John R. Anderson as an update of his prior ACT and ACT* theories. [3]. Since then, it has gone through several iterations adding new cognitive modules, refining simulations, and usability updates [28]. Per Whitehill [36], ACT-R "models how humans recall 'chunks' of information from memory and how they solve problems by breaking them down into subgoals and applying knowledge from working memory as needed."

ACT-R has demonstrated correlation between its major modules and activation of areas of the human brain [25]. It has been previously modified to be used in human-robot interaction (HRI) research [33]. An ACT-R based model of Working Memory has been developed [21], and the ACT-R website includes 22 research papers studying working memory which utilized ACT-R [1].

$$1 - \Pi_i^{chunks} \frac{1}{1 + e^{\frac{\tau - A_i}{t}}} \tag{1}$$

[36]

Per John Anderson, the founder of the ACT-R system, and Phillip Pavlik, "ACT-R's retrieval memory system is based on a unitary trace that is composed of a sum of a number of individual strengthenings" [24]. The following formula for the activation of a chunk (which in this experiment are victims) expresses this sum of "strengthenings":

$$A_i = \ln \Sigma_{j=1}^{n} t_j^{-d} \tag{2}$$

Where j represents a particular interaction with that chunk (in this case, including discovery, triaging, and transporting) and t_j is the time since that interaction. This formula is based directly on the ACT-R formula for working memory node activation [21].

I also utilized “Cognitive Load,” a modification of the above “Probability of forgetting” model to instead express working memory in terms of number of chunks:

$$\Sigma_i^{chunks} e^{\frac{\tau - A_i}{t}} \tag{3}$$

Where i represents a given victim, t is a scaling temperature parameter, τ is the memory retrieval threshold, typically set to 1, and A_i represents the “Activation” for the victim, previously defined above.

The testbed calculates and reports these measures in real-time over the course of a trial. Due to ACT-R’s usage in measuring working memory, I consider it a valid metric.

While Cognitive Load and Probability of Forgetting are correlated with the number of victims discovered or evacuated, both do not simply reflect raw counts of the number of victims discovered or evacuated. Rather, Cognitive Load and Probability of Forgetting both consider these ”chunks” in terms of their impact on memory. Each chunk is considered both in the aggregate and individually, with each chunk separately being ”activated” and ”reactivated” as the player encounters it over the course of the experiment, and with each chunk decaying in activation level over time. In the context of my proposed model I consider each victim ”chunk” to be a separate task instance. Thus, what these ACT-R measures add to my overall thesis is providing an expression of how different task strategies impact overall performance with respect to how those task strategies impact the cognitive burden of managing related task instances in working memory.

I compared the average cognitive load and average probability of forgetting of each trial. In addition, I divided the trials into three groups, depending on whether the number of Phase counts for that trial was in the lower quartile (Low), in the upper quartile (High) or between the upper and lower quartile (Mid). I then analyzed how the average cognitive load for each group changed on a minute-by-minute basis.

5.4.4 Final Score

Final score refers to the total number of points awarded to a team for the evacuation of victims, with 10 points awarded for every regular victim successfully evacuated, and 50 points awarded for every critical victim successfully evacuated. Because score is the metric which the team is attempting to maximize, it is considered a satisfactory measure of overall performance.

5.4.5 Evacuated-Discovered Ratio

The Evacuated-Discovered Ratio was measured as the number of victims evacuated divided by the overall number of victims discovered. A ratio of 1 indicates that all victims discovered were evacuated, while a ratio of 0 would indicate that none of the victims discovered were evacuated.

A lower Evacuated-Discovered Ratio was considered an indicator of inefficiency and error, indicating that the team either

1. Misjudged how many victims they could evacuate in the time remaining
2. Forgot where discovered victims were and couldn't find them in the time remaining
3. Or forgot that the discovered victims existed at all.

To simplify matters, I classified victims that were discovered but not evacuated as "forgotten".

I additionally evaluated how many victims that were forgotten were triaged by the Medic. Because a victim must be triaged before it can be evacuated, it is important to know how many victims were triaged in order to identify the culpability of the Transporter in any forgotten victim.

These metrics are imperfect, both because the Transporter is not the only person who can evacuate, and because the Medic could theoretically triage several victims in rapid succession near the end of the mission, before the Transporter could reasonably respond. However, because it is assumed that the majority of triages would happen with time for the

Transporter to respond, and because the Transporter is expected to evacuate the most out of any roles, these metrics are considered satisfactory rough approximations.

5.4.6 Time Between Discovery and Evacuation

Discovery-Evacuation Time was measured as the average distance between when a victim was discovered (either by the victim appearing for the first time in a player's field of view, or by a Transporter's signal going off for the room that victim was in for the first time) and when that victim was evacuated. When compared against the Phase count and Phase-Evacuation Ratio, this could lend insight into the Transporter's overall efficiency with evacuations on an individual basis, and demonstrate the material effects of different Phase strategies on one element of performance – how quickly any individual victim is evacuated, once found.

5.4.7 Transporter-Medic Distance

The degree to which the Transporter and Medic are in proximity to each other was thought to provide insight into the degree that they shared role responsibilities with each other – for example, the Medic evacuating some victims, and the Transporter assisting the Medic in triaging critical victims. Thus, the average distance between the Medic and the Transporter was considered.

5.4.8 Total Distance Travelled by Transporter

I additionally examined Transporter behavior through overall distance traveled by the Transporter, on the hypothesis that more Phase switches would correlate with reduced distance traveled, implying greater efficiency. The testbed periodically recorded the player position, typically multiple times within a given second. I calculated the Euclidean distance between each of these updates for the same player, and then summed the total of these inter-update distances to obtain the total distance traveled by the Transporter.

5.4.9 Marker Block Placement

Outside of verbal communication, marker blocks were one of the primary ways players could communicate with other players as well as themselves. A marker block could be placed to signal the presence of an item of interest to another player, or as a reminder to the player themselves for later utilization.

Because placed marker blocks appeared on the dynamic map, it is possible that the amount of marker block utilization could be suggestive of utilization of the dynamic map. If true, and if the dynamic map, as I hypothesized, does reduce working memory load, we would expect more marker block placement to increase with fewer Phases and decrease with more Phases, indicating that by placing more of the burden for "remembering locations" on the dynamic map, players have less reason to Phase-switch more frequently, as they have reduced risk of working memory decay impacting their ability to find and retrieve the victim.

5.4.10 Competency

To what degree did pre-existing competency of the player in a Minecraft environment predict success, and to what degree did that competency predict number of Phases?

Both survey and experimental data were utilized to lend insight to this. Pre-experiment survey questions included questions regarding overall experience with videogames, self-assessment of videogame-related skills, and familiarity with Minecraft. Experimentally, a competency test was performed before the experiment as well, where players were individually assigned to complete a set of tasks comparable to the tasks they would be assigned in the experimental trials, with their time to completion recorded.

I separately investigate both data sources to see how each correlates with final team score as well as with Phase strategy.

5.4.11 Agent/Advisor

Because different agents had different architectures and offered different interventions at different times, the possibility existed that these interventions had variable impacts on overall team behavior and on Phase strategy. I thus examined these agents in the following manners:

- Whether interventions from these ASIs significantly impacted player strategy
- A qualitative analysis of how agents differed in the content of interventions.
- How performance differed between different ASI conditions in terms of final score, average Phase count, and average Phase-evacuation ratio.
- Whether there is a significant difference between advisor groups, both in terms of competency as well as date of experiment for advisor group.

5.4.12 Practice Effect

Each team participated in two trials, with the two trials separated by a short period for filling out additional survey questions. A substantial practice effect was thus expected. I examined how final score, Phase count, and the Phase-Evacuation Ratio differed between the first and second trials, and whether this has implications for conclusions about what influences the use of different Phase strategies.

5.4.13 Santa Barbara Sense of Direction (SBSOD)

The Santa Barbara Sense of Direction is a self-assessment of respondent memory and sense of direction [16]. This assessment was given to players prior to the trials, and is considered with regards to whether any responses on the questionnaire are associated with cognitive load, Phase count, or Phase-evacuation ratio.

5.4.14 Percent of Evacuations Performed by Transporter

Because all players have the ability to evacuate victims, it is important to know how many victims actually *were* evacuated by the Transporter, as that would lend insight into the degree that the Phase Count and Phase-Evacuation Ratio of the Transporter could plausibly directly impact the overall performance of the team. Thus, the percent of evacuations performed by the Transporter, as opposed to by the Medic or Engineer, was considered.

5.4.15 Percent of Medic Assists Performed by Transporter

When considering the number of Phases and the Phase-evacuation ratio, additional context was gleaned by considering the degree that the Transporter was performing other actions not directly related to its primary role – most prominently, how often it assisted the Medic in triaging critical victims. Thus, Phase count and Phase-Evacuation Ratio were compared with the percent of Medic assists by the Transporter (out of all Medic assists shared between the Transporter and the Engineer).

6.0 Results

6.1 Distribution of Results

6.1.1 Phase Count

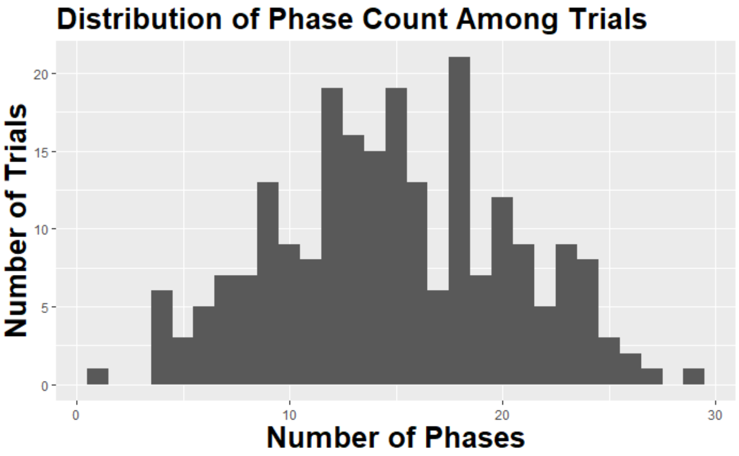


Figure 2: Phase Count Distribution

The mean Phase count among trials was 14.86, with a median of 15.00. The minimum number of Phases was 1 and the maximum was 29. Distribution is roughly symmetric, with a Pearson Skew of 0.074.

6.1.2 Phase-Evacuation Ratio

The mean Phase-Evacuation Ratio among trials was 0.7689, with a median of 0.7860. In other words, this means that on-average, for every three Phases, four victims were evacuated. The minimum Phase-Evacuation Ratio was 0.16667 (between 8 and 9 evacuated for

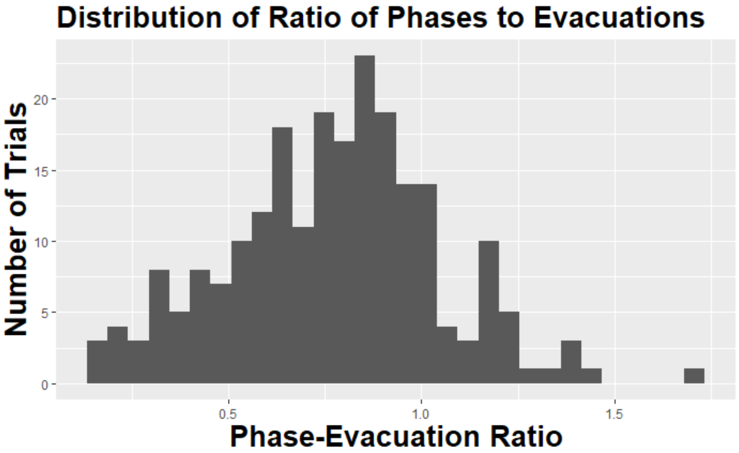


Figure 3: Distribution of Ratio of Phases to Evacuations

every Phase), while the maximum was 1.7143 (nearly two Phases for every one evacuation). Distribution is roughly symmetric, with a Pearson skew of 0.072

6.1.3 Average Cognitive Load

The mean average cognitive load over the course of a trial was 1.25 with a median of 1.03. Minimum average cognitive load was 0.4053 and maximum was 4.2861. Distribution was significantly positively skewed, with a Pearson skew of 2.02.

6.1.4 Average Probability of Forgetting

The mean average probability of forgetting over the course of a trial was 0.54 with a median of 0.51. The minimum average probability of forgetting was 0.28 while the maximum was 0.89. Distribution was positively skewed, with a Pearson skew of 0.71.

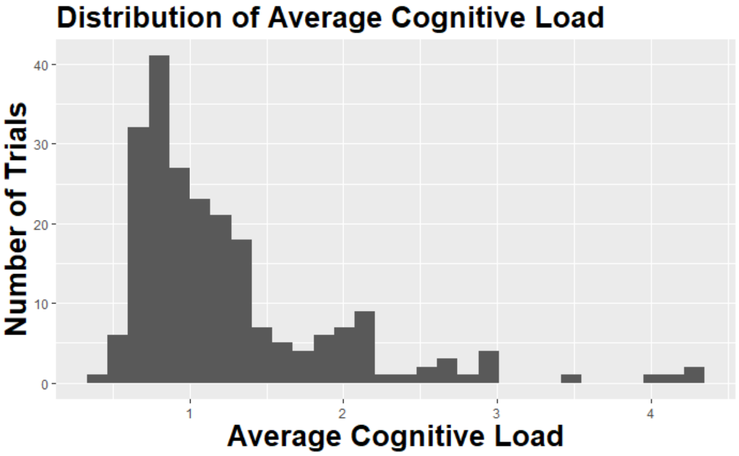


Figure 4: Average Cognitive Load Distribution

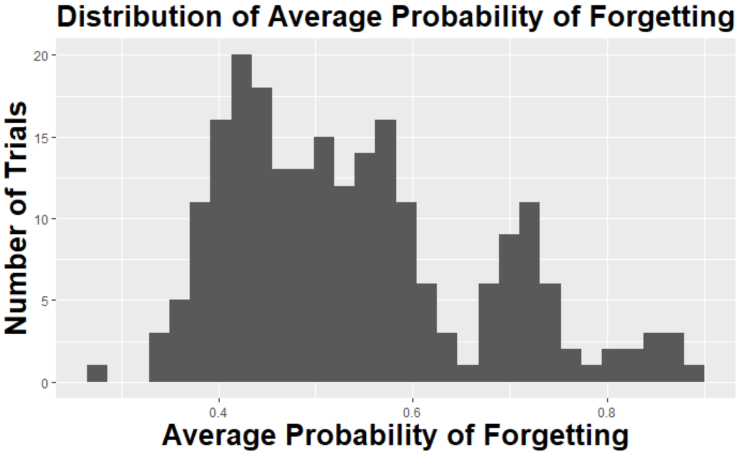


Figure 5: Average Probability of Forgetting

6.1.5 Final Score

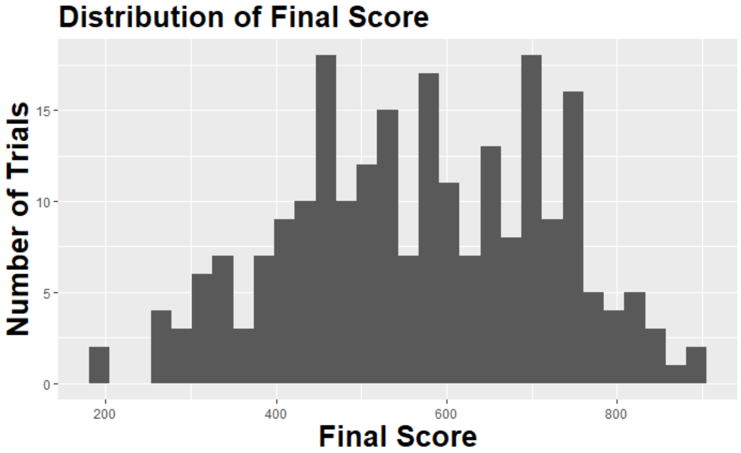


Figure 6: Final Score Distribution

Final score for teams at the end of a trial ranged from 190 to 890, with a mean score of 567 and a median score of 570. Distribution was somewhat negative, with a Pearson skew of -0.104.

6.2 Phase Count and Phase-Evacuation Ratio

The number of Phases showed a positive relationship with evacuation count, $R^2 = 0.12$.

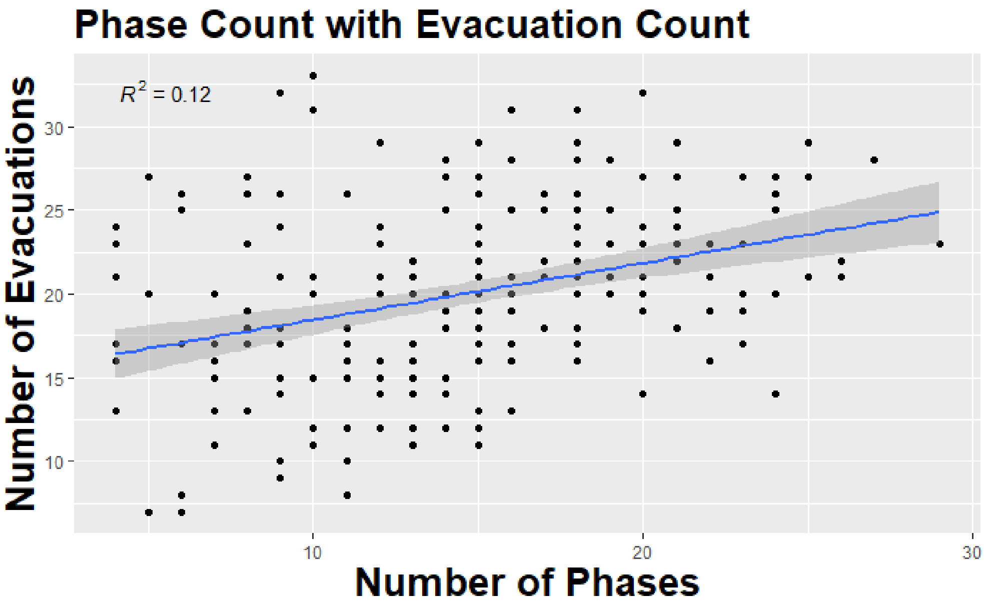


Figure 7: Phase Count v. Evacuation Count

6.3 Cognitive Load and Probability of Forgetting

6.3.0.1 Phase Metrics with Average Cognitive Load and Average Probability of Forgetting

Both Phase Count and Phase-Evacuation Ratio showed a substantially negative relationship with average cognitive load – as either increased, average cognitive load decreased. For Phase-Evacuation Ratio, the correlation was $R^2 = 0.22$

This was similarly shown with average probability of forgetting, which is simply a different manner of representation of cognitive load, instead emphasizing the likelihood of forgetting an individual chunk in memory. For Probability of Forgetting, the correlation was $R^2 = 0.20$.

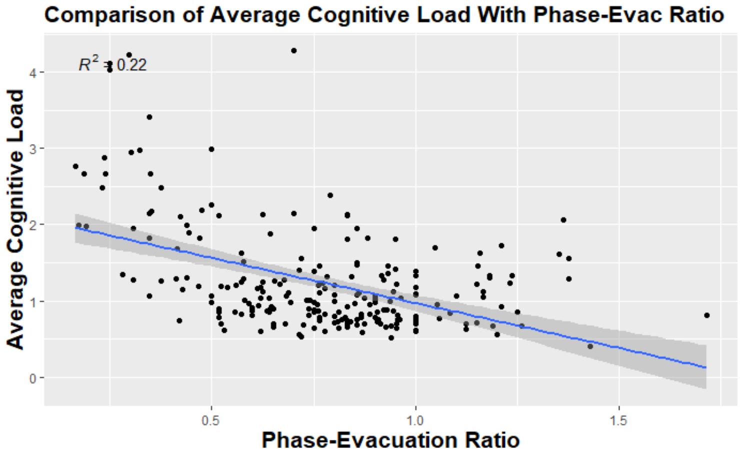


Figure 8: Phase-Evacuation Ratio v. Average Cognitive Load

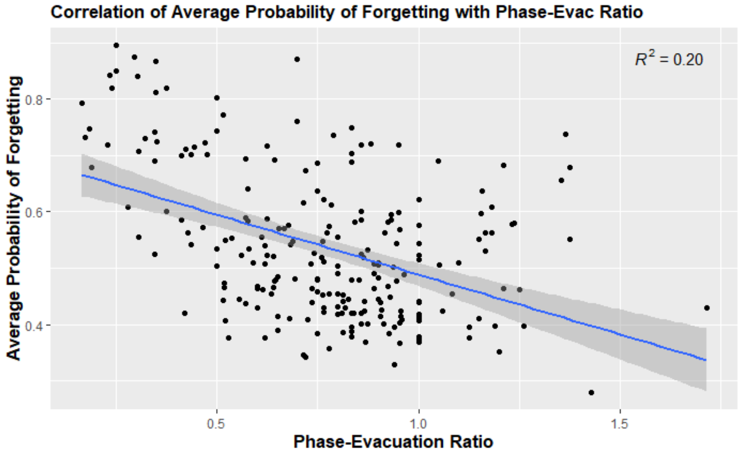


Figure 9: Phase-Evacuation Ratio v. Average Probability of Forgetting

6.3.0.2 Average Cognitive Load Over Time, By Phase Group

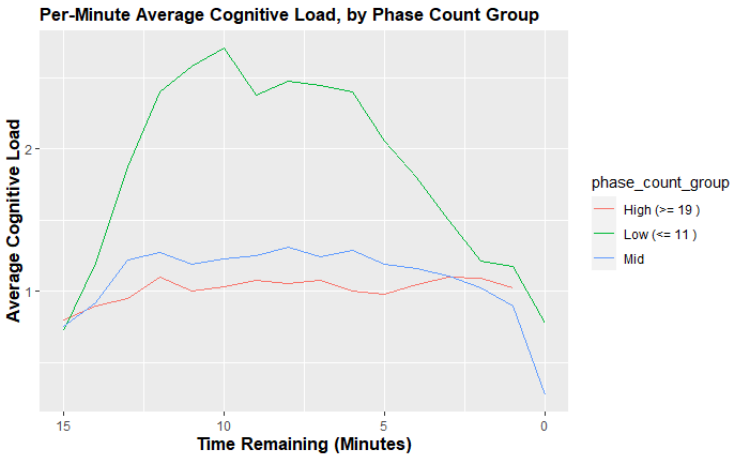


Figure 10: Per-Minute Average Cognitive Load, by Phase Count Group

Lower Phase count and lower Phase-Evacuation groups showed substantially higher cognitive load throughout the middle of the trial, diverging early on from the Low and Mid phase count groups and converging only near the end of the trial, which most victims found are expected to be evacuated.

6.4 Final Score

While Phase count shows a roughly 11% correlation with final score, the Phase-Evacuation Ratio shows the opposite result, inversely correlated 11% with final score, suggesting that fewer Phases per evacuation were associated with higher final score.

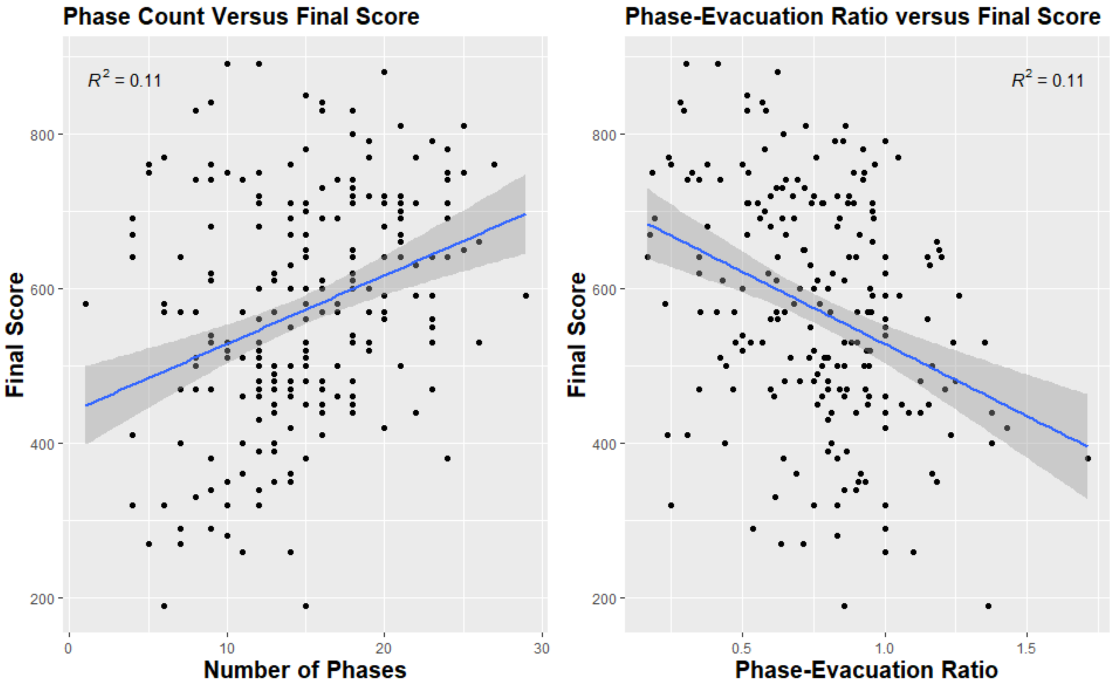


Figure 11: Phase Count and Phase-Evac Ratio v. Final Score

6.5 Evacuated-Discovered Ratio

Comparing the Phase count and Phase-Evacuated Ratio with the Evacuated-Discovered Ratio produced mixed and weak results. While Phase count suggested that an increase in the number of Phases was weakly ($R^2 = 0.08$) associated with an increase in the Evacuated-Discovered Ratio, the Phase-Evacuated Ratio showed the opposite at roughly the same level of fit ($R^2 = 0.06$).

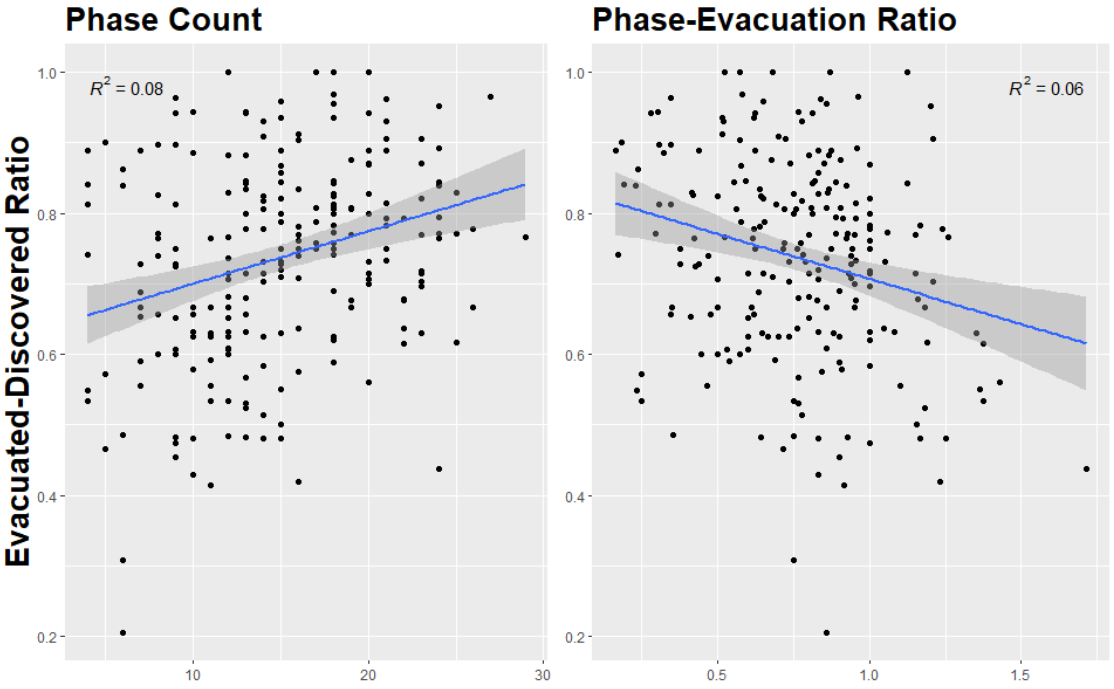


Figure 12: Phase-Evacuation Ratio and Phase Count v. Evacuated-Discovered Ratio

6.5.1 Percent of Victims Discovered and Triaged but Not Evacuated

Most victims discovered but not evacuated – which I classify as ”forgotten” – were not triaged. The mean percent of triaged forgotten victims was 30%, the median was 22%, while the 1st and 3rd quartiles were 9% and 45%, respectively. In other words, most forgotten victims were not in a state where the Transporter could have evacuated them.

Relatedly, there was no observable relationship between the number of Phases and the overall percent of victims forgotten, with $R^2 < 0.01$.

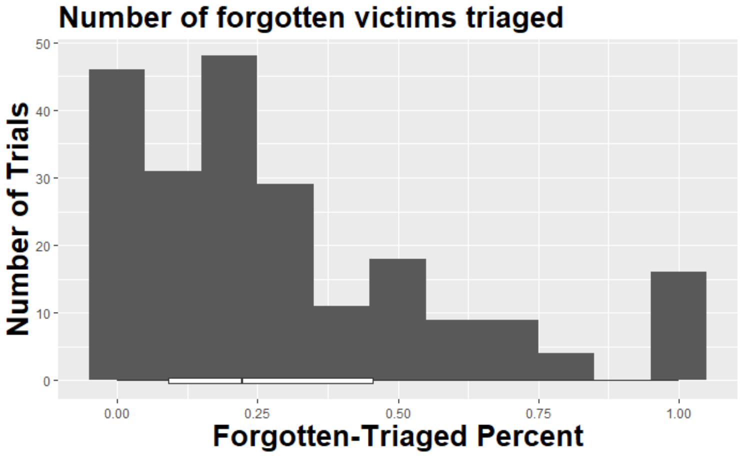


Figure 13: Forgotten-Triaged Percent

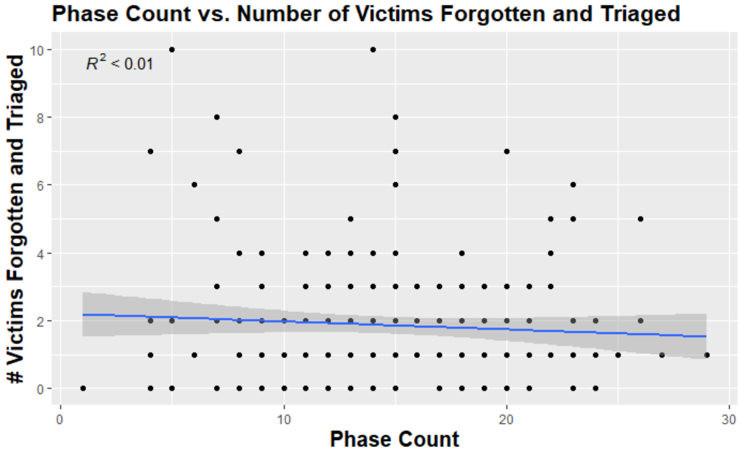


Figure 14: Phase Count vs. Forgotten and Triaged

6.6 Time Between Discovery and Evacuation

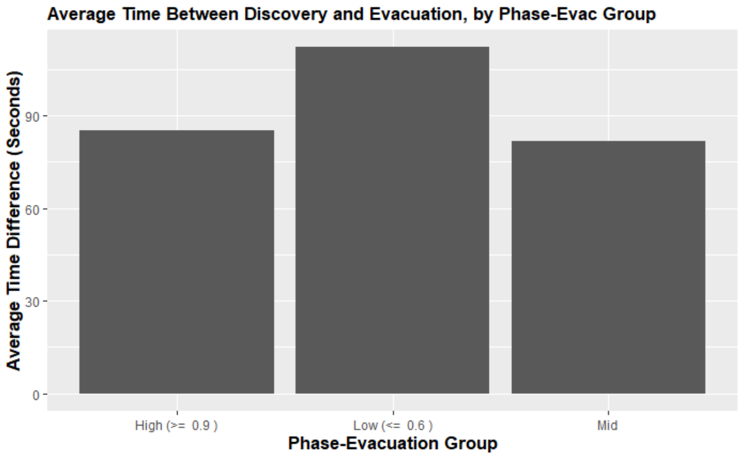


Figure 15: Average Time Between Discovery and Evacuation, by Phase-Evac Group

When more Phases occurred per evacuation, the time between when a victim was discovered and when that victim was evacuated was substantially smaller than when fewer Phases occurred per evacuation. Teams with a Phase-Evacuation Ratio of greater than 0.93 had an average time difference of 85 seconds 0.6 showing at most an average time difference of 85 seconds, while teams with a Phase-Evacuation Ratio between 0.60 and 0.93 showed an average time difference of 81 seconds, and teams with a Phase-Evacuation ratio of 0.6 or less showing an average time difference of 112 seconds.

6.7 Transporter-Medic Distance

Phase-Evacuation Ratio showed an inverse relation with the average distance between the Transporter and the Medic over the course of the trial, with $R^2 = 0.13$

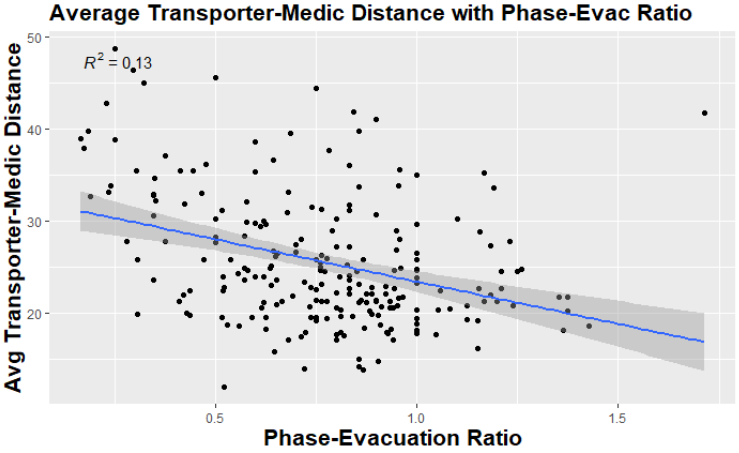


Figure 16: Phase-Evacuation Ratio v. Transporter-Medic Distance

6.8 Total Distance Traveled by Transporter

Total distance traveled by Transporter showed a strong positive correlation with final score, with $R^2 = 0.39$. Conversely, Transporter Phase-Evacuation Ratio showed a strong negative correlation with final score, with $R^2 = 0.24$.

6.9 Marker Block Placement

Transporters placed marker blocks substantially more frequently than Medic (56% of all marker blocks placed for Transporter, compared to 27% for Medic), which in turn placed marker blocks substantially more frequently than Engineers (16%).

Marker blocks were of eight possible categories, five of which unambiguously related to the presence or absence of victims. "abrasion" and "bonedamage" denoted the two different kinds of regular victim, and correspondingly denoted the evacuation zones on the map which

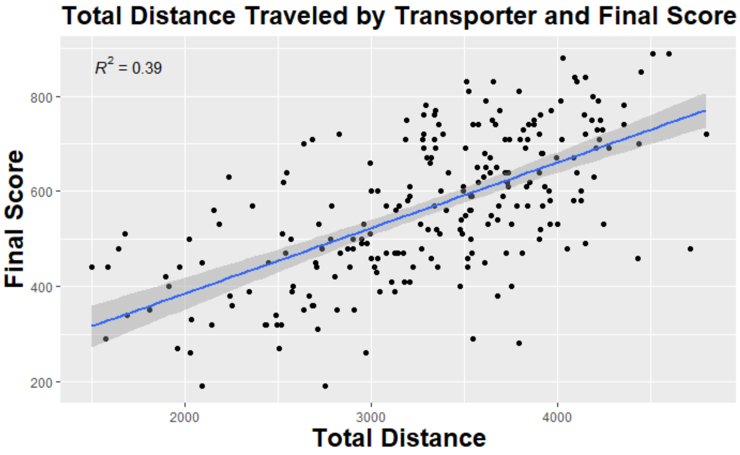


Figure 17: Transporter Distance v. Final Score

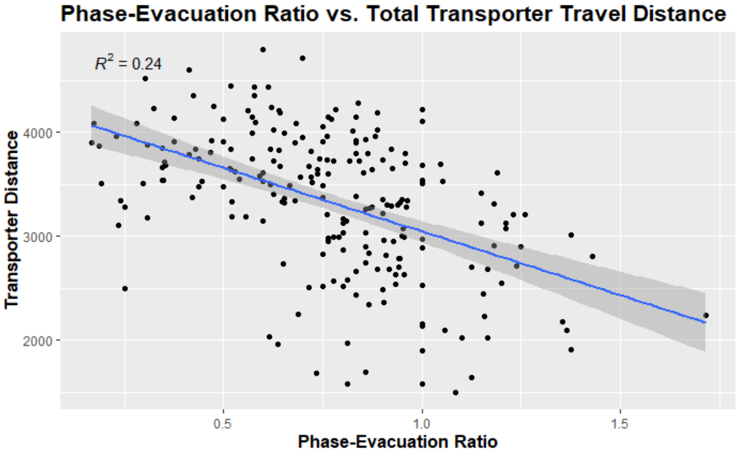


Figure 18: Phase-Evacuation Ratio v. Transporter Distance

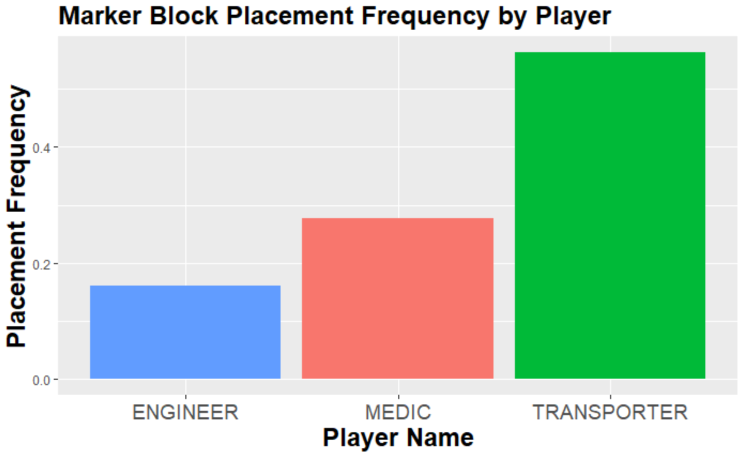


Figure 19: Marker Block Placement Frequency by Player

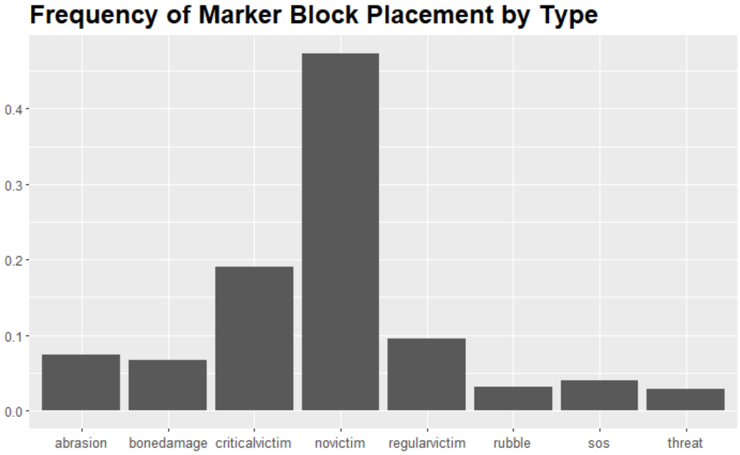


Figure 20: Frequency of Marker Block Placement, by type

the Transporter would need to evacuate victims to. Only the Medic has the tools needed to identify if a victim is “abrasion” or “bonedamage”, and thus these markers were most often placed by the Medic.

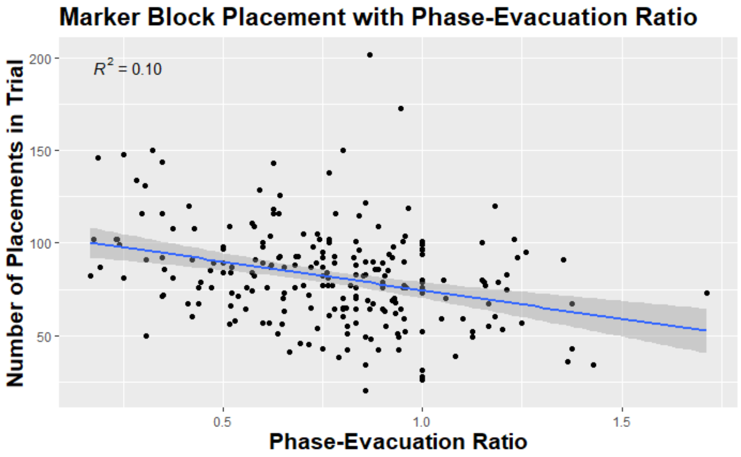


Figure 21: Marker Block Placements with Phase-Evacuation Ratio

Marker block placement tended to decrease as the number of Phases per evacuation increased, with an $R^2 = 0.10$.

There is at least some evidence ($R^2 = 0.11$) of a relationship between how many marker blocks were placed with final score.

6.10 Competency

6.10.1 Survey

Two questions from pre-trial surveys were of interest in this analysis:

The first, hereafter referred to as “Game Experience”: “Please rate your experience level

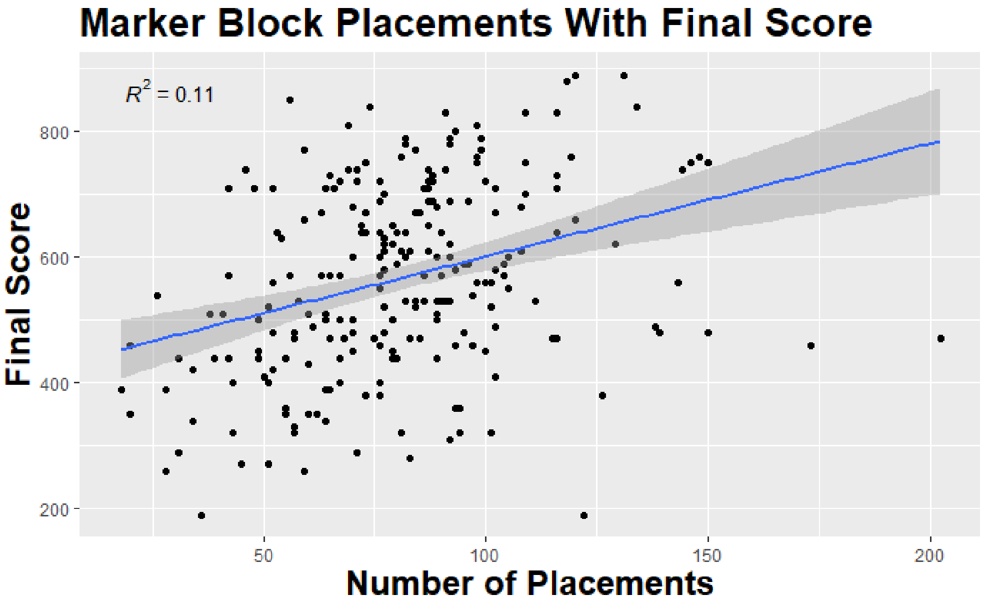


Figure 22: Marker Block Placement With Final Score

in playing video games. Please be as honest and objective as possible.", with responses ranging from "Novice" to "Expert".

The second, hereafter referred to as "Minecraft Experience": "Please indicate how many years of experience you have with the following: - Years playing Minecraft (any versions or styles of play)."

Responses to these questions were numericized and then compared both in terms of average team values (i.e. average game experience and average Minecraft experience), and in terms of the Transporter themselves.

Both average game experience ($R^2 = 0.09$) and average Minecraft experience ($R^2 =$

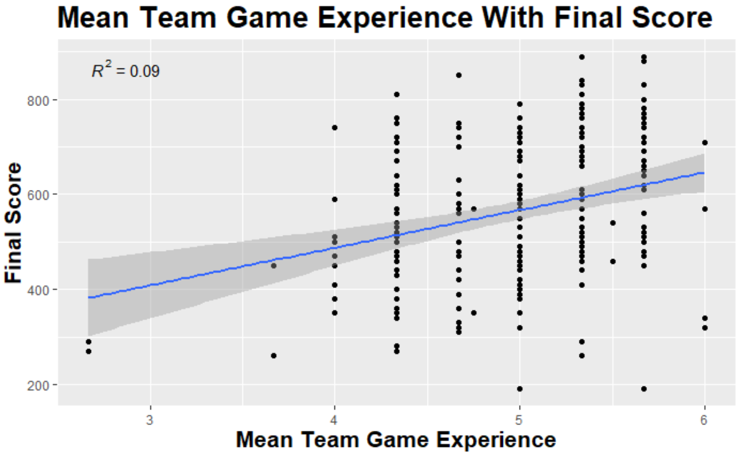


Figure 23: Mean Team Game Experience With Final Score

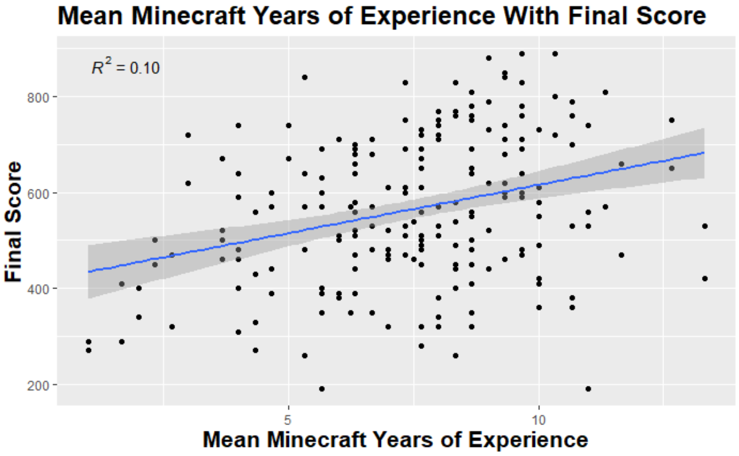


Figure 24: Mean Minecraft Years of Experience with Final Score

0.10) was positively correlated with final score. Similarly, game experience and Minecraft experience of the Transporter correlated with final score, although to a lesser degree.

In contrast to the comparisons with final score, the relation of results to Phase count and Phase-Evacuation Ratio were insubstantial. Whether evaluating on a team level or on the level of the Transporter alone, whether considering Game experience or Minecraft experience, and whether considering Phase Count or Phase-Evacuation Ratio, results were slight if present at all, with R^2 of 0.06 or less.

6.10.2 Competency Task

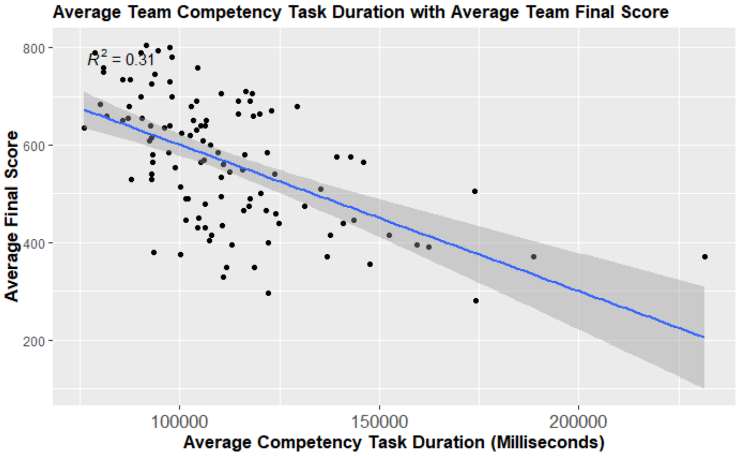


Figure 25: Average Team Competency Task Duration with Average Team Final Score

The longer players within a team took to complete a competency task on average, the poorer that teamed perform on both trials, with $R^2 = 0.31$.

While Phase count shows an insubstantial relationship with average completion time for the competency task ($R^2 = 0.03$), teams with a lower average completion time for the

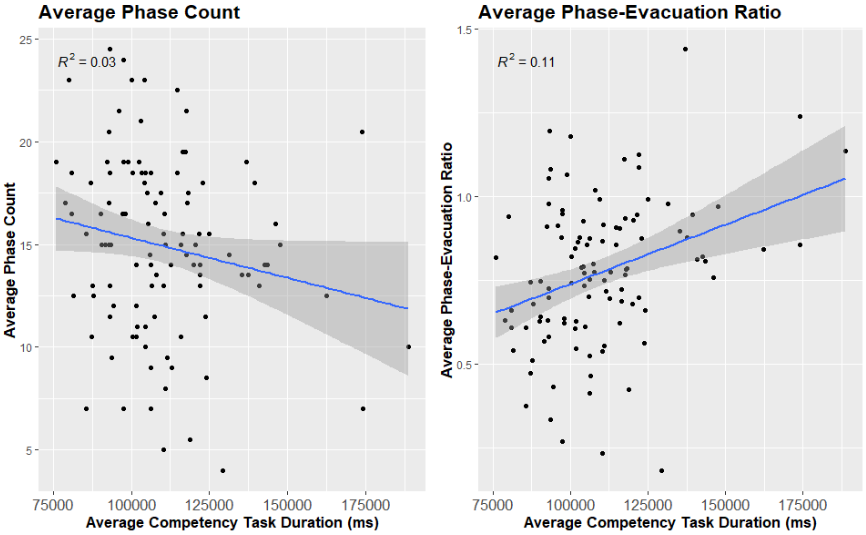


Figure 26: Average Team Competency Task Duration with Average Phase Count and Phase-Evacuation Ratio

competency task also tended to have a lower average Phase evacuation ratio, evacuating more victims within the same number of Phases ($R^2 = 0.11$).

6.11 Agent/Advisor

6.11.1 ASI Advisor Commonalities

CMU, CRA, SIFT, and possibly DOLL, in particular, encourage utilizing marker blocks and other forms of communication that could plausibly impact working memory load regarding the location of victims. For example, the following intervention was deployed by CMU's ATLAS agent 14 times over the course of the 28 trials it was associated with:

Medic, you may have identified that victim with the wrong marker block.

Table 5: Advisor Commonalities

	CMU	CRA	DOLL	UAZ	SIFT	USC
Identify missed/incorrect markers	YES	YES	YES	NO	YES	NO
Encourages removing obsolete/stale markers	YES	NO	YES	NO	NO	NO
Encourages communication after placing markers	NO	NO	NO	YES	YES	NO
Encourages marker placement	YES	YES	YES	NO	YES	NO
Encourages informing other players about placed/triaged victims	YES	NO	NO	YES	NO	NO
Update on total victims evacuated/needing evacuation	NO	NO	YES	NO	YES	NO
Encourage team communication (general)	YES	YES	YES	YES	YES	YES
Encourage using role-specific skills	YES	YES	YES	NO	NO	NO
Encourage prioritizing critical victims	YES	YES	NO	NO	NO	NO
Encourage Medic-Engineer Cooperation	YES	YES	NO	NO	NO	NO
Corrects players on missing/incorrect marker block placement	YES	YES	NO	NO	YES	NO

6.11.2 Results by Advisor Condition

All advisors performed worse than either human advisor or no-advisor conditions. No-Advisor and Human Advisor conditions had average final scores of 621 and 622, respectively; Non-ASI final scores ranged from 501 (USC) to 559 (SIFT).

Comparing Phase Count and Phase-Evacuation Ratio: While human advisor still showed the highest average Phase count at 16.32 Phases, CMURI showed the second highest at 16.09 Phases, and No Advisor was the fourth-highest at 14.967 Phases. With Phase-Evacuation Ratio, UAZ showed the highest Phase-Evacuation Ratio, while Human Advisor was third highest at 0.811, and No Advisor was second lowest at 0.702.

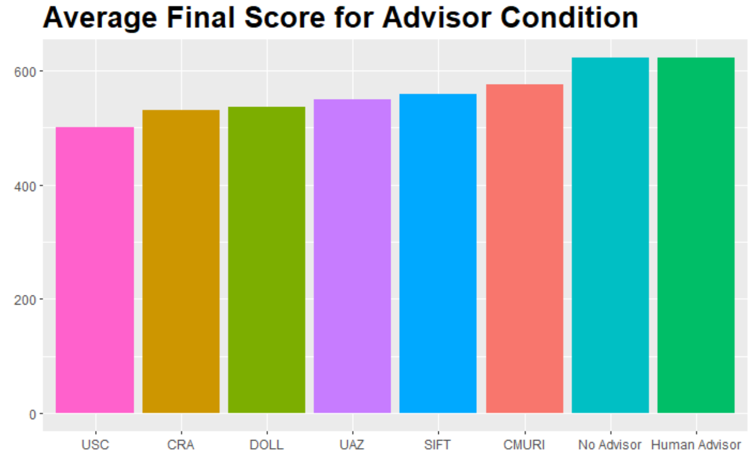


Figure 27: Average Final Score for Advisor Condition

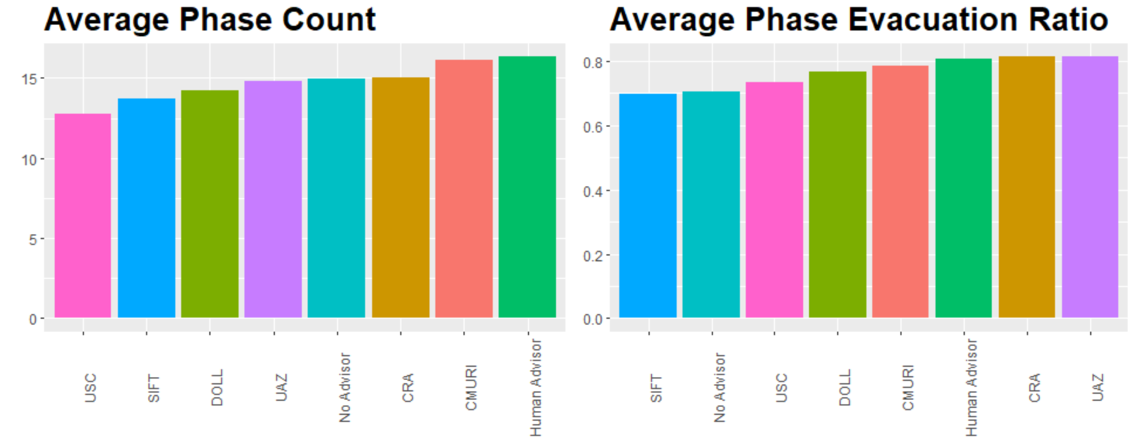


Figure 28: Average Phase Count/Phase-Evacuation Ratio for Advisor Condition

6.11.3 Competency by Advisor Group

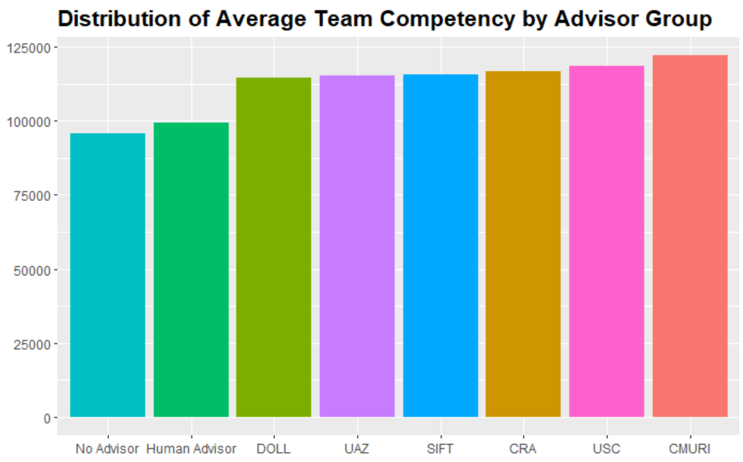


Figure 29: Distribution of Average Team Competency Completion Time by Advisor Group

Taking competency into account, participants for the Human and No Advisor conditions showed a substantially lower time to completion for the competency task compared to the advisor trials.

Applying the Welch t-test to average team competency (as measured by completion times on the competency task) between ASI and non-ASI groups yielded a t-score of -5.0461 and a p-value of 2.546e-06, indicating extremely high confidence of these two populations being substantially distinct.

6.11.4 Date of Advisor Conditions

Human and no-advisor trials were hosted in April and early May, prior to all other advisor trials, which were randomized in the period of May through July.

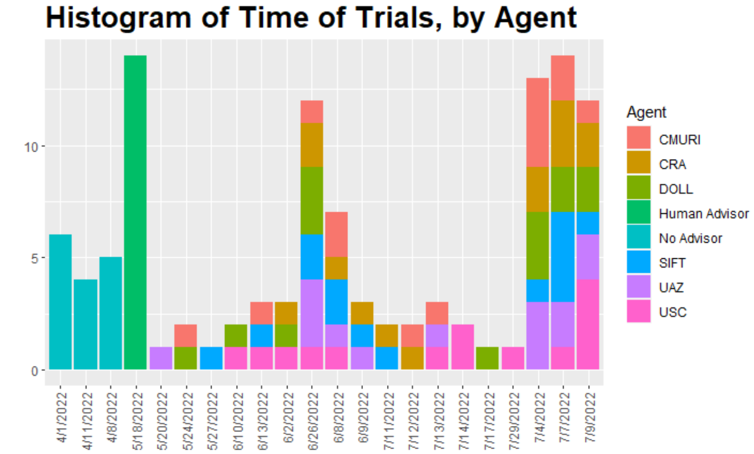


Figure 30: Histogram of Time of Trials, by Agent

6.12 Practice Effect

With an average and median increase of 100 points (521 to 621) and an IQR of 17.5 to 190.0, teams typically saw improved performance in the second trial of on average 19%, as would be expected by the practice effect.

Phase count, meanwhile, saw a median increase of 2 Phases and an average increase of roughly 0.4, from 14.7 to 15.1, a more modest increase of 2.7%, with an IQR of -4 to 5.

Finally, the average Phase-Evacuation Ratio changed from 0.828 to 0.710, a decrease of roughly 14%, with an IQR of -0.29 to 0.08.

6.13 Santa Barbara Sense of Direction (SBSOD)

To make data comparisons easier, I've created a key to shorten SBSOD survey questions, to better compare answers side-by-side.

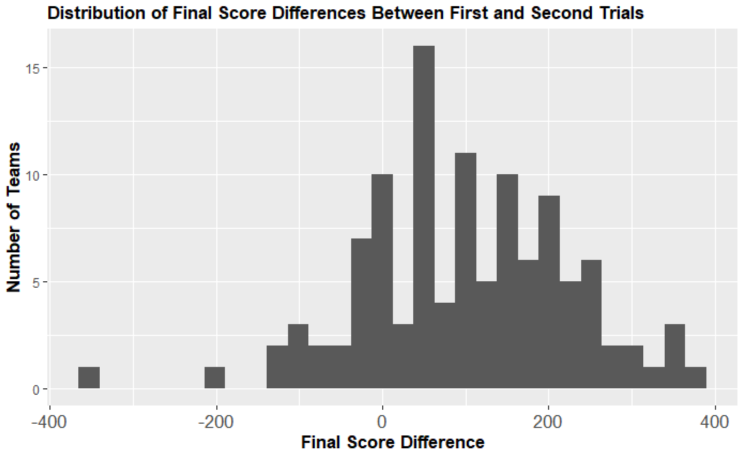


Figure 31: Distribution of Final Score Differences Between First and Second Trials

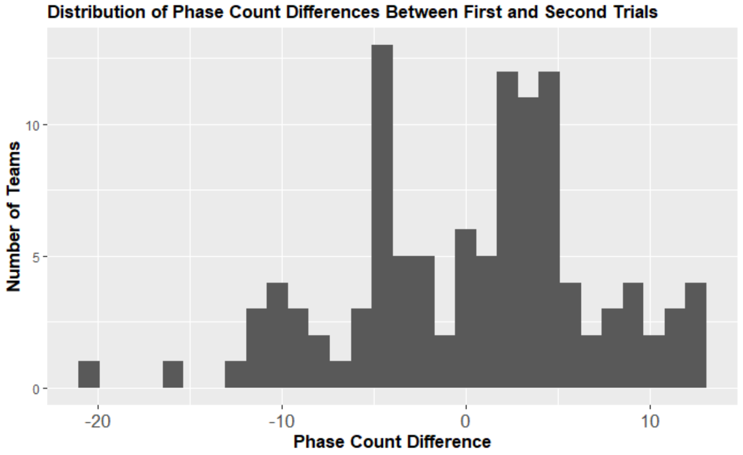


Figure 32: Distribution of Phase Count Differences Between First and Second Trials

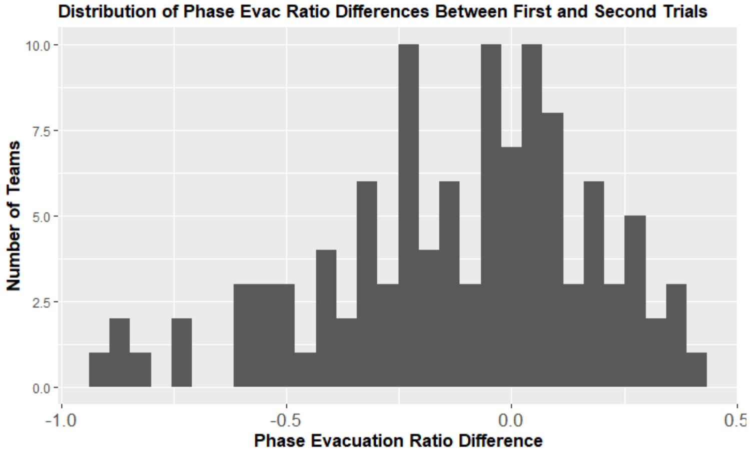


Figure 33: Distribution of Phase-Evac Difference Between First and Second Trials

Distribution of values giving in figure. Being a 7-point Likert Scale, 4 represents “Neither Agree nor Disagree”, greater than 4 represents “Agree”, and less than 4 represents “Disagree”.

I additionally took the correlations between survey questions and Phase Count, Phase-Evacuation Ratio, and Average Cognitive Load.

Q1, Q5, Q7, Q11, Q12, and Q13 all relate to sense of direction, and show the largest relationships when taking Phase count into consideration, although these responses become more muted when normalized by number of evacuations, leaving only Q5 and Q7 above R levels of 0.1. Average cognitive load also shows low relationships, with only Q12 (“It’s not important for me to know where I am”) above $R = 0.1$, and Q2 (“I have a poor memory for where I left things”) closest to that threshold of the remaining questions with $R = -0.09$.

Notably, nearly all questions with a significant relationship show a substantially skewed distribution, either primarily agree or primarily disagree (Q1 is 90% agree, Q7 is 82% agree, Q11 is 23% agree, Q12 is 7% agree, and Q13 is 20% agree). Of the previously mentioned six sense of direction questions, only Q5 is not heavily skewed either heavily in favor of agree or heavily in favor of disagree, although responses were nonetheless more likely to be agree than disagree (58% agree).

When considering SBSOD results on a numeric scale (not necessarily advisable for a Likert Scale, but considered nonetheless), results thus generally fail to achieve significance, with questions relating to direction coming the closest to significance, although whether this is because of a genuine relationship between Phase and sense of direction, or a mere statistical byproduct is unclear.

Considering results instead in terms of how many agreed or disagreed with each question, which may be more appropriate for a Likert Scale, also did not yield particularly relevant results. When performing Welch's T-Test on "Agree/Disagree" distributions for each question, the highest t-value was Q9 ("I am very good at reading maps") at 0.163, which corresponds to a P-value of 0.435, falling far short of significance. In other words, there is no evidence that Agree and Disagree populations substantially differed in their Phase-Evacuation Ratios.

6.14 Percent of Evacuations Performed By Transporter

Whether evaluated via raw Phase count or Phase-evacuation ratio, an increase in the number of Phases was associated with a decrease in number of evacuations performed by the transporter, with R^2 between 0.11 for Phase count and 0.20 for Phase-Evacuation Ratio.

Correspondingly, the number of evacuations performed by the Medic and Engineer increased, with a greater increase by the Medic ($R^2 = 0.19$) than the Engineer ($R^2 = 0.12$), who as the slowest member of the team is least productively used transporting victims across large distances on the map.

6.15 Percent of Medic Assists Performed By Transporter

No relationship of significance was found when comparing the Phase-Evacuation Ratio with percent of Medic assists by the Transporter ($R^2 = 0.06$).

Table 6: SBSOD Survey Questions

Q1	I am very good at giving directions.
Q2	I have a poor memory for where I left things.
Q3	I am very good at judging distances.
Q4	My sense of direction is very good.
Q5	I tend to think of my environment in terms of cardinal directions (North, South, East, West).
Q6	I can easily get lost in a new city.
Q7	I enjoy reading maps.
Q8	I have trouble understanding directions.
Q9	I am very good at reading maps.
Q10	I don't remember routes very well while riding as a passenger in a car.
Q11	I don't enjoy giving directions.
Q12	It's not important to me to know where I am.
Q13	I usually let someone else do the navigational planning for long trips.
Q14	I can usually remember a new route after I have traveled it only once.
Q15	I don't have a good mental map of my environment.

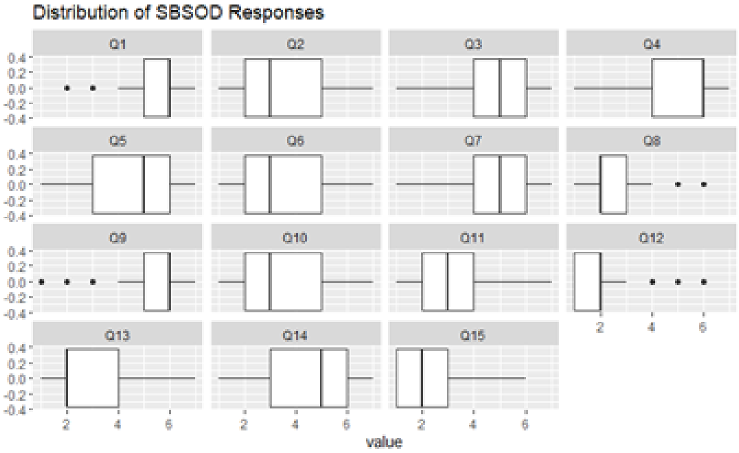
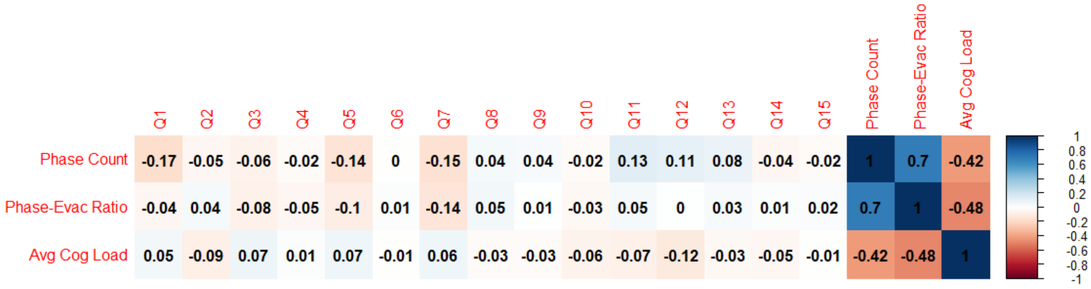


Figure 34: Distribution of SBSOD Responses



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Phase Count	Phase-Evac Ratio	Avg Cog Load
Phase Count	-0.17	-0.05	-0.06	-0.02	-0.14	0	-0.15	0.04	0.04	-0.02	0.13	0.11	0.08	-0.04	-0.02	1	0.7	-0.42
Phase-Evac Ratio	-0.04	0.04	-0.08	-0.05	-0.1	0.01	-0.14	0.05	0.01	-0.03	0.05	0	0.03	0.01	0.02	0.7	1	-0.48
Avg Cog Load	0.05	-0.09	0.07	0.01	0.07	-0.01	0.06	-0.03	-0.03	-0.06	-0.07	-0.12	-0.03	-0.05	-0.01	-0.42	-0.48	1

Figure 35: SBSOD Correlations

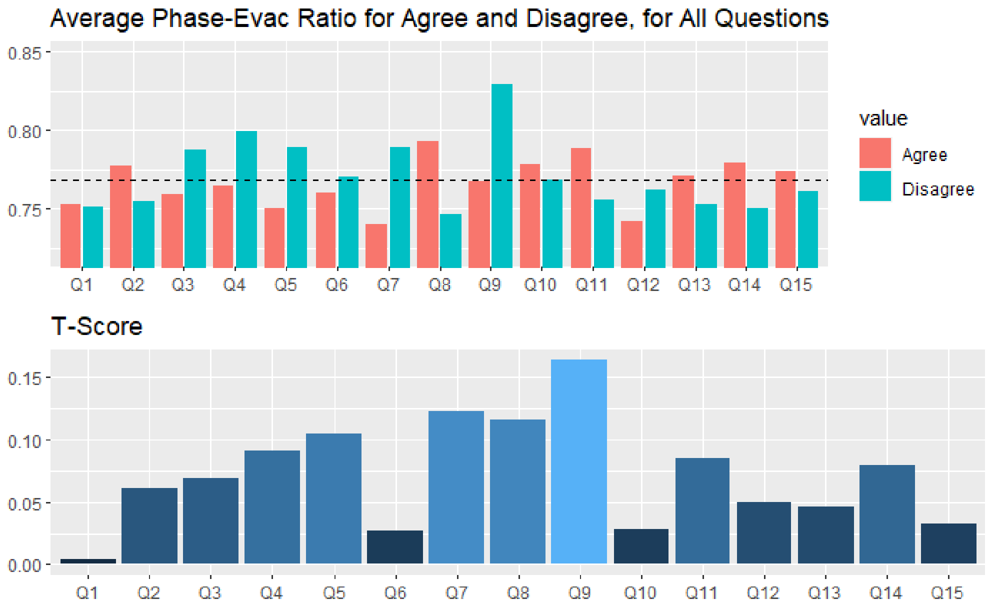


Figure 36: Average Phase-Evac Ratio for Agree and Disagree, for All Questions, and associated T-scores for each. Dashed line represents overall mean.

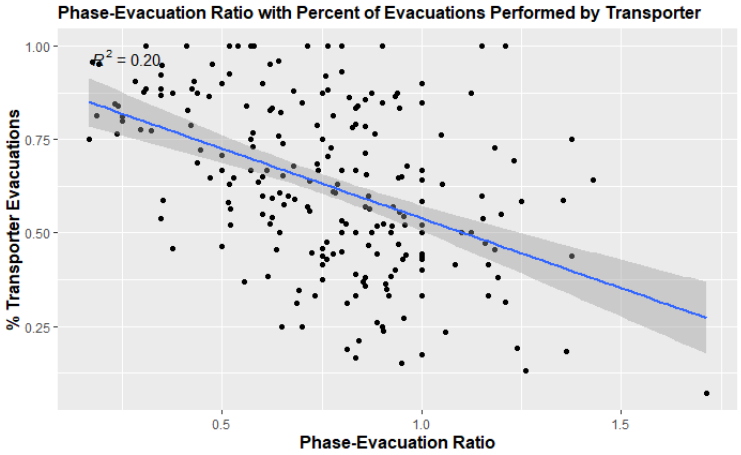


Figure 37: Phase-Evacuation Ratio v. Transporter Evacuation Count

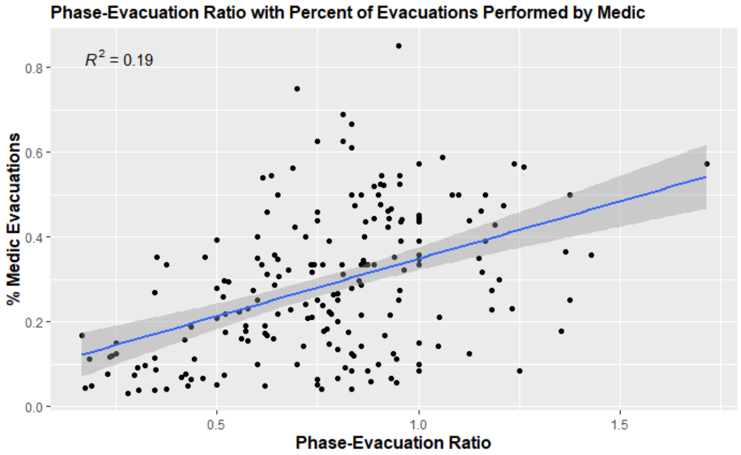


Figure 38: Phase-Evacuation Ratio v. Medic Evacuation Percentage

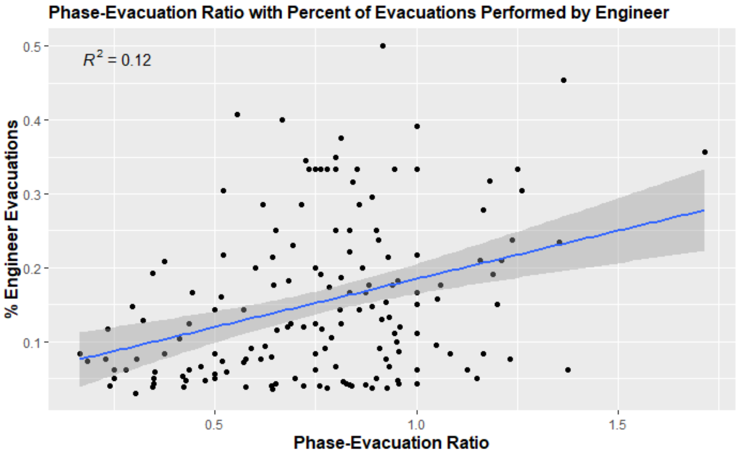


Figure 39: Phase-Evacuation Ratio v. Engineer Evacuation Percentage

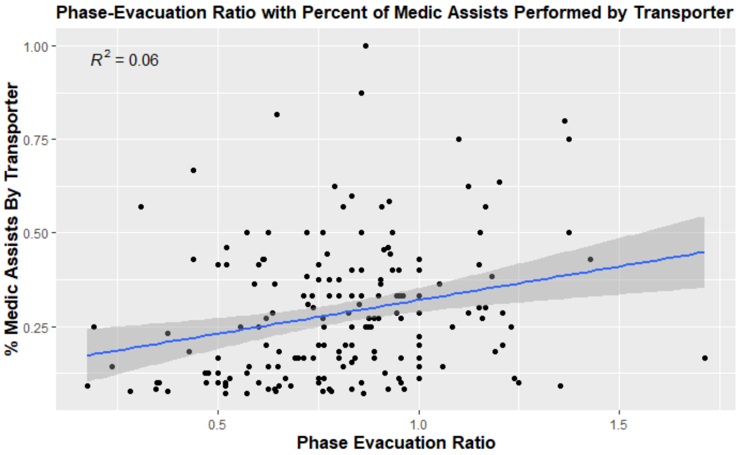


Figure 40: Phase-Evacuation Ratio v. Transporter Medic Assists

7.0 Discussion

7.1 Limitations

Any analysis of this data must take into account the considerable limitations of the experimental environment and the data produced from it.

The first and most obvious is that this experiment was not designed to study multitasking at all, but was instead meant to evaluate how players responded to several different Artificial Social Intelligences (ASIs). As a consequence, the design of the study introduced a number of features which, while they may make sense for evaluating ASIs, do not benefit an analysis of multitasking.

Among these is the use of multiple players who interacted, both verbally and nonverbally, with each other. Players discussed strategy, cooperated or failed to cooperate with each other, and modified the progress of each testing session in ways that, individually, each confound the results of the study.

However, barring a more in-depth and likely inherently subjective analysis of how the unique remarks by each player in each team confounded the overall trial, the influence of individual player coordination will be disregarded for the purposes of this thesis. Instead, I will focus on confounds which have more quantifiable metrics: The ASI agents and the Dynamic Map.

7.1.1 Agents

All advisors performed worse than either human advisor or no-advisor conditions, but there are considerable confounds which throw doubt onto whether these variations are the result of advisors or other causes.

Notably, one of the most significant determinators of experiment success was prior Minecraft experience and overall competency. It is highly probable that the ASI and non-

ASI groups should be considered different populations, throwing analyses based on whether a given team is assigned to the ASI vs. non-ASI condition into substantial doubt.

Competency could also plausibly affect the comfort players had with how quickly they switched Phases, and their overall capacity to remember features about the Minecraft environment. However, no noticeable pattern was found when comparing Phase count and Phase-Evacuation Ratio between different advisor conditions.

A partial answer to why these populations are so different is may be found in examining the times at which participants participated in different advisor conditions, with the Human and None conditions occurring first and the subsequent ASI conditions occurring in randomized order afterwards. With Welch's t-test showcasing a strong possibility of the early human and no-advisor trials being distinct from the advisor trials in terms of competency, it is highly likely that these two populations were distinct.

With considerable discrepancies between different advisor conditions, data analysis thus becomes substantially more fraught. Each advisor condition has only 14-15 teams, and thus only 28-30 trials. Separating by each advisor condition reduces overall significance of any findings. Attempting to compensate for the reduction in sample size by collating similar groups together is also fraught – while Human and No-Advisor conditions likely have a similar participant population, human advisors may plausibly confound results compared with no-advisors (although prior data finds their average final scores to be roughly the same). Similarly, while the ASI conditions are randomized, each agent gave a different set of interventions under different conditions.

Though the final scores did not vary significantly among participants in different ASI conditions (excepting the human and None conditions), it is still possible that ASI interventions influenced overall performance. Some ASIs, for example, recommended players either start or stop focusing on discovering new victims and instead focus on triaging and evacuating existing victims, which could thus prompt Phase changes.

7.1.2 The Dynamic Map

Perhaps the most problematic element, however, was the dynamic map, which when used properly can obviate much of the need to retain information about the status and location of victims in working memory. While data relating marker block placement to number of Phases suggests the degree of utilization of the dynamic map (presumably players would not place marker blocks if they did not intend themselves or other players to view them on the map), the studies did not record eye movement or other metrics that would indicate how often the players glanced at the map, and so how much the dynamic map was utilized, and how much it would affect working memory, can only be conjectured.

The utility of the marker block is limited to two contexts: either the player sees them on the dynamic map, or the player sees them in the Minecraft environment. There is thus a plausible argument that players benefited from marker block placement by seeing the icons representing the marker blocks appear on the dynamic map. This would enable players to keep track of what rooms contain victims and which do not, as well as what the status of those victims are, enabling players to "offload" some amount of information regarding the location and status of victims onto the dynamic map and marker blocks.

Complicating this picture, however, is internal research conducted by CMU which found the dynamic map had limited efficacy [12]. With the intent to determine "whether the use of real-time information aids can improve performance by improving the participant's strategic approach", experimenters analyzed the effectiveness of adding a real-time dynamic map midway through a Search and Rescue task performed within the Minecraft Environment. The intent of the intervention was explicitly to allow cognitive offloading and "reduce the limitations of working memory (attentional load) and reduce computational effort (the strategic approach)".

However, results contradicted initial expectations. The report found no evidence that the dynamic map intervention improved performance compared to a control. Furthermore, while it was expected that the dynamic map would moderate the effects of low pre-intervention performance, the opposite effect was found: players performing poorly prior to the intervention performed worse after the addition of the intervention, and players performing well prior to the intervention performed better after the intervention, ultimately “canceling” out each other and producing no effect overall.

Caveats should be taken with these results. The first was that this dynamic map was placed within the Minecraft interface, rather than adjacent to it, as in the case of the Study 3 Dynamic Map. This may, indeed make it easier for a player to utilize the dynamic map, as they must only look at part of the Minecraft window, rather than look away to a separate window. The second is that this experiment involved only a single participant at a time; other players did not populate the dynamic map with information which the participant may not possess. Third is that this dynamic map recorded the location and trajectory of the player, not information such as victim placement or location. In other words, players could not store information about the location of victims and did not need to, nor did they have other players whom they could share information with.

Still, they throw into doubt the actual efficacy of the dynamic map and raise questions about why these marginal and even negative effects occurred. One possibility is that contrary to a GPS, the dynamic map – in both the internal study and in Study 3 – does not rotate with the perspective of the player, nor does it indicate which direction the player is facing relative to the map. This could confuse players, particularly players with lower spatial ability and hence, supposedly, reduced ability to navigate with the help of a map. Another possibility is that the dynamic map adds to the cognitive load of players, particularly those inexperienced at Minecraft. Thus, while players experienced at Minecraft would be able to make proper use of it, inexperienced players would be further overwhelmed by the additional stimuli. Consequently, the confounding effect of the dynamic map is not clear.

7.2 Interpretation of Results

7.2.1 Total Distance Travelled by Transporter

Against my initial expectations, I found that total distance traveled by the Transporter was strongly correlated with final score – the more the Transporter traveled, the higher their score tended to be. Rather than an indicator of inefficiency, total distance travelled suggests a Transporter spending more time utilizing its greater speed to discover and evacuate victims.

Also against my initial expectations, I found that fewer Phases per evacuation (a more subtask-focused strategy) was associated with greater distance traveled, and more Phases per evacuation (a more instance-focused strategy) tended to correlate with lower overall distance.

One interpretation of the above is that the costs of context switching too frequently – in particular, the time to move between a region of the map with undiscovered victims and a region with discovered and triaged victims – outweighed the benefits of a low cognitive load.

7.2.2 Percent of Evacuations Done By Transporter

A larger number of Phases (higher instance-focused) corresponded with a decrease in the percent of evacuations done by the transporter. One possible explanation is that more Phases also meant the Transporter was engaged in activities other than Discovery and Evacuation, such as assisting the Medic in triaging critical victims. However, when comparing Phase count and Phase-Evacuation Ratio with percent of Medic assists, no relationship was found.

When considering the inverse relationship between Phase-Evacuation Ratio and Transporter-Medic distance, this suggests that the Transporter and Medic spent more time in proximity with one another, which may increase the degree that the Medic shared evacuation responsibilities with the Transporter.

7.2.3 Phase-Evacuation Ratio v. Evacuated-Discovered Ratio

The mixed results and low level of fit when comparing the Phase-Evacuation Ratio with the Evacuation-Discovered Ratio suggest inconclusive evidence in support of my hypothesis that an increase in the number of Phases would increase overall accuracy. Assuming that hypothesis is not simply wrong, possible explanations are that the overall task was not difficult enough, and cognitive load not high enough, for errors to occur on a significant level–a larger map with more victims may have made an effect. It is also possible that the

Dynamic Map had an influence on this result; however, as previously discussed, the effect of the Dynamic Map is not clear.

7.2.4 Cognitive Load

The substantial negative relationship between Phase-Evacuation Ratio and cognitive load (and the corresponding positive relationship with probability of forgetting) is in keeping with my hypothesis that greater number of Phases (and thus a more instance-focused strategy) would enable one to reduce overall working memory load by enabling participants to resolve outstanding victim evacuation tasks whose information would otherwise need to be retained within working memory. This is further supported by results showing how more Phases occurring per evacuation (high instance-focus) led to a smaller time between when a victim was discovered and when that victim was evacuated – the longer it takes to evacuate a victim, the longer that victim would have to remain in working memory.

However, this cognitive load did not show a substantial impact on actual performance. In addition to there being no relationship between Phase count and percent of forgotten victims, the majority of victims who were forgotten had not been triaged, meaning the Transporter was not the last point of failure for the evacuation of these victims, precluding the use of the number of forgotten victim as a metric for Transporter performance.

7.2.5 Marker Block Placement

Transporters placed marker blocks substantially more frequently than Medics, which in turn placed marker blocks substantially more frequently than Engineers. This meets my general expectations; the Transporter, explicitly tasked and best equipped to explore the map, would likely place more marker blocks so as to demarcate areas of the map that are and are not useful for the Medic and Engineer to access.

One possible interpretation of the decrease in marker block placement as the Phase-Evacuation Ratio increased (indicating a more instance-focused strategy), as previously discussed, is that utilization of marker blocks and/or the dynamic map allowed players to "offload" the burden of working memory, reducing the risk of the decay of victim location contents in working memory negatively impacting overall performance, because one could simply glance at the dynamic map to see where a victim is. Using marker blocks less, by contrast, would increase the burden on working memory, requiring more Phase switches to mitigate the risk of working memory decay negatively impacting performance.

7.2.6 Practice Effect

While the number of Phases tended not tended not to change between the first and second trials for a given team, players tended to evacuate more players within the same number of Phases, and accordingly saw a performance increase.

One possible explanation for this is that as players become more competent and familiar with the environment, they become more efficient at finding and evacuating victims within the same number of Phases. However, if we assume number of Phases to be mediated by cognitive load and cognitive load capacity, the lack of change in the total number of Phases would imply that cognitive load capacity remains relatively static.

Combined with the results finding a positive relationship between competency task completion time and Phase-Evacuation Ratio (but an insubstantial relationship with Phase count), this suggests that competency had little effect on the total number of Phases, but a more substantial effect on the efficiency of those Phases.

7.3 Overview of Results

Reviewing my hypotheses, I found evidence that performance of subtask A – Discovery increased instance-relevant information in memory, and that performance of subtask C – Evacuation – decreased instance-relevant information in memory. However, I did not find evidence that more frequent switching between subtasks A and C were associated with better performance – indeed, I found evidence pointing to the opposite conclusion.

All results were of relatively limited significance, with at most twenty percent of the data being explicable by the Transporter's use of Phases. Because of the confounding factors of the previously-discovered limitations, the overall trend of the data provides a suggestive, but not conclusive, idea of a more complicated relationship between multitasking and productivity than which is typically discussed in the experimental literature.

Table 7: Summary of Results

Variable	**Phase Count**	**Phase-Evacuation Ratio**
Final Score	Positive, $R^2 = 0.11$	Inverse, $R^R2 = 0.11$
Marker Block Placement	Inverse, $R^2 < 0.01$	Inverse, R^2=0.10
Total Transporter Travel Distance	Positive, $R^2 = 0.01$	Inverse, $R^2 = 0.24$
Evacuated-Discovered Ratio	Positive, $R^2 = 0.08$	Inverse, $R^2 = 0.06$
Average Cognitive Load	Inverse, $R^2 = 0.18$	Inverse, $R^2 = 0.22$
Average Probability of Forgetting	Inverse, $R^2 = 0.16$	Inverse, $R^2 = 0.20$
Percent of Medic Assists Performed by Transporter	Inverse, $R^2 < 0.01$	Positive, $R^2 = 0.06$
Competency	Inverse, $R^2 = 0.03$	Positive, $R^2 = 0.11$
Transporter-Medic Distance	Inverse, $R^2 = 0.11$	Inverse, $R^2 = 0.13$
Percent of Evacuations Performed by Transporter	Inverse, $R^2 = 0.11$	Inverse, $R^2 = 0.20$

Several analyses yielded weak or no results. In particular:

- No meaningful relationship was found between Phase-Evacuation Ratio or Phase Count and Santa Barbara Sense of Direction (SBSOD) survey responses.
- No meaningful relationship was found between Phase Count and Phase-Evacuation Ratio with percent of Medic Assists

- No meaningful relationship was found between agent/ASI condition and Phase Count or Phase-Evacuation Ratio.

Combined, this suggests that increased Phase-switching (i.e. a more instance-focused strategy) negatively impacts performance but reduces overall working memory burden by limiting the overall number of open task instances the Transporter has to keep track of. The probability of error measurement suggests that fewer Phases per evacuations (a more subtask-focused strategy) should increase the likelihood of error; however, that did not occur in the data at a significant level. This is likely because the players, but not the probability of forgetting measurement, accounted for the dynamic map, using it as a kind of cartographic checklist to track where victims were and whether they had been triaged or evacuated. This is supported by an increased use of marker blocks being correlated with final score ($R^2 = 0.11$) as well as with the ratio of Phases to evacuations increasing as marker block usage decreased, implying less utilization of the dynamic map.

One question worth asking is whether the absence of the dynamic map would produce a substantial change in results. I hypothesize that this would be the case, and that the lack of a dynamic map would force players to rely more on working memory and thus increase the pressure of a high cognitive load on the players. Whether this would be enough to substantially impact error at the cognitive load levels measured, however, is unknown.

Regardless, these results match overall with existing literature on multitasking, finding that more frequent task switching decreases overall performance. However, it adds wrinkles into that research.

7.4 Multiple Task Axes

Controlling for evacuation count, fewer Phases (more subtask-focused) correlated with greater Transporter distance but higher cognitive load. The opposite was true with more Phases. In other words, fewer Phases meant greater *logistical* efficiency but lower *cognitive* efficiency. The cost in cognitive efficiency was not realized with a corresponding decrease in performance in these experiments. However, though a performance cost via cognitive cost not realized, it is plausible that such a cost could exist in a different scenario.

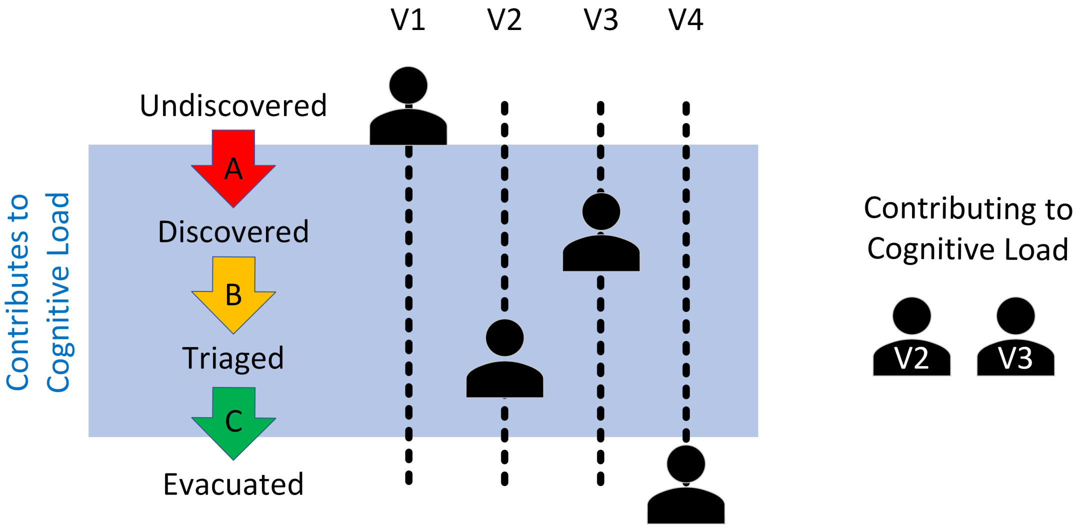


Figure 41: Illustration of Example Transporter State

Consider a scenario where the Transporter has only four victims: V1, which is undiscovered. V2, which has been triaged by the Medic. V3, which the Transporter has just discovered. And V4, which the Transporter previously evacuated.

Because V3 has not yet been triaged, the Transporter, though in close proximity to V3, cannot do anything else with it right now. We will consider simply waiting for the Medic to be a non-option due to the extended idle period that would require. Thus, the Transporter is confronted with two different choices:

- Continue searching for remaining victims (in this case, V1)
- Move to evacuate V2, whose location is known and is ready to be evacuated.

While this can be conceptualized simply as a task switch, it can also be considered in the sense that the Transporter is engaging in two concurrent types of tasks – which I will call "Task Axes".

One task axis is the "Shared Subtask" task axis, where the Transporter continues the discovery subtask. On this axis, the Transporter is performing the same sequence of actions but accruing additional task instances – it is monotasking on the subtask, but multitasking on task instances.

The other task axis is the "Shared Instance" task axis, where the Transporter evacuates known victims but has to switch subtasks. Here, the multitasking is from the discovery to evacuation subtask (and to some degree switching focus to a different task instance), but the Transporter is resuming focus on existing task instances and reducing the number of task instances currently being managed.

In either case, multitasking is occurring, but along different task axes. In each case of axis-switching, a multitasking cost is incurred, but the cost differs depending on the axis.

Assuming that V3 is in close proximity to an undiscovered region of the map, it would take little time to continue to search this region for V1. By contrast, if V2 is a substantial distance away, the Transporter would need to go to V2, pick up and move V2 to an evacuation zone, and then return to the undiscovered region to resume the search. In the short term, choosing to search the undiscovered area near to where the Transporter already is, is a logistically superior choice.

However, if the Transporter finds V1, then V1 will be added to the Transporter's overall cognitive load, which V2 and V3 are already contributing to. In the current experiment, where a dynamic map can be used to keep track of the victims, this cognitive cost may be trivial. However, if there is no dynamic map, or if, instead of 2 victims currently contributing to cognitive load there are 20, the cognitive cost becomes more pronounced – there would already be a high number of victims that the Transporter would need to keep track of, and there is likely already a high risk of the Transporter misremembering the locations of victims, misremembering their states, and miscalculating the most optimal path to evacuate them.

As previously stated, this A-B-C task configuration has parallels in many modern workplace tasks, and the optimal rate of axis-switching depends on configurations unique to the environment as well as to the person. For example, cognitive load as measured in this exper-

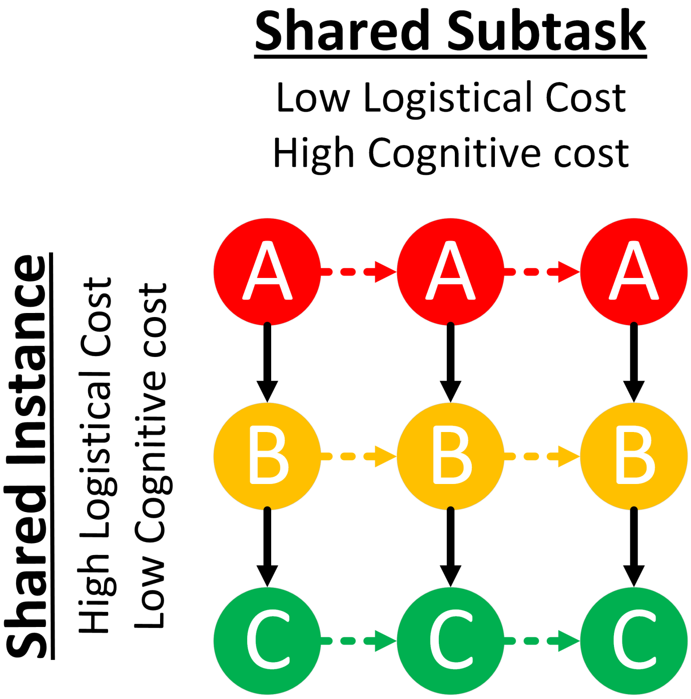


Figure 42: Multiple Task Axes with their associated costs

iment does not take into account the individual's overall working memory capacity, which may vary from person to person.

Similarly, parameters of the task environment may also matter. In this experiment, optimal behavior may depend on how many victims there are, how dense or sparse the victims are within the environment, and whether the dynamic map is present or absent. This is compounded by additional factors: What information is needed for a task instance, and how much? How sensitive is the task to cognitive error? How much time does a third party need to complete the intermediate subtask separating its initiating and concluding subtasks?

One could argue that assigning the same individual to complete two different subtasks is itself inefficient, and that efficiency is improved by delegating different individuals to complete the different subtasks. However, an issue pops up when information is shared between the subtasks. In that case, the first party must communicate that information to the party performing the concluding subtask, or to the party performing the intermediate subtask, who must then communicate that information to the party performing the concluding subtask. That may be less practical, depending on the situation, than the first party simply doing it themselves.

To return to the waiter example, one could certainly attempt to streamline restaurant productivity by delegating one person to receiving the orders, and another to delivering the orders (and this is often done in fast food restaurants). However, part of what draws people to more traditional restaurants is often the establishment of a personal relationship (or at least the illusion of one) between the server and the serviced. The waiter must not only know the location of the table they're serving, but who ordered what, as well as the personalities and overall disposition of the customers and how their needs change over the course of the meal. The customer, in turn, may appreciate the familiarity of the same person attending to their needs throughout their meal, and may feel more disconnected (to the disadvantage of the restaurant) if a different person appears to provide a different sub-service to the customers. In this circumstance, it is not only the raw information that is retained by the waiter but nuanced inter-personal information which cannot be easily transmitted to another worker.

Similarly, in a tech support situation, many a customer has been frustrated at being transferred to another agent, with slightly different jurisdiction and expertise, and having to re-explain their problem, and often be asked many of the same questions they were asked by the previous agent. And it seems questionable whether having a manager oversee the initiation of a subordinate's task, and a different manager overseeing the conclusion of that same subordinate's task, is adding anything except for an additional layer of bureaucracy.

Multitasking is unavoidable in many situations simply because it is more efficient to retain the same individual across multiple subtasks–even if those subtasks are staggered by time and other unrelated tasks–than to delegate different parties to managing each subtask and having to communicate the relevant information between each.

7.5 Avenues for Further Research

If these results prove valid, they would have implications for a number of fields. To give a few examples:

- Workplace management, in deciding division of labor, in particular how much the labor should be divided before they begin to suffer consequences
- In a psychological context, the question of where this tradeoff between working memory load and productivity in managing multiple tasks occurs, and how to measure it
- In AI theory of mind, which ASIST was investigating, the question of how and when an AI advisor might request a task switch to a player.

The noise in the data precludes drawing strong conclusions. However, the data is nonetheless suggestive of a relationship between the degree of Phase-switching and working memory load, as well as of behaviors that are themselves related to working memory load. Stronger conclusions would require experimental conditions that model the task configuration previously described and which eliminate confounds such as the dynamic map and the presence of advisory agents.

One possible such experiment could even maintain the Minecraft setup, but with the following modifications

- The dynamic map is eliminated
- Agent interventions are eliminated
- All victims are worth the same number of points
- The Engineer role is eliminated
- The Transporter, rather than placing marker blocks, instead signals the location to the Medic via a button press.
- The Medic role is static and programmatically controlled, rather than human controlled, and is not a physical presence on the map. Once the Transporter reports on the location of a victim, a timer elapses at the end of which the victim is considered triaged.

Such an experimental setup would eliminate or substantially reduce the most considerable confounds, allowing a more thorough investigation of the relationship between Phase-switching and working memory load.

8.0 Conclusion

The evidence reviewed here is far from conclusive, and merely points in a direction that has gone relatively uninvestigated in the field of multitasking research. And yet point it still does, suggesting that multitasking may have more reasons for its prevalence in modern life than contemporary experimental research literature suggests.

Multitasking, among other reasons, is necessitated in some circumstances by an inability of the individual to complete a task in a single sitting, particularly when a task must be passed off to another party for an extended period of time. Under such circumstances, an individual must take on additional tasks or risk wasting productive time.

What this data suggests is that performance of the same type of subtask, while being generally more efficient than switching between two types of subtasks, can gradually increase the burden on working memory, while switching between those subtasks and a subtask which *eliminates* the amount of information that must be tracked in working memory, reduces that burden.

While the presence of confounds in this experimental data (most notably the dynamic map) plausibly prevented higher cognitive loads from taxing the player, the limited capacity of working memory are well-documented, and cognitive load cannot increase indefinitely. The use of task-switching suggests a possible means by which an individual may manage that cognitive load, trading an immediate cognitive inefficiency for a longer-term benefit in preventing a cognitive load that would make the individual more prone to errors. Such an explanation could plausibly explain the prevalence of multitasking in the work place, and suggest why multitasking is unlikely to vanishing from the modern workplace in the near future.

Bibliography

[1] ACT-R. Publications and models. `http://act-r.psy.cmu.edu/publication/`.

[2] Rachel F. Adler and Raquel Benbunan-Fich. Juggling on a high wire: Multitasking effects on performance. *International Journal of Human-Computer Studies*, pages 156–168, 2012.

[3] John R. Anderson. *Rules of the Mind*. Erlbaum, 1993.

[4] Sinan Aral, Erik Brynjolfsson, and Marshall W. Van Alstyne. Information, technology and information worker productivity: Task level evidence. 2011.

[5] Catherine M. Arrington and Gordon D. Logan. The cost of a voluntary task switch. *Psychological Science*, pages 610–615, 2004.

[6] R. Bendell, J. Williams, S.M. Fiore, and F. Jentsch. Towards artificial social intelligence: Inherent features, individual differences, mental models, and theory of mind. *Advances in Neuroergonomics and Cognitive Engineering*, 2021.

[7] Emma Beuckels, Guoquan Ye, Liselot Hudders, and Veroline Cauberghe. Media multitasking: A bibliometric approach and literature review. *Frontiers in Psychology*, 2021.

[8] Arnaud Boutin and Yannick Blandin. On the cognitive processes underlying contextual interference: Contributions of practice schedule, task similarity and amount of practice. *Human Movement Science*, pages 910–920, 2010.

[9] P. W. Burgess. *Real-World Multitasking from a Cognitive Neuroscience Perspective*, pages 465–472. MIT Press, 2000.

[10] T Buser and N Peter. Multitasking. *Exprimental Economics*, pages 641–655, 2012.

[11] DARPA. Artificial social intelligence for successful teams (asist) proposers day (archived). `https://www.darpa.mil/news-events/artificial-social-intelligence-for-successful-teams-proposers-day`, 2019.

[12] Fade Eadeh, Anita Woolley, Cleotilde Gonzalez, Henny Admoni, Ngoc Nguyen, and Pranav Gupta. Hypothesis/capability pre-registration: Asist fall 2020, cmu ta2. 2020.

[13] Joshua Elliott. Artificial social intelligence for successful teams (asist). `https://www.darpa.mil/program/artificial-social-intelligence-for-successful-teams`.

[14] J. Engstrom, M. L. Aust, and M. Vistrom. Effects of working memory load and repeated scenario exposure on emergency breaking performance. *Human Factors: The Journal of Human Factors and Ergonomics Society*, pages 551–559, 2010.

[15] Victor M Gonzalez and Gloria Mark. Constant, constant, multi-tasking craziness: managing multiple working spheres. pages 113–120. Association for Computing Machinery, 2004.

[16] Mary Hegarty, Anthony E Richardson, Daniel R Montello, Kristin Lovelace, and Ilavanil Subbiah. Development of a self-report measure of environmental spatial ability. *Intelligence*, 30(5):425–447, 2002.

[17] C. Kapadia and S. Melwani. More tasks, more ideas: The positive spillover effects of multitasking on subsequent creativity. *Journal of Applied Psychology*, pages 542–559, 2021.

[18] Diwas Singh KC. Does multitasking improve performance? evidence from the emergency department. *Manufacturing Service Operations Management*, pages 168–183, 2013.

[19] Glenn J. Lematta, Christopher C. Corral, Verica Buchanan, Craig J. Johnson, Anagha Mudigonda, Federico Scholcover, Margaret E. Wong, Akuadasuo Ezenyilimba, Manuel Baeriswyl, Jimin Kim, Eric Holder, Erin K. Chiou, and Nancy J. Cooke. Remote research methods for human–ai–robot teaming. *Human Factors and Ergonomics in Manufacturing Service Industries*, pages 133–150, 2022.

[20] Assar Lindbeck and Dennis J. Snower. Multitask learning and the reorganization of work: From tayloristic to holistic organization. *Journal of Labor Economics*, pages 353–376, 2000.

[21] M. C. Lovett, L. M. Reder, and C. Lebiere. *Modeling working memory in a unified architecture: An ACT-R perspective.*, pages 135–182. Cambridge University Press, 1999.

[22] Stephen Monsell. Task switching. *Trends in Cognitive Sciences*, pages 134–140, 2003.

[23] Harold Pashler. Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, pages 220–244, 1994.

[24] Phil Pavlik and John Anderson. An act-r model of the spacing effect. 2003.

[25] Y. Qin, D. Bothell, and J. R. Anderson. Act-r meets fmri. *Web Intelligence Meets Brain Informatics*, pages 205–222, 2006.

[26] B.C.W. Ralph, P. Seli, and K.E. Wilson. Volitional media multitasking: awareness of performance costs and modulation of media multitasking as a function of task demand. *Psychological Research*, pages 404–423, 2020.

[27] V.J. Rideout, U.G. Foehr, and D.F. Roberts. Generation m2: Media in the lives of 8-to 18-year-olds. Technical report, 2010.

[28] Frank E. Ritter, Farnaz Tehranchi, and Jacob D. Oury. Act-r: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2018.

[29] J. S. Rubinstein, D. E. Meyer, and J. E. Evans. Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, pages 763–797, 2001.

[30] Srna S, Schrift RY, and Zauberman G. The illusion of multitasking and its positive effect on performance. *Psychological Science*, pages 1942–1955, 2018.

[31] Dario D. Salvucci and Niels A. Taatgen. Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, pages 101–130, 2008.

[32] D. L. Strayer, S. C. Castro, J. Turrill, and J. M. Cooper. The persistence of distraction: The hidden costs of intermittent multitasking. *Journal of Experimental Psychology*, pages 262–282, 2022.

[33] J. Gregory Trafton, Laura M. Hiatt, Anthony M. Harrison, Franklin P. Tamborello, Sangeet S. Khemlani, and Alan C. Schultz. Act-r/e: an embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, pages 30–55, 2013.

[34] Y. Wang, N. Gurney, J. Zhou, D. V. Pynadath, and V. Ustun. Neural heuristics for route optimization in service of a search and rescue artificial social intelligence agent. 2021.

[35] JI Westbrook, MZ Raban, and SR Walter. Task errors by emergency physicians are associated with interruptions, multitasking, fatigue and working memory capacity: a prospective, direct observation study. *BMJ Quality Safety*, pages 655–663, 2018.

[36] Jacob Whitehill. Understanding act-r - an outsider's perspective. 2013.