AN EXPLORATION OF DISTRIBUTED PARALLEL SORTING IN GSS

by

Christopher B.R. Diller

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF MANAGEMENT

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN MANAGEMENT INFORMATION SYSTEMS

In the Graduate College

THE UNIVERSITY OF ARIZONA

2013

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Christopher B.R. Diller, titled “An Exploration of Distributed Parallel Sorting in GSS” and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Date: November 15, 2013
Jay F. Nunamaker, Jr.

Date: November 15, 2013
Paulo B. Goes

Date: November 15, 2013
Joseph S. Valacich

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: November 15, 2013
Dissertation Director: Jay F. Nunamaker, Jr.

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Christopher B.R. Diller

ACKNOWLEDGEMENTS

I would like to extend my most sincere thanks to Joel H. Helquist and John Kruse for all their selfless effort in support of this project. I couldn’t have done this without you.

I also want thank Nathan W. Twyman, Thomas O. Meservy, Mark W. Patton, and William T. Neumann for everything they have done to help me toward the successful completion of my degree. A man never stands as tall as when he stoops to help another in need. Thank you for lifting me up as you have.

Finally, I want to thank my committee members for their dedication to me throughout my time at the University of Arizona. Thank you for not giving up on me and allowing me to finish what I started.

DEDICATION

I have always believed that every living human should thoughtfully consider the lives of their family and their ancestors on a daily basis. Families will endure and thrive only if we recognize and learn from the sacrifices made, the tribulations endured, and the victories won by our family members in their struggle to survive and succeed – as each living person is the cumulative result of all that effort. It is the solemn duty of the living to honor their ancestors with work that benefits their posterity and the family's future generations, in turn.

To that end, I dedicate this small work to all of the members of my family – past, present, and future.

I have always been supremely proud of my immediate family – Almyra, Richard, John, Kelly, Jenny and Emily. I deeply love you all and appreciate everything you have done to support me throughout my lifetime. However, as a result of my tireless attempts to meaningfully contribute to our family legacy, I feel as though the decisions I made unfortunately have hurt you the most. Because you know me so well, I am certain that you understand… but I wish that I had been a better son, brother, and father to you.

Wasting a talent is a terrible sin, and I pray that my Creator will judge my life in a better light than my results might indicate. I tried to use my talents in the best way I knew how, and tried to help everyone I could whenever I could… I just wasn't able to make the lasting impact I believe my family's legacy deserves. I do hope that my life can serve as a cautionary tale to our descendants – live a noble and productive life, put God first and others before yourself, but always be sure to count the cost.

Dad, I'm sorry. I've tried my best. I'll keep trying to be a man of honor, as you were. I've missed you so.

TABLE OF CONTENTS

LIST OF TABLES 9

LIST OF FIGURES 10

ABSTRACT 11

1 Introduction 12

1.1 Purpose of the research 15

2 Achieving convergence in collaborative environments 17

2.1 The facilitation bottleneck 18

2.1.1 The characteristics of a skilled facilitator 19

2.1.2 The costs of a skilled facilitator 20

2.2 The importance of backchannel feedback 21

2.3 The next generation of collaboration methods and design 22

2.3.1 ThinkLets 22

2.3.2 Participant-driven GSS (PD-GSS) 23

2.4 Cognitive load 26

2.4.1 The measurement of cognitive load 29

2.5 Normalized clustering error (NCE) 31

3 Methods 34

3.1 Research questions and hypotheses 34

3.2 Independent variables and experimental treatments 35

3.3 Dependent variables and data collection 36

3.3.1 Measuring sort effectiveness via normalized clustering error (NCE) 37

3.3.2 Measuring time in a parallel task 38

3.3.3 Measuring cognitive load via the NASA-TLX method 38

3.4 Experiment details 40

3.4.1 Research setting 40

3.4.2 Experimental task overview 40

3.4.3 Participants 41

3.4.4 Experimental procedures and administration 43

4 Analysis and results 45

4.1 Descriptive statistics of sort quality/effectiveness (NCE) 45

4.2 Descriptive statistics of activity completion times 46

4.3 Comparing cognitive load across treatments 47

4.3.1 Comparing cognitive load (Adjusted TLX) across treatments 48

4.3.2 Comparing perceived task difficulty across treatments 49

4.3.3 Comparing self-reported comfort levels across treatments 52

4.3.4 Comparing self-reported fatigue levels across treatments 54

4.4 Comparing task completion times across treatments 56

4.5 Comparing sort effectiveness (NCE) across treatments 59

5 Discussion 62

5.1 Limitations 64

5.2 Future research 65

5.2.1 Measuring cognitive load in collaboration environments 65

5.2.2 Designing experimental treatments for convergence activities 66

5.3 Practical recommendations 68

5.3.1 The “sort-your-own” method of facilitating convergence 68

5.3.2 Technology-enabled sort aggregation to generate a “consensus sort” 70

6 Conclusion 71

APPENDIX A – DAILY EXPERIMENT OPERATIONS GUIDE 72

APPENDIX B – SCRIPT FOR SUBJECT QUALIFICATION 74

APPENDIX C – PARTICIPANT TASK CHECKLIST 75

APPENDIX D – PARTICIPANT WORKSTATION DESKTOP 76

APPENDIX E – PRE-EXPERIMENT SURVEY QUESTIONS 77

APPENDIX F – POST-EXPERIMENT SURVEY QUESTIONS 80

APPENDIX G – TRANSCRIPT OF INSTRUCTIONAL VIDEO 83

APPENDIX H – COMPLETE LIST OF ITEMS TO BE SORTED 86

APPENDIX I – ANOVA CONTRASTS OUTPUT (RAW TLX) 90

APPENDIX J – ANOVA CONTRASTS OUTPUT (ADJ. TLX) 93

References 95

LIST OF TABLES

Table 1: Descriptive statistics of the experiment's participants (by treatment condition) 42

Table 2: Descriptive statistics of subgroup sort quality/effectiveness (NCE) 45

Table 3: Completion times (in seconds) for key experimental tasks 46

Table 4: Descriptive statistics of the subgroups averaged Raw TLX scores (by condition) 48

Table 5: Descriptive statistics of Adjusted TLX measures (by condition) 49

Table 6: Descriptive statistics of self-reported task difficulty (by condition) 50

Table 7: Results of ANOVA (one-way) contrasts on task difficulty (by condition) 51

Table 8: Descriptive statistics of pre- and post-experiment comfort levels (by condition)......... 52

Table 9: Results of paired-samples t-test on self-reported comfort levels (by condition)........... 53

Table 10: Descriptive statistics of pre- and post-experiment fatigue levels (by condition) 54

Table 11: Results of paired-samples t-test on self-reported mental fatigue (by condition)......... 56

Table 12: Descriptive statistics of sort task completion time (by condition) 57

Table 13: Results of ANOVA (one-way) contrasts on sort completion time (by condition) 58

Table 14: Descriptive statistics of the NCE values (by treatment condition)............................. 60

Table 15: Results of ANOVA (one-way) contrasts on sort effectiveness (by condition) 61

Table 16: Summary of Hypothesis Support 62

LIST OF FIGURES

Figure 1: A timeline of user satisfaction levels during typical collaborative meetings............... 13

Figure 2: An illustration of the desired effects of the research.. 16

Figure 3: How normalized clustering error is calculated 32

Figure 4: An illustration of the basic experimental design of the study...................................... 36

ABSTRACT

When the members of a group work collaboratively using a group support system (GSS), they often "brainstorm" a list of ideas in response to a question or challenge that faces the group. The satisfaction levels of group members are usually high following this activity. However, satisfaction levels with the process almost always drop dramatically when the group is forced to sort, distill, or classify all of the brainstorming feedback in a synchronous, serially-conducted activity, held immediately after the brainstorming activity. Past explanations for the drop in satisfaction often point to the increased time required to complete a sort and to the mental difficulty in sorting large lists (i.e., increased "cognitive load"). The experiment conducted in this study was designed to expose the participants to conditions featuring different levels of cognitive demand, achieved by varying the number of items to be sorted. This design simulates an asynchronous method of sorting group feedback – a process that can be viewed as a "distributed parallel sort." This dissertation explores methods for measuring the cognitive load experienced by a participant during a sorting activity (using the NASA Task Load Index), evaluating the effectiveness of having group members sort partial lists of items instead of working synchronously on the same full list (objectively measured using normalized clustering error against a "gold standard" result), and proposes new methods for mitigating the drop in satisfaction levels that regularly occur in these collaborative settings without compromising the effectiveness of their sorting results. The experimental results imply that an individual's perceived difficulty of the task may rely on other factors, rather than just the length of a list. The results also imply that the NASA-TLX framework to measure cognitive load may need to be refined further (or implemented differently), if it is to be used in GSS research contexts. Finally, two methods are proposed (a facilitation-based recommendation and another technology-enabled option) that may help to mitigate the drop in satisfaction levels, improve a group's effectiveness, and reduce the time required for that group to effectively sort their feedback in collaborative GSS sessions.

1 Introduction

There are both tangible and intangible benefits associated with using group support systems (GSS) for group collaboration activities (de Vreede, Vogel, Kolfschoten, & Wien, 2003). Tangible benefits include reducing the amount of time and resources required to complete the meeting's requirements as compared with traditional, non-GSS supported meetings (Grohowski & McGoff, 1990). Other tangible benefits include such things as generating an increased number of higher-quality brainstorming ideas (Dennis & Valacich, 1990; Gallupe & Dennis, 1992).

The intangible benefits associated with GSS usage, however, are more difficult to quantify. These benefits include such things as improved problem definition, improvements in the level of group cohesion, and more group commitment to the solution (Nunamaker & Briggs, 1996).

Despite all of the benefits and process innovations touted by researchers over the past three decades of collaboration research, the biggest problem in collaborative work continues to be convergence – the process of sorting through a set of potential alternatives to find feasible solutions, and then building an adequate level of consensus among group members as a meeting progresses towards a final decision.

Many of the methods and tools prescribed in recent and historical literature show modest success in some applications, but to date, these successes have not been generalizable to all groups in all settings. This problem is exacerbated in asynchronous and/or distributed modalities, where a skilled facilitator is not present to moderate the group's effort (Briggs & de Vreede, 2003). However, even in facilitated collaborative sessions, groups find it difficult to build consensus and converge on a set of actionable solutions to the issues they are attempting to resolve.

When groups work collaboratively, their activities generally involve the following stages (Nunamaker & Dennis, 1991):

- Idea Generation ("brainstorming" a variety of potential solutions)

- Idea Organization ("sorting/classifying" the contributions into similar clusters/categories)
- Prioritizing (reviewing the strengths/weaknesses of the idea clusters and proposed solutions)
- Policy Development (deciding on the most appropriate alternatives to implement and assigning specific tasks, roles and directions to appropriate parties)

From a practical perspective, the convergence process appears to begin in the "idea organization" stage. Evidence of the difficulty of convergence can be seen simply by considering the satisfaction levels of group members during this time (see Figure 1), where those levels reach their lowest point (Chen, Hsu, Orwig, Hoopes, & Nunamaker, 1994). This is also the point where "free-riding" becomes more prevalent among group members, usually signaling a loss of confidence in the process or a reticence to continue on a course of action that they don't feel is meaningful (or worthy of their effort).

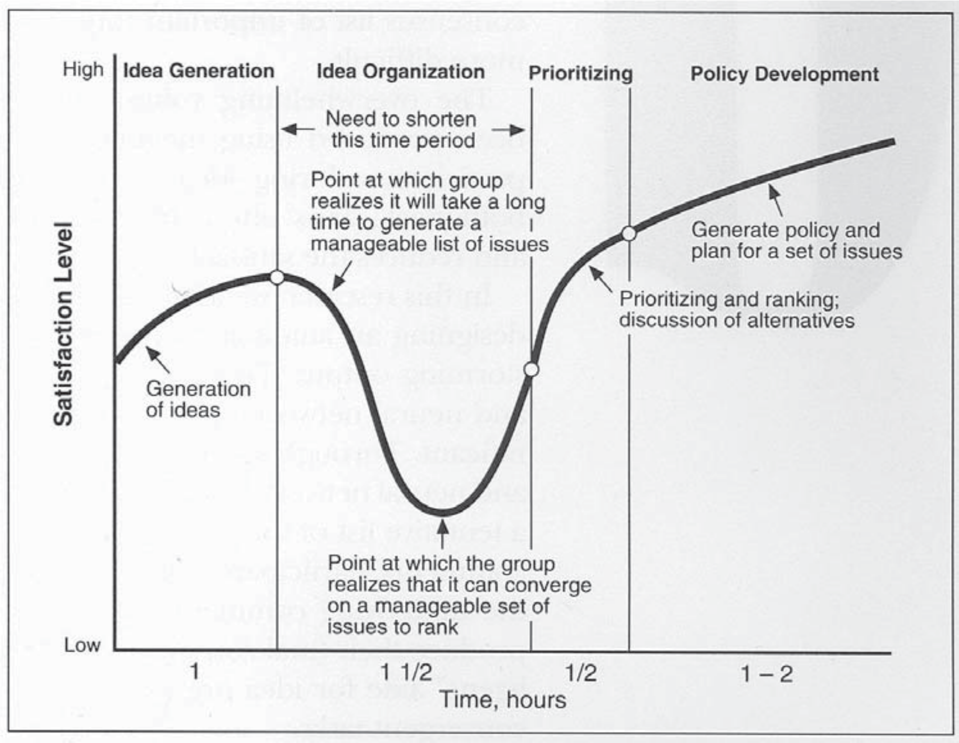


Figure 1: A timeline of user satisfaction levels during typical collaborative meetings

But the nature of the activities in each of the stages of collaboration is curiously related to member satisfaction levels. Idea generation, prioritizing, and policy development can all be viewed as "parallel" activities, meaning that group members are able to communicate simultaneously and work "individually" (or focus on a particular aspect of a complex task), while other group members are free to focus on other aspects. Research has shown that parallel communication is a function of GSS that often improves the efficiency of the group process (Dennis, George, Jessup, Nunamaker, & Vogel, 1988). As a result, during parallel activities, the satisfaction levels of GSS-enabled group members are more likely to increase over time.

However, the idea organization stage is often conducted as a serial activity, requiring the entire group to focus on the same aspect of the problem, in order to bring its collective wisdom to bear and leverage the group's effort on refining that aspect as they move forward in their search for an optimal solution. In contrast to parallel activities, serial activities like these often seem to cause member satisfaction levels to decrease.

Some of the explanations that have been posited to explain the decline in satisfaction levels observed during serial tasks state that:

- Users don't like receiving critical/negative feedback regarding their ideas;
- Users don't like seeing their contributions "diluted," or "lumped in" with other ideas;
- Users are intimidated by the cognitive difficulty of sorting large sets of feedback data; and
- Users' lose their individual enthusiasm during organization because the process is lengthy.

The first two explanations are rooted in the psychological concept of "evaluation apprehension," a well-studied phenomenon in human behavior (Cottrell, 1972; Rosenberg, 1965). Unfortunately, the effects of this phenomenon are not likely to be significantly changed with respect to changes in collaborative processes as it seems to be a deeply-rooted principle of human behavior.

The last two explanations, however, are rooted in the concept of “cognitive load” (Tarmizi & Sweller, 1988; Hilmer & Dennis, 2000). This represents an area of particular interest to collaboration research, as it can be manipulated much more easily than entrenched psychological characteristics like evaluation apprehension. With additional study, academic research may be able to identify methods to transform traditional serial tasks into parallel tasks, and configure them in a way to reduce the perceived cognitive load of the convergence process and its various tasks.

However, as Briggs (2003) notes, skilled facilitators are in short supply and are expensive to acquire – a fact that is still very true today. Historical research has also clearly demonstrated that the success of a collaborative group is often highly dependent upon the skill of the facilitator. As a result, Briggs suggests that future academic research in collaboration and group support systems should focus upon downplaying (or eliminating) the facilitator entirely and instead focusing on “collaboration engineering” and “thinkLets” (e.g., pre-defined automated tools or routines) that enable groups to benefit from the best practices of facilitation, without requiring a facilitator to be present.

1.1 Purpose of the research

This study is an attempt to investigate the scientific validity of the cognitive load-based explanations for the decline in satisfaction observed in a sorting task, with respect to the organizational effectiveness of their effort. It is also designed to simulate an asynchronous, distributed approach to an open sort task – by allowing a group’s members to process a data set in parallel (independent of other members).

The ultimate objective of the study is to identify methods and strategies that help reduce the time required to complete the idea organization phase of a collaborative meeting while maintaining higher satisfaction levels throughout the convergence process, which may enable groups to make better decisions faster (see Figure 2). Ideally, the end results would foster improved methods and strategies that could be applied to many collaborative environments (synchronous/asynchronous and co-located/distributed, regardless of whether the group is facilitated by a human or not), and also lead to more effective technology-based tools that enable participant-driven GSS (PD-GSS) success as well.

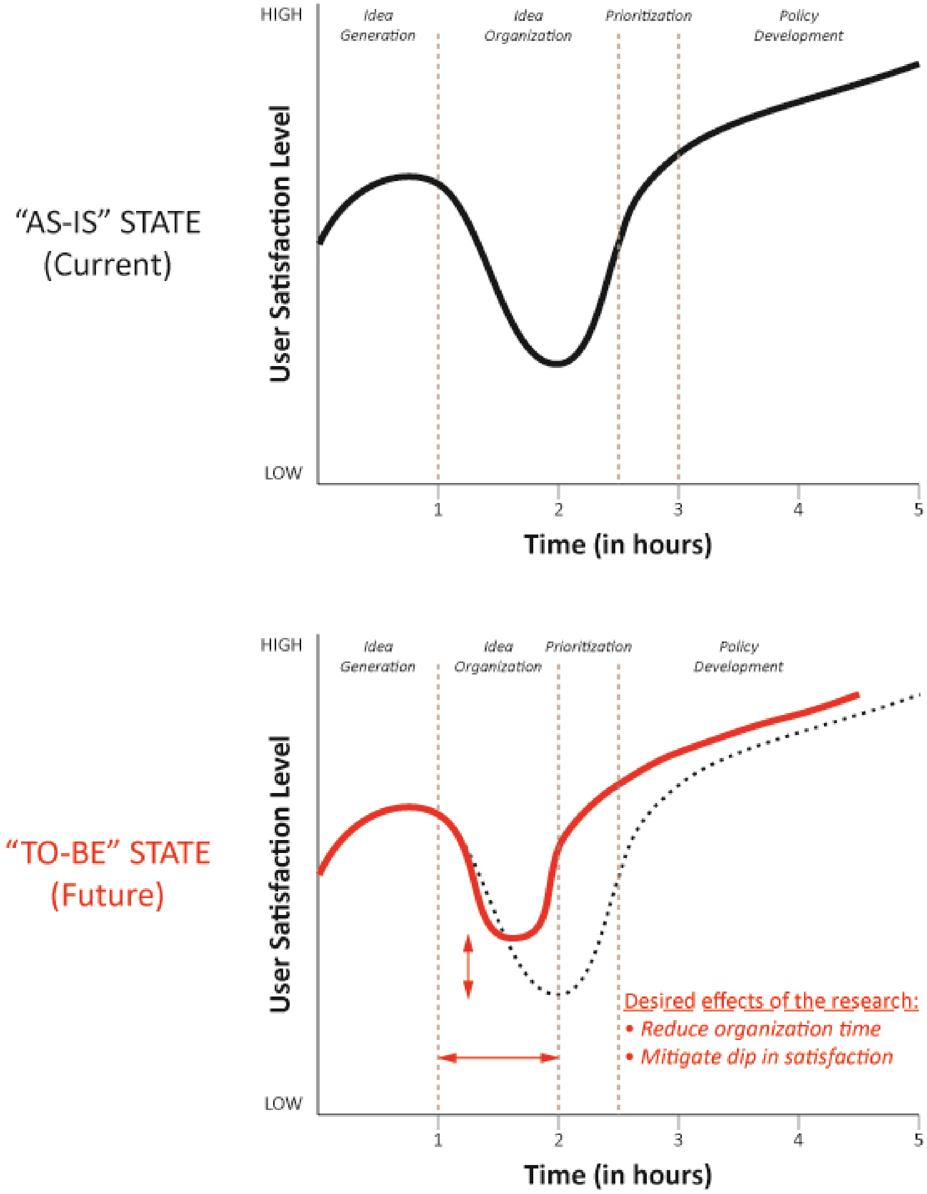


Figure 2: An illustration of the desired effects of the research

2 Achieving convergence in collaborative environments

A significant amount of research has been done on the value of collaborative divergence – the process of eliciting a greater number of quality ideas from a group, as this has been shown to directly improve the effectiveness of the group's work product (Briggs et al, 2003).

Past studies have provided effective guidance to help maximize divergence when conducting collaborative meetings by addressing matters such as:

- Meeting environment characteristics (e.g., seating arrangements for proximal groups, media "richness" for distributed or virtual groups, etc.) (Nunamaker & Briggs, 1996; Romano, Nunamaker, Briggs & Mittleman, 1999);
- Group size and composition (i.e., the inclusion of all stakeholders and/or parties affected by a particular problem) and the optimal number of participants required to achieve a goal (Valacich, Dennis & Nunamaker, 1992; Lowry, Roberts, Romano, Cheney & Hightower, 2006; Roberts, Lowry & Sweeney, 2006);
- Task design and agenda organization (i.e., developing more effective templates that can be used in typical situations) (Anson, Bostrom & Wynne, 1995; Briggs et al, 2003); and
- Participant motivation (i.e., keeping the group focused on the task at hand, as they move towards a desired outcome) (Romano et al, 1999; de Vreede, Boonstra & Niederman, 2002; Briggs et al, 2003).

However, the biggest problem in collaborative group work remains convergence (Briggs et al, 2003) – the process of distilling and organizing the brainstorming feedback, evaluating alternatives, obtaining consensus and helping a group quickly develop actionable plans to implement. These convergent tasks have proven to be much more problematic than brainstorming. They are often marked by reduced participant satisfaction, declining motivation levels, as well as increased frustration and dissent among participants as their ideas are evaluated (see Figure 1).

To combat the issues that often hinder convergence, prior research has also generated effective facilitation guidance for the leaders of collaborative meetings – methods and strategies that can help to:

- Enhance participant satisfaction (e.g., making sure "every voice is heard" by the group and that all ideas are given equal importance and due consideration) (Dennis, Haley & Vandenberg, 1996);
- Enhance participant motivation (i.e., keeping the group focused on the task at hand, as they move towards a shared desired outcome) (Briggs & Nunamaker, 2006); and
- Identify potential areas of compromise quickly and negotiate mutually-beneficial solutions (e.g., methods to achieve satisficing, persuasive leveraging of a tipping point, etc.).

2.1 The facilitation bottleneck

Prior research has consistently asserted that the skill of a human facilitator has a direct, positive effect on the effectiveness of a group and the quality of their work product (de Vreede et al, 2002; Briggs et al, 2003). A group led by a highly-skilled facilitator can work together more effectively and produce better results than a group led by an amateur facilitator via factors such as:

- Better meeting design and time management (i.e., the facilitator asks the right questions, in the right way, in the right order, and knows how to avoid wasting the group members' time);
- Improved meeting disposition levels (i.e., the facilitator can "keep things positive," quell negative behavior among participants, and promote individual satisfaction);
- Faster resolution of disputes (i.e., the skilled facilitator knows how to tactfully respond to dissent and deal with arguments when they arise);

- Impartial interpretation of the results (i.e., an impartial facilitator's objective view can help the group overcome emotionally-charged obstacles and/or mediate high-stakes issues); and
- More effective consensus-building (i.e., the skilled facilitator has a demonstrated ability to quickly consolidate and process all of the available information, as well as the ability to lead the group members to agreement on a minimal set of results).

2.1.1 The characteristics of a skilled facilitator

It is also important to note that a sizable component of a facilitator's skill is rooted in their language abilities, their experience with leading other collaborative groups, and their practical understanding of basic social psychological principles (Briggs et al, 2003). Many people mistakenly believe that a cursory knowledge of these complex concepts makes them capable of serving as a good facilitator – only to realize later (after a failed session) that these are the defining characteristic of what makes a skilled facilitator so valuable.

Language abilities play an enormous role in the ultimate success of a collaborative session. If the meeting's activities aren't carefully worded or presented to the participants, the session may not generate the desired results (Briggs, Crews & Mittleman, 1988). An old adage says: "You only get the right answer when you ask the right question." Skilled facilitators are able to write questions and provide instructions that precisely address the principal concerns of the group, all while using a carefully-chosen economy of words. One poorly-crafted sentence or improperly-worded question has the capability to completely derail a group's progress and lead them towards a sub-optimal result.

Furthermore, skilled facilitators rely upon their past experience to develop a sense for assessing the limits of a collaborative group session. They know how to structure an agenda to lead a group to its desired goal, without unnecessarily taxing the participants' limited resources (de Vreede et al, 2002). In other words, the skilled facilitator is able to predict the physical and psychological effects that every activity will have on the group, as the meeting progresses. Amateur facilitators often under-estimate the fatigue of certain collaborative activities, and over-estimate the group's stamina or ability to successfully navigate through a series of tasks. (In other words, the amateur facilitator is usually over-ambitious and demands too much from a group by preparing too many activities). These meetings usually fail to meet the expectations of the group, and the frustration grows as the session nears its end, manifesting itself in behaviors that emphasize expedience over quality. In the worst cases, the participants stage a "user mutiny," and erode the facilitator's implicit ability to control the group's actions. In contrast, a skilled facilitator would know exactly what activities to schedule, how to order those activities to minimize fatigue and discomfort, how to meet the goals of the session within a given time frame, and without losing control of the session.

2.1.2 The costs of a skilled facilitator

The benefits of a skilled facilitator always pose a significant cost to the group (Briggs et al, 2003). Historically, facilitated convergence activities tend to require a serial meeting structure. Where divergence activities can always be done in parallel (i.e., every participant can work independently to develop their own thoughts), facilitated convergence activities are performed in serial (i.e., in a step-by-step process performed one activity at a time). This, in turn, requires more time to complete the group's activities – effectively negating the benefits of time management mentioned earlier. In addition, facilitated activities tend to require the synchronous presence of a group's participants – in other words, the group needs to meet at the same time and be working on the same items. (Note that "synchronous presence" doesn't always imply physical proximity or a single meeting location. But working synchronously with a facilitator in a virtual/distributed environment will usually require the use of high-bandwidth, two-way conferencing technologies, which imposes a different technological cost on the group.) Finally, skilled facilitators are in short supply and their time is a prized commodity – there are expected costs associated with acquiring a facilitator that may rise in relation to their demonstrated skill and experience.

Regardless of the meeting modality (synchronous vs. asynchronous and proximal vs. distributed), the inclusion of a facilitator into a collaborative process can be viewed as inefficient, relative to the other alternatives, simply because of these higher costs. Much of the recent academic research in collaboration is aimed at removing the facilitator from the process. By eliminating a group's need for synchronicity, a central meeting location and/or teleconferencing technologies, and serial meeting task design, the convergent activities of the future may be conducted without a facilitator more quickly and cheaply than is possible today. Ideally, this would occur without any negative effect on the quality of the group's work product – but thus far, the majority of the proposed technological tools and methods for augmenting convergence without human facilitation have fallen short in that regard.

In sum, there remains a demonstrated need for additional research on collaborative convergence activities. The costs of achieving convergence will remain higher until such time as new facilitator-less technologies or methods are developed that meet or exceed the results of those generated with the presence of a skilled human facilitator.

2.2 The importance of backchannel feedback

Practical experience indicates that human facilitators (regardless of skill) usually find it more difficult to lead larger groups, especially those in distributed and/or asynchronous modalities. As group size increases, there is more participant feedback for the facilitator to attend to and more "voices" that expect to be heard.

Skilled facilitators are trained to be highly attuned to the nonverbal behaviors of their participants – they rely on "backchannel feedback" (a variety of unspoken communication cues, such as eye gaze, facial expressions and posture) to make judgments about their meeting participants. Facilitators continuously monitor this backchannel data to gauge the current mood/emotion of the group, determine interest levels of ideas, assess the quality of proposals, and make judgments on the timing and progression of the meeting's tasks.

Critically important in this regard is the concept of facilitator impartiality and content neutrality. A facilitator with a "hidden agenda," an ulterior motive, or a personal predisposition towards a particular outcome can misinterpret (or completely ignores) the backchannel feedback being sent by the participants. This poses a dangerous proposition to the group, as the facilitator may be intentionally leading them astray from the collective, optimal outcome. Practical experience and qualitative research has shown that impartial outsiders are perceived to be more successful than vested stakeholders in facilitating successful results, all other things being equal (de Vreede et al, 2002).

While the actual threshold is always unknown (since it varies as a result of dozens of contextual factors), every facilitator has a practical limit to the number of participants that they can successfully attune to – and as that threshold is exceeded, they will experience diminishing returns on their performance. These problems are exacerbated in distributed environments, where the amount of backchannel feedback is reduced, and amplified significantly in asynchronous modalities where the amount of backchannel feedback is minimal, if present at all (Briggs et al, 1998).

2.3 The next generation of collaboration methods and design

As the science of collaboration continues to progress, more people will find themselves being engaged with more distributed, asynchronous technologies. The next generation of collaborative tools and methods should be designed to require less human facilitation and empower a collaborative group's participants to "drive" themselves to their own optimal solution without the need for a human facilitator.

There are three academic approaches that have surfaced in recent years that are worth noting, since they directly address this practical concern: ThinkLets, participant-driven GSS (PD-GSS), and dynamic collaboration.

2.3.1 ThinkLets

Briggs et al (2003) proposed a framework called collaboration engineering, which aims to reduce a collaborative group's need for a facilitator by empowering the participants with pre-packaged tools that leverage the expertise of a skilled facilitator and enable them to structure and execute refined collaborative processes. These tools would enable any stakeholder to create their own collaborative session to complete a series of basic tasks.

The primary component of that framework is the thinkLet, which is defined as “the smallest unit of intellectual capital required to create one repeatable, predictable, pattern of collaboration among people working toward a goal" (Briggs, de Vreede, Nunamaker & Tobey, 2001). Each thinkLet contains:

- The tool itself (all required hardware and/or software used to complete the unit of work);
- All configuration details (guidance on how to set up the thinkLet to meet a group’s needs); and
- A script (the procedures and instructions that should be provided to all group members).

The inherent wisdom of such an approach is that its modular design helps novice facilitators design collaborative sessions that utilize the best practices of facilitation (since they are part of the thinkLet’s package). Customized collaborative sessions can be constructed quickly using a combination of thinkLets to accomplish specific goals. In addition, thinkLet architects could even pre-configure a string of thinkLets to accommodate often-used or repetitive collaborative workflows (e.g., divergence, convergence, consensus-building, etc.).

The flexibility of these components and the ease with which they can be transformed into templates and integrated into existing GSS installations makes these thinkLets a very useful building block for future collaborative work – without the need for a skilled facilitator to manage the process, although even experienced facilitators can use thinkLets, too.

2.3.2 Participant-driven GSS (PD-GSS)

Participant-driven GSS (or PD-GSS) is another proposed modular collaboration engineering framework that promises to empower participants and alleviate the need for skilled facilitation. But where thinkLets are often intended for use in synchronous, proximal collaborative settings, PD-GSS is entirely focused on asynchronous and/or distributed group work (Helquist, 2007).

While PD-GSS remains little more than a nascent notion today (the foundational technologies and processes are still in development), the concept has slowly built traction in research and practice. The proposed PD-GSS applications are designed to be intelligent systems, intended to leverage the skills and abilities of a group to gain the perspective and insight of a variety of people (Helquist, 2007). They will be capable of analyzing the behavior and productivity of a group's participants so that the system can automatically route them to the activities or tasks that would benefit the group most. This automated routing is made possible because the underlying structure of PD-GSS sessions is more complex than traditional GSS systems, which are designed to be much more serial in nature.

For example, the system might recognize that a certain participant is posting so much content into a particular idea generation activity that they are effectively dominating that "conversation." In response, the next time they log on to the system, it may prevent them from re-entering that activity, forcing them to perform an initial sort of feedback from another activity or review other participants' contributions to another activity (Helquist, 2007).

The hallmarks of the PD-GSS framework are the 24/7 availability of its resources, its capability to effectively manage very large groups (well beyond what a human facilitator could handle), and the variety of structural controls that tap into the system's analytical intelligence to keep the participants actively engaged and motivated to contribute (Helquist, Kruse & Adkins, 2006a).

PD-GSS applications will likely be significantly slower than their GSS counterparts, in terms of task completion times, and will likely feature a lower volume of responses to idea generation tasks because of the proposed peer-reviewed nature of feedback approval and revision process (Helquist, Kruse & Adkins, 2006b). However, the additional time required to complete these tasks may prove to be invaluable.

One expectation of these proposed systems is a significant reduction of low-quality feedback (often referred to as "noisy input") in idea generation tasks (Helquist, 2007). This may help to alleviate the initial shock associated with convergence and idea organization tasks, because the input to be sorted through will feature fewer items, but those items will likely be of higher quality. Additionally, every task in the collaborative session will be designed to be performed in parallel – sorting, alternative analysis, voting and even policy development will be able to be performed in parallel (Helquist, Kruse & Adkins, 2008).

Although PD-GSS sessions are designed to be participant-driven, there is still a need for a human facilitator to configure the intelligent system effectively. From instructional dissemination to system-level threshold configuration, a facilitator of some kind will still be required to effectively use PD-GSS tools. In other words, the role of the PD-GSS facilitator is more akin to a technological facilitator than the process facilitator usually seen leading traditional GSS sessions.

The ultimate objective of PD-GSS is to help a group's members identify (on their own) possible ways to achieve consensus within the group and proactively help focus the group's collective attention on the most appropriate ideas and alternatives (Helquist et al, 2008). However, there is also the perilous possibility that certain participants (or even entire groups) will not be able to recognize those opportunities without facilitated guidance. Since the proposed systems will likely not have a means for a human facilitator to provide such guidance, there is always a chance that a critical mass of participants may lose faith in the system quickly as a result, causing the effort to go to waste.

2.3.3 Dynamic collaboration

Dynamic collaboration is a new approach to coordinate the complex collaborative tasks of a virtual team through constant process alignment and product refinement (Helquist, Deokar, Meservy & Kruse, 2011). It utilizes highly flexible workflows to create an agile environment for group interaction. Rather than following a pre-defined set of procedures and processes, dynamic collaboration systems rely upon the judgment and feedback of the participants to generate the group's workflow "on the fly."

As participants interact with a dynamic collaboration system, they are routinely asked for their opinion on issues such as: Overall work product quality, data sufficiency, and even proposed next steps. In other words, the collective wisdom of the group determines what tasks are performed when and for how long – rather than relying on a human facilitator or an intelligent agent to make those judgments.

Thus, dynamic collaboration uses constant polling to enable a virtual group to quickly alter its processes/tasks and, in turn, refine its work product in response to a complex (or even dynamically-changing) task.

This agile approach is thought to be most valuable in ad-hoc environments (where virtual teams are formed quickly and the participants do not have well-established relationships) and in any situation where there is so much ambiguity in the problem that a pre-defined collaborative workflow is not easily determined (e.g., "we can't utilize a thinkLet if we really don't understand the nature of the crisis yet").

Helquist et al (2011) describe the benefits of dynamic, composable collaboration as a means to:

- Improve agility through rapid composability;
- Provide specific task support through focused modules;
- Create flexible workflows that will help to move the distributed team through efficient and effective processes; and
- Enhance usability through simple, intuitive user interfaces.

Dynamic collaboration is intended to be fully automated (responding to the polling results), although a human facilitator could be incorporated into any session, if one is available. It is also intended to accommodate a virtual team's asynchronous activity, possibly through a PD-GSS system (provided that the intelligence of the PD-GSS agents is directed by the results of the participants' polling).

The concept of dynamic collaboration is highly promising, but the ultimate success of every session is dependent upon the availability of a highly- scalable system architecture that features an intuitive interface providing clear situational awareness for the participants. However, the work on dynamic collaborative systems is still entirely conceptual at this point – since such specialized distributed parallel collaborative systems do not yet exist.

2.4 Cognitive load

As mentioned earlier, the participants of a collaborative session usually view "idea organization" as the single most difficult phase in the traditional, serially-structured collaborative meeting. The root cause of this sentiment is thought to be entrenched in task complexity, or the mental demand placed upon the participants. To perform this task successfully, each participant must focus their attention for a considerable amount of time in order to help the group process the feedback and develop a collective schema that best organizes the content. It requires not just simple reading of responses, but forces each participant to make comprehensive judgments/decisions on each item in order to store it effectively. As the group size increases, and more feedback is generated, the task becomes even more mentally challenging for the average participant. As the task goes longer, a participant's enthusiasm with the collaborative process (which peaked after the idea generation phase) may drop as well.

The complexity of idea organization and the categorizing of feedback in a GSS is a difficult mental task because it may lead a participant to feel a sense of "information overload" – which occurs when an individual perceives the volume of information to be greater than their mental faculties can handle.

Cognitive load theory (Sweller, 1988) posits that humans can store and process only a limited amount of information at any given time, as defined by the individual's cognitive abilities and the capacity of their (short-term) "working memory." An individual's total cognitive load is actually the sum of three components (Sweller, Van Merriënboer & Paas, 1998):

- *Intrinsic cognitive load* (the inherent demand of the task itself);
- *Germane cognitive load* (demand associated with processing and schema development); and
- *Extraneous cognitive load* (demand associated with the presentation of information).

Some people perform certain mental tasks faster or more efficiently than others (so the capacities vary from person to person), but a given individual's cognitive load capacity is fixed, regardless of the task (Voorhies & Scandura, 1977). Whenever the inflows of sensory information exceed an individual's fixed capacity to manage or process that information, their brain signals their body with general (physical) signs of anxiety, fatigue, and/or discomfort. In practice, most skilled facilitators are able to detect these subtle symptoms and make adjustments to the session (or refine their methods on-the-spot) in an attempt to minimize the negative effects of this phenomenon.

Sweller et al, (1998) assert that cognitive load refers to a human's executive control of their working memory during a learning activity. In order for a person to "learn" something, ALL of the information currently stored in their working memory must be successfully processed. If any part of the information was processed unsuccessfully, that person must repeat the task (until they successfully process it all) in order to have learned something.

In practical terms, the idea organization phase of a collaborative meeting is indeed a learning activity – the group is learning (collectively) from the data they generated during their brainstorming activity. They are working together to convert raw data into information, in order to generate knowledge. To accomplish this task as a group requires the participants to read an item in the list, gain an understanding of that item, and then they “learn” it by developing a logical “schema” for storing that item in the GSS tool (i.e., find or create an appropriately-named folder to contain it). This process is repeated hundreds of times during a brief time period, and since the ideas may address literally dozens of topics, the task is indeed significantly difficult (even when distributed amongst the membership of the entire group).

Viewing the categorization task in this light (as a learning activity) enables the application of cognitive load theory to collaborative sorting situations. Since two of the components of total cognitive load (intrinsic and extraneous cognitive load) are held fixed in GSS-enabled collaboration sessions, subsequent scientific investigation in this arena will focus on the third component, germane cognitive load (schema development). Thus, all subsequent references to cognitive load in this dissertation will refer specifically to the germane component, unless otherwise noted.

Experimentation of the cognitive load theory has shown that reduced levels of cognitive load increase an individual’s learning efficiency (Sweller et al, 1998). If collaborative sessions are designed to minimize the effects of cognitive load, it stands to reason that the group (as a whole) might also be more efficient in the completion of their tasks, which would likely lead to more effective decision-making.

However, when the categorization of items takes place in a shared GSS tool (where the participants all see the same items at the same time), the design and functionality of the categorization feature may also cause additional frustration (perhaps increasing the intrinsic cognitive load). For example, in the categorization feature of GroupSystems ThinkTank software, the meeting participants share the same screen and all additions/edits/moves to the items on that screen are updated in real time, as all of the participants work to process the data. This may cause some participants to feel heightened levels of frustration when an item they are reading (or pondering) suddenly disappears from their view – after another participant has moved it. When this happens, the participant often feels as though they wasted their time on that item, triggering a small, unsatisfactory psychological response. If this happens often enough, the group's dissatisfaction with the process may build and (if taken to an extreme) undermine the perceived satisfaction with the tool or the session.

2.4.1 The measurement of cognitive load

Obviously, any scientific exploration of the effects of cognitive load would require some instrumentation or tool that could provide objective measurements of the cognitive load experienced by a participant in a particular experimental treatment condition. The measurement method that was selected for use in this experiment was the NASA Task Load Index (NASA-TLX), created by Hart & Staveland (1988). The tool itself and the rationale for its selection will be discussed at length in the next section.

Several other measurement alternatives were also considered for use in this exploratory study. Since this experiment represents the first attempt to isolate a particular collaborative activity and measure the effects of cognitive load resulting from its completion (across three treatment conditions), it seemed logical to incorporate manipulation checks into the measurement instruments. These checks would help to verify the validity of the measurements recorded and serve as a backup measure in case of implementation or measurement errors. Ultimately, the most suitable manipulation checks were discovered by selecting items from the other measurement instruments that were not selected.

Paas and Van Merriënboer (1993) developed relative condition efficiency, which is one of the most noteworthy measurement tools to date. However, this measure, which combines mental effort ratings with performance scores, is intended to quantify perceived mental effort via simple comparisons of instructional conditions (i.e., learning modalities for classroom curriculum). This measure was briefly considered for use in this dissertation experiment, but ultimately it was rejected – simply because the emphasis of this particular experiment focuses on the effect of the volume of content and not the manner in which the content is presented to the participants.

DeLeeuw & Mayer (2008) conducted an experiment to "examine the sensitivity of three commonly used techniques for measuring cognitive load – response time to a secondary task during learning, effort ratings during learning, and difficulty ratings after learning - to each of the three aspects of cognitive load." They simply recorded self-reported participant responses (on a 9-point scale) to direct questions regarding effort, difficulty, etc. Their findings indicated that the three components of cognitive load may be measured independently. However, their methodology was not duplicated as a measure of cognitive load for this experiment either.

The rationale for this decision was due to an operational concern and the challenge of implementation. Operationally, the participant's individual sort completion time was a particular variable of interest and was predicted to vary greatly across treatment conditions. As such, interfering with the proposed experimental task (even for a critical measurement of the cognitive load variable) was deemed to be intrusive and contrary to the primary objectives of the study. In addition, the particular version of the ThinkTank GSS tool that was utilized for this experiment did not allow for easy "polling" of the participants during the experiment. Thus, DeLeeuw & Mayer method was also rejected for use in this particular study, although it remains a very viable alternative that should (and will) be considered for future experimental study in this arena.

Paas, Tuovinen, Tabbers & Van Gerven (2003) provide an intriguing analytical comparison of other cognitive load measurement alternatives. Several of these methods were also considered for implementation in this particular experiment, and may still prove to be highly effective in this context. But it was ultimately decided that the validation of the NASA-TLX data could be best accomplished with a parsimonious solution – the simple delivery of pre- and post-experiment surveys that collected participants' self-report measures of comfort/fatigue levels, task difficulty, mental effort, sort effectiveness, and general satisfaction (see Appendix E and Appendix F.)

2.5 Normalized clustering error (NCE)

In this experiment, task effectiveness is a critical element in the evaluation of a collaborative group’s work product. In order to facilitate the objective comparison of group sorting effectiveness in this collaborative experimental setting, another quantitative measure is required.

Although the experimental design features three treatment conditions, the participants assigned to the various conditions ultimately sort items from the same original set. Thus, the conditions can be compared to one another, provided an objective assessment of a group's sorting performance can be made for each instance. A quantitative clustering accuracy algorithm was determined to be the ideal metric to determine a sort's relative quality.

The measure chosen for this purpose is known as the normalized clustering error, or NCE (D. Roussinov & Zhao, 2003; Roussinov & Chen, 1999). This metric has been used in prior collaborative research to evaluate the quality and accuracy of automated clustering algorithms and approaches.

The calculation of an NCE value consists of comparing a treatment group's submitted sort to a pre-determined "gold standard" sort -- which, for this experiment, was created by a highly-skilled, certified facilitator who is a subject matter expert on the original content. (Note: The original list of feedback items that was sorted was originally generated in 2007, and the "gold standard" sort was prepared at the same time.)

The NCE metric evaluates the associations between the items contained in the various clusters (or folders, in ThinkTank terminology) of a particular sort. A quantitative measure of accuracy is calculated by comparing the associations of a submitted sort to the gold standard set, as shown in Figure 3.

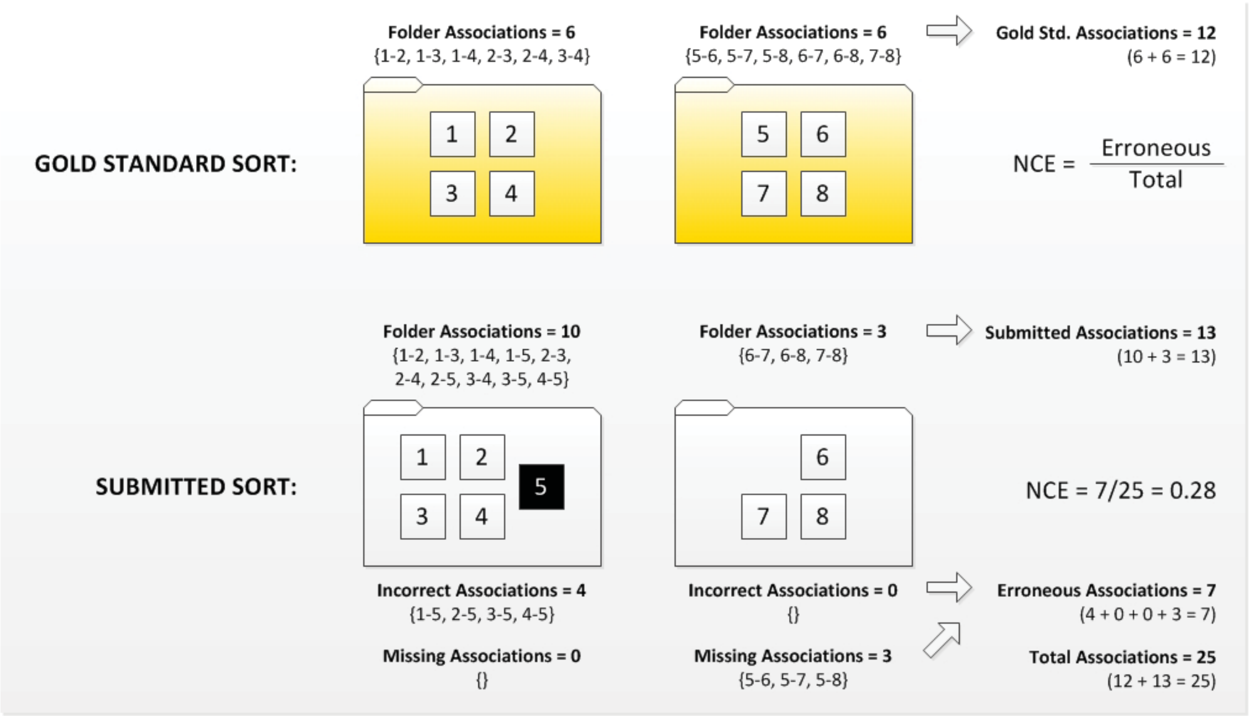


Figure 3: How normalized clustering error is calculated

To begin, the formula for calculating an NCE value is:

$$NCE = \frac{Erroneous\ associations}{Total\ associations}$$

An association can be defined as a "link" between two items in the same folder. The formula for calculating the number of a particular folder's associations (where *n* represents the number of items contained in that folder) is:

$$Folder\ associations = \frac{n(n-1)}{2}$$

In Figure 3, notice that there are four items in each folder of the gold standard sort, thus there are six associations within each folder. (4 * 3 ÷ 2 = 6)

Therefore, the denominator of the NCE figure is simply the sum of the number of folder associations for ALL of the folders being compared (in both the gold standard folders and the submitted sort). Thus, in the example shown in Figure 3, the denominator would be: 6 + 6 + 10 + 3 = 25.

The numerator of the NCE figure is the summation of the number of “incorrect” associations (those found in a submitted folder that do not appear in the gold standard folder) as well as the number of “missing” associations (those found in the gold standard folder that do not appear in the submitted folder).

$$Erroneous\ associations = incorrect\ associations + missing\ associations$$

In the example shown in Figure 3, the submitted sort folder on the left contains item number 5, but that item is not in the gold standard’s folder. Thus, there are 4 incorrect associations and 0 missing associations in that submitted sort folder. Since item 5 is not contained in the submitted sort folder on the right, there are 0 incorrect associations there, but there are 3 missing associations in that folder. These incorrect and missing figures are simply added together to calculate the total number of erroneous associations: 4 + 0 + 0 + 3 = 7.

Hence, the NCE value for the example shown in Figure 3 would be 7 ÷ 25 = 0.28.

The NCE metric has a range between 0 and 1. An NCE value of 0 represents a submitted sort that perfectly matched the gold standard– there are no incorrect and no missing associations. Whereas an NCE value of 1 represents a submitted sort that is completely different from the gold standard – the comparison of the two featured no associations in common.

This objective NCE value will be used to measure the effectiveness of the groups’ sorted results in this dissertation experiment, where lower values (those closer to 0) will indicate higher-quality sorting performances.

3 Methods

The experiment conducted for this dissertation is an attempt to explore the scientific validity of the cognitive load-based explanations for the decline in satisfaction observed in a sorting task (with respect to the organizational effectiveness of a group's effort), using a distributed (or parallel) task design.

The core activity of the experiment involved using a commercial collaborative software application (GroupSystems ThinkTank) that was configured to provide individually-customized "sessions" for each participant to perform their sorting task. The sessions were all pre-configured to feature some or all of the idea generation feedback of a previous group (see Appendix H), which addressed an issue that would be reasonably familiar to any of the potential subjects who were eligible to participate in this study.

Each individual was required to watch a brief training video that provided instruction on how to use the ThinkTank software to complete the sorting task, and also given a printed set of instructions that reinforced the training (to help insure that all subjects could complete the task independently, without significant difficulty).

In addition, two online surveys were created to collect participants' self-report measures (both before and after the sorting task) on a variety of topics, such as: Cognitive load, computing skill/experience, psychological state (e.g., comfort and fatigue), collaborative attitudes, estimates of effectiveness, and other evaluative dimensions (see Appendix E and Appendix F).

3.1 Research questions and hypotheses

The primary objectives of this research were to answer the following research questions:

RQ_1: Is it cognitively "easier" for individual group members to sort/classify a smaller subset of a group's entire pool of ideas, rather than the entire set?

RQ_2: In a collaborative setting, would a group be able to sort a set of ideas faster, if it were broken up into smaller subsets and sorted individually by group members working in parallel, rather than working serially?

RQ_3: If a group's members sorted smaller, distinct subset of ideas in parallel, would this adversely affect the effectiveness of the overall result?

Therefore, the central hypotheses of this study were simply:

H_1: An individual's perception of the cognitive load associated with sorting a set of ideas is positively related to the number of items in the set to be sorted.

H_2: The speed of sorting smaller subsets of a larger pool of ideas is significantly faster than sorting the entire set.

H_3: Sorting smaller subsets of a larger pool of ideas is as least as effective as sorting the entire set.

3.2 Independent variables and experimental treatments

The collaborative design utilized in this study was selected primarily because it was highly generalizable and could easily be implemented in field settings. The basic approach merely entails selecting (at random) a predefined number of items from the aforementioned full list of feedback and assigning them to individuals to process/sort. Since this can be done without a facilitator and involves individual effort, the experiment is representative of an asynchronous/distributed collaborative group with its members working in parallel.

Thus, the independent variable is the size of the set of ideas to be processed by each person in their ThinkTank session, and three treatment conditions were implemented:

- Condition A required a participant to sort all 110 items;
- Condition B required a participant to sort a randomly-selected set of 55 items;
- Condition C required a participant to sort either 36 or 37 randomly-selected items.

Figure 4 (below) graphically depicts this design and highlights some of the preliminary predictions of the dependent variable relationships, which will be discussed in the next section.

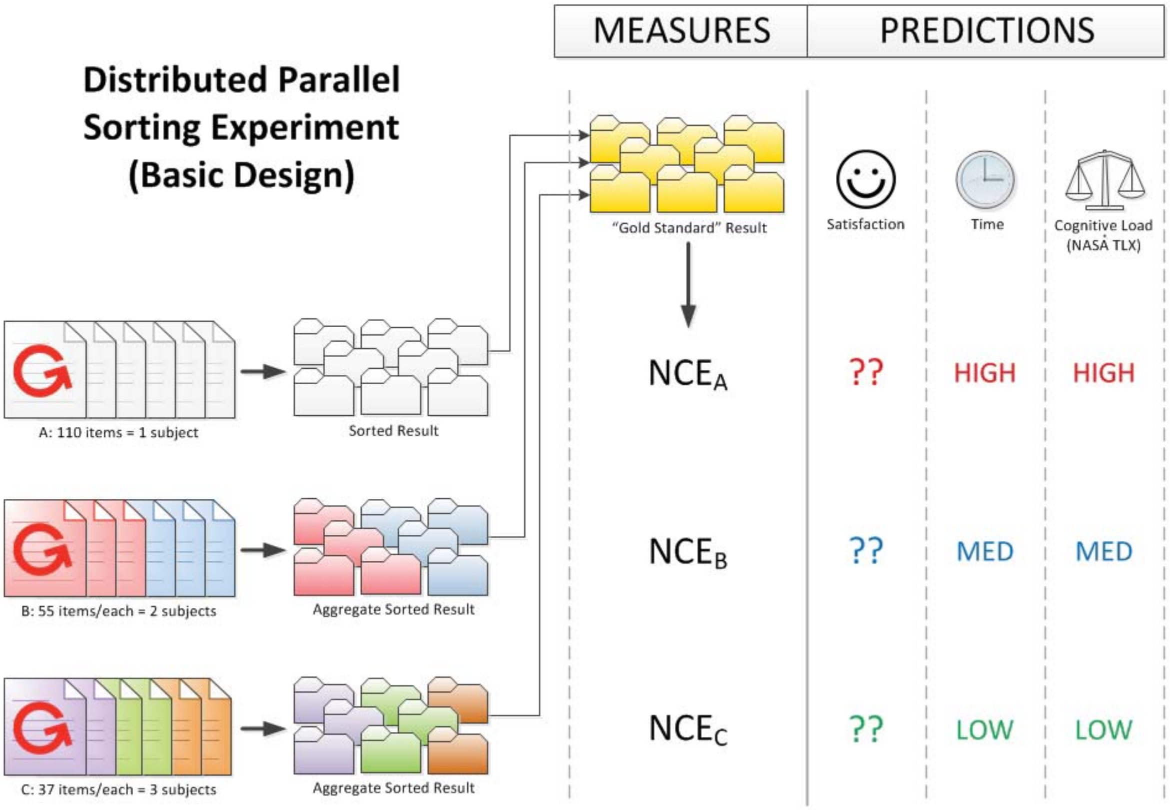


Figure 4: An illustration of the basic experimental design of the study

3.3 Dependent variables and data collection

Since this particular research project is somewhat exploratory in nature (and since cognitive load is notoriously difficult to measure concretely without implementing invasive technologies), a variety of information was gathered about the participants and their perceptions of the process. Initially, there were three dependent variables of primary importance to this study:

- The quality/effectiveness of a participant's sort;
- The time required to complete the sort; and
- The cognitive load experienced (or the perceived difficulty of the task).

The survey instruments utilized in this study provided other supporting data, including user satisfaction reports, manipulation checks, and cross-validating measures. While this data may prove useful in follow-up experiments, much of it will not be addressed in this dissertation.

However, by aggregating/consolidating the sort results of some participants, this experimental design can be viewed as simulating an asynchronous group's parallel task in a distributed environment – a task that was once considered to be best performed serially, via the effort of a synchronous group in a proximal setting. It was hoped that the participants in this study could achieve the "wisdom of the crowds" by allowing them to work autonomously in this setting.

3.3.1 Measuring sort effectiveness via normalized clustering error (NCE)

Of principal concern to the study is the effectiveness of the sorts created by the participants, thus the first dependent variable to be calculated for each response set was a normalized clustering error (NCE) score, as it provides an ordinal measure of sort quality (Roussinov & Chen, 1999).

The NCE score generation process, however, requires measuring each participant sort against a "gold standard" result. The sort used as the "gold standard" in this research was performed in 2007 by a human "oracle" (an expert facilitator who happens to be a subject matter expert) just after the original list was originally generated.

Additionally, it is important to note that generating an NCE score requires that the two sets be comprised of the same items – in other words, both sorts in the comparison need to contain the same 110 items. This computational requirement is the reason why the independent variable values were chosen.

The participants in Condition A serve as a control group and sort all 110 items in the list, which makes generating an NCE score simple as their sort result can be directly measured against the gold standard.

However, for those in Condition B, two participants will need their 55-item sorts to be consolidated in order to create a full sort of 110 items and the items they sort will have to be distinctly different. So, the sessions created for Condition B were created in pairs, featuring one randomly-selected set of 55 items and another set with the remaining 55 items that were not selected to appear in the first.

Similarly, for those in Condition C, three participants' results would be needed to make a full sort of 110 items. Thus, one set of 37 items was randomly selected from the original list of 110 items… from the remaining 73 items, another set of 37 was randomly selected to form another participant's set… and the remaining 36 were used for a third participant.

In other words, the work product of some participants (those in Conditions B or C), unbeknownst to them, were consolidated with the work product of other participants to form a representation of a distributed team's result. These aggregated results were then compared to the gold standard result to generate an NCE score for that ad-hoc subgroup.

3.3.2 Measuring time in a parallel task

Since the design of the experiment is intended to replicate a distributed, parallel scenario, calculating the completion time of an ad-hoc subgroup is simply the longest time of any member of that subgroup.

In practical terms, if you were to assume that all of a subgroup's participants were given the task to sort their particular items at the same time and instructed to work in parallel, the group (as a whole) would not be done with the task until the last person finished. Therefore, for the purposes of this analysis, the time variable is the maximum completion time of the ad-hoc group's individual times.

3.3.3 Measuring cognitive load via the NASA-TLX method

Many experimental methods were considered in an attempt to quantify the cognitive load experienced by the participants of this experiment, but the NASA Task Load Index (or NASA-TLX) method (Hart & Staveland, 1988) was ultimately selected and implemented.

The reasons for the selection of NASA-TLX over other alternatives include:

- It is a non-invasive method, requiring no specialized instruments or equipment;
- It is well-regarded in academia and used often in Human Factors research;

- The dimensions of the method seemed to be a good match to this study's variables;
- It is simple to implement and administer; and
- It could be completed quickly and easily by the experiment's participants.

NASA-TLX is used by NASA researchers (and other scientists) to objectively evaluate the overall difficulty of particular tasks, and its results can be used to discern the cognitive load demands that a particular task imposes upon an individual. Essentially, it is self-reported feedback on the various demands experienced as a result of performing a task. The basic approach features survey questions using 100-point scales (in increments of 5) to classify the following aspects of an activity of research interest:

- Mental demand (which includes cognitive load, computational difficulty, etc.);
- Physical demand (which can also apply to computer-based activities, e.g., eye strain);
- Temporal demand (time-based pressure or pace issues);
- Performance (self-reported evaluations of success and/or failure);
- Effort (self-reported measures of the amount of work required); and
- Frustration (feelings of irritation, stress, annoyance, etc.).

3.3.3.1 Administration of the NASA-TLX method

There are several ways to implement the NASA-TLX method – paper-and-pencil versions, computer-based versions, and even variably-weighted versions have been used by researchers in the 25 years of its existence. The most widely-used versions seem to be the pencil-and-paper versions that feature a "pairwise comparison" method of weighting – where participants are asked a series of 15 questions which determine the ideal "ranking" of importance given to the six dimensions of the survey.

To date, there is no clear consensus on the ideal administration method of the TLX approach, as most academic analyses of the tool tends to focus on specific applications or environments (which only loosely apply to the task in this study and the primary phenomena of interest).

Thus, a computerized version of what is called the “Raw TLX” method was put into use for this research study. “Raw TLX” scores are not weighted and simply averaged together. The approach was selected due to:

- Its ease of implementation (the survey tool enabled proper question formatting);
- The reduced chance of data errors (in transcribing pencil-and-paper responses); and
- The relative speed of the task, in comparison (adding 15 questions to the post-experiment survey didn’t seem to be an ideal choice at the time).

Furthermore, the “Raw TLX” seemed to be more appropriate when the participants’ response data would need to be combined (in Condition B and Condition C) to enable a consistent unit of analysis (i.e., the ad-hoc subgroups).

Thus, the weighting element of the NASA-TLX tool was not implemented in this particular experiment. The effects of this decision will be discussed later in this dissertation.

3.4 Experiment details

3.4.1 Research setting

The research was conducted on the main campus of the University of Arizona in McClelland Hall Room 214. This facility (formally known as the “Arizona Public Service Technical Classroom,” but hereafter referred to as MCLD 214) is a small auditorium-style facility that features 29 identically-configured PC workstations for use by the participants. It is a typical electronic classroom environment that is ideally suited for the collaboration tasks proscribed in this research, as it has been the location of hundreds of collaboration research projects since the building’s opening in 1992.

3.4.2 Experimental task overview

Each individual in the research study was processed as follows:

- Each person was asked a series of questions to verify their eligibility for the survey (see Appendix B) and given a customized set of anonymized user credentials for the collaborative software (see Appendix C for a sample);
- They were seated at a workstation in the MCLD 214 facility which displayed a consistent desktop image to assist them in the completion of their tasks (see Appendix D);
- They were then presented with consent information which they were required to acknowledge and completed a pre-experiment survey (see Appendix E);
- Then, they were shown a brief training video demonstrating how to use the collaborative software (see Appendix G for the transcript);
- They then logged in to the collaborative software (using the credentials provided earlier), and sorted a list of items (see Appendix H);
- After logging out of ThinkTank, they completed a post-experiment survey (see Appendix F);
- Finally, they reported back to the Investigator to return the task checklist they were given upon arrival (as a safety precaution to avoid re-accessing the application and tainting the data at a later time) and given a printed confirmation receipt signed by the Investigator (as tangible proof to give to their instructor for the extra credit) and promptly left the facility.

3.4.3 Participants

All of the participants were adult-aged college students (without regard to race, ethnicity, or gender) who confirmed that they were 18 years of age or older before beginning. The study did not target any vulnerable populations.

Table 1below provides the descriptive statistics of the experiment’s participant population, by treatment condition. The treatment conditions are clearly balanced, in terms of gender and age.

Table 1: Descriptive statistics of the experiment's participants (by treatment condition)

	N	GENDER		AGE					
		MALE	FEMALE	MEAN	MEDIAN	MODE	MAX	MIN	STD. DEV.
A	56	29	27	21.0	21	21	33	18	3.32
B	122	61	61	20.9	21	21	38	18	3.44
C	174	88	86	21.4	21	21	58	18	5.07
TOTAL	352	178	174	21.2	21	21	58	18	4.30

It should also be noted that the number of participants increases with each condition (from A to B to C).This is because only ONE person is required to create a “full sort” for Condition A, but TWO people are required to create a “full sort” for Condition B and THREE people are required for Condition C. Thus, Condition B should have twice as many participants as Condition A… and Condition C should have three times as many participants as Condition A… which is reflective of the participant population shown.

3.4.3.1 Eligibility requirements

Since the study required that each subject be able to read text on a computer screen, blind or severely visually-impaired subjects were not eligible to participate (unless they provided their own enabling technology/equipment).

In addition, subjects who had inadequate English skills were also not eligible to participate, since the study requires subjects to have college-level English reading, writing and comprehension skills.

Finally, cognitively-impaired subjects were not eligible to participate, since the sorting activity is generally regarded as a cognitively-challenging task.

3.4.3.2 Recruitment efforts

The recruitment of participants was accomplished exclusively via in-class announcements made by various professors in the Eller College of Management who routinely offer extra course credit to enrolled students who participate in doctoral research studies. The professors read the study’s recruitment solicitation at the beginning of each class, and showed a PowerPoint slide featuring a QR code link to a website that summarized the opportunity. The professors also posted transcripts of the announcement on their class websites, and some sent e-mail to their students.

3.4.3.3 Obtaining consent

There was minimal risk involved to participate in this study. There are no known physical, psychological, social, legal, or economic risks that applied. All subjects were instructed that they could walk away from the experiment at any time with no adverse consequences. As a result, this research did not consent any subjects as “the research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside of the research context”. In lieu of the consent form, participants were notified of their rights by an online page before participating in the initial survey (see Appendix E).

3.4.4 Experimental procedures and administration

The study was conducted in the MCLD 214 facility over a period of approximately two weeks in the fall semester of 2013. The facility was open for 12-14 hours each day, ready to handle any subjects that elected to participate. No appointments were necessary, and qualified “walk-ins” were welcomed.

The Investigator managed all procedural tasks personally throughout the duration of the study. To insure that all subjects were processed in a consistent manner, a daily experiment operations guide was created (see Appendix A) and a standardized script for subject qualification was employed to properly verify each potential participant’s eligibility (see Appendix B).

Subjects were recruited from the Eller College of Management’s student population (graduate and undergraduate) who were willing to participate in experiments.

Specialized equipment was not required for this research. Only the standard PC workstations in MCLD 214 were used by the participants.

3.4.4.1 Data assurance safeguards to preserve privacy and confidentiality

All of the data generated as a result this study was stored on physically and logically secure servers, located within the MCLD 214 facility, and only accessible by the Investigator.

The sorted data sets in the ThinkTank software and the participants' responses to the two surveys were the only data that was collected and stored. No other records or information were accessed and no follow-up contact occurred.

In terms of confidentiality, it is important to note that the only identifiable information provided by the participants (their university e-mail address, entered during the pre-experiment survey) was permanently deleted immediately after that information was given to the appropriate course instructor (allowing the student to receive course extra credit for their effort). Only anonymized data was stored and used for analysis.

Protection of participant privacy was accomplished in two ways: 1) No observation or recording of subjects was conducted in or around the facility during the experiment; and 2) The MCLD 214 facility features computer monitors that are protected with "privacy screens,", which guard against people attempting to read another user's monitor from nearby positions. As an additional measure to protect privacy and confidentiality, the only people allowed in the facility during the experiment were the Investigator and the participants themselves.

4 Analysis and results

The survey data (pre- and post-experiment) that was collected over the course of the experiment was consolidated into a single spreadsheet for statistical analysis.

4.1 Descriptive statistics of sort quality/effectiveness (NCE)

The sorting results of each participant were filtered and processed separately – i.e., the text portion of each feedback item was removed, leaving only its numerical identifier, which (for those in Conditions B and C) was consolidated with the other ad-hoc subgroup members' results. This enabled the generation of the NCE score via an automated Python script. The program individually compared each ad-hoc subgroup's sort to the gold standard and created a numerical NCE value, which was then added to the consolidated data spreadsheet. The number of category folders created by each participant's sort was also recorded, and a statistical summary of all of the subgroups' sort data appears below in Table 2.

Table 2: Descriptive statistics of subgroup sort quality/effectiveness (NCE)

	N subgroups	SORT GROUP'S NCE SCORE				NO. OF FOLDERS CREATED PER PERSON			
		MEAN	MAX	MIN	STD. DEV.	MEAN	MAX	MIN	STD. DEV.
A	55	0.7434	0.8674	0.5376	0.082	7.55	21	2	3.50
B	60	0.7963	0.8729	0.7007	0.035	6.58	14	2	2.48
C	58	0.8260	0.9262	0.7500	0.034	6.16	13	2	2.29
TOTAL	173	0.8028	0.9262	0.5376	0.053	6.52	21	2	2.62

It should be noted that two sets of data were invalidated and removed from all analysis. Due to undetected server errors at the time of the ThinkTank session's configuration, two participants (one from Condition A and one from Condition B) were not given the correct set of ideas to be sorted (the server did not process the last few items correctly in each case). Therefore, the affected Condition B subgroup (the erroneous set and its corresponding partner set) was eliminated, as was the single Condition A sort.

Since lower NCE scores indicate a better quality sort (a 0 score is a perfect match), the mean NCE scores in Table 2 above imply that the sorting performance of those in Condition A was better than that of the subgroups in Condition B, which was better still than the results of the subgroups in Condition C.

4.2 Descriptive statistics of activity completion times

The completion times of the survey tasks were automatically recorded by the server – and these times were added to the spreadsheet. However, the ThinkTank application server does not make its time records available to an administrator. Thus, the participants were required to manually enter a "START" and "FINISH" submission in separate ThinkTank activities which had time-stamps enabled… and these values were copied to the spreadsheet to calculate the required completion time for the sorting activity.

Table 3: Completion times (in seconds) for key experimental tasks

	N	PRE-SURVEY COMPLETION TIME (in secs)			
		MEAN	MAX	MIN	STD. DEV.
A	56	140.7	507	64	70.4
B	122	163.2	594	56	94.5
C	174	143.1	414	56	71.1
TOTAL	352	149.6	594	56	80.3

	N	SORTING COMPLETION TIME (in secs)			
		MEAN	MAX	MIN	STD. DEV.
A	55	1458.5	2758	878	443.6
B	120	1062.9	2631	452	363.6
C	174	840.7	3171	308	357.0
TOTAL	349	1016.0	3171	308	431.8

NOTE: Three participants' results were removed due to server errors.

	N	POST-SURVEY COMPLETION TIME (in secs)			
		MEAN	MAX	MIN	STD. DEV.
A	55	119.0	216	55	34.5
B	118	134.9	290	66	42.9
C	173	145.8	946	48	93.5
TOTAL	346	137.8	946	48	72.5

NOTE: Six participants' responses were not correctly submitted.

Note that six participants did not correctly submit their post-experiment survey responses. This will explain any differences in the N values that may appear in the results tables in this section. Thus, the six affected subgroups' data was ignored from any analysis that required any missing post-survey data.

However, as Table 3 shows, it is curious to note that while the pre-experiment survey times for Conditions A and C are somewhat similar, the mean post-experiment survey completion times are different between those two conditions – with participants assigned to Condition A spending nearly 30 seconds less to complete it than those participants assigned to Condition C. The large outlier in Condition C (the individual who took nearly 16 minutes to complete their survey) made this statistically insignificant… but with three times as many participants in Condition C as Condition A, this still merits a notice. This curiosity will be addressed later, in the discussion of cognitive load.

4.3 Comparing cognitive load across treatments

The first statistical analyses that were performed attempted to provide an answer to the study's first research question:

> *RQ_1: Is it cognitively "easier" for individual group members to sort/classify a smaller subset of a group's entire pool of ideas, rather than the entire set?*

The initial hypothesis stated:

> *H_1: An individual's perception of the cognitive load associated with sorting a set of ideas is positively related to the number of items in the set to be sorted.*

Table 4 shows the descriptive statistics of the subgroups' averaged Raw TLX scores, by condition.

Table 4: Descriptive statistics of the subgroups averaged Raw TLX scores (by condition)

	N	Mean	Std. Dev.	Std. Error	95% Confidence Interval for Mean		MIN	MAX
					Lower Bound	Upper Bound		
A	49	41.1	11.95	1.71	37.7	44.5	13.3	63.3
B	49	38.8	9.59	1.37	36.0	41.5	23.8	74.2
C	75	39.0	7.49	0.87	37.3	40.8	22.5	63.3
Total	173	39.5	9.53	0.72	38.1	41.0	13.3	74.2

There were very slight differences in the reported mean Raw TLX scores for the subgroups across the three treatments. An ANOVA (one-way) was attempted to explain the variance in the Raw TLX scores, but because the Levene test for equality of variance was violated in this analysis, F(2,170) = 6.999, *p* = .001, we cannot assume that the variance is equal. However, all of the contrast tests that assume unequal variance failed to yield any statistically significant results. (See Appendix I for full results.)

On the basis of this initial analysis, the initial hypothesis (H_1) would be rejected. This result was fairly surprising, so another analysis was attempted using an "Adjusted TLX" score instead of the "Raw TLX"

4.3.1 Comparing cognitive load (Adjusted TLX) across treatments

Since the "Raw TLX" score was comprised of six separate measures, it seemed logical that perhaps some statistical noise was being introduced into that figure may have caused the failure to find significant results. So, four of the TLX components were dropped to create a simplified "Adjusted TLX" score, comprising only two components to be averaged – mental demand (post-survey question 8.1) and frustration (post-survey question 8.6). This score was re-calculated for all of the subgroups and another ANOVA (one-way) performed.

Table 5 shows the descriptive statistics of the Adjusted TLX scores, by condition.

Table 5: Descriptive statistics of Adjusted TLX measures (by condition)

	N	Mean	Std. Dev.	Std. Error	95% Confidence Interval for Mean		MIN	MAX
					Lower Bound	Upper Bound		
A	49	52.6	20.4	2.92	46.7	58.4	12.5	95.0
B	96	47.0	22.5	2.30	42.4	51.6	5.0	87.5
C	201	47.7	21.4	1.51	44.8	50.7	0.0	95.0
Total	346	48.2	21.6	1.16	45.9	50.5	0.0	95.0

Once again, there were slight differences in the mean Adjusted TLX scores for the subgroups across the three treatments. However, in this case, the Levene test for equality of variance was not violated, F(2,343) = 0.768, p = .465, so equal variances could be assumed. Yet, still none of the results reached the threshold required to be deemed statistically significant. (See Appendix J for full results.)

Thus, based upon the results of the first two analyses, there is no support for the initial hypothesis (H_1) when using the NASA-TLX data. However, a few manipulation checks were employed as backups in the surveys (in the event of discovering insignificant findings with the NASA-TLX scores) and their analyses yielded different results.

4.3.2 Comparing perceived task difficulty across treatments

The first manipulation check variable that attempts to identify potential differences in cognitive load experienced by the participants of this experiment is a self-reported evaluation of the task difficulty. This was asked in the post-experiment survey (question 10.1), and featured a seven-point Likert scale of response options (where 1 represented “not at all difficult” and 7 represented “very difficult”). The descriptive statistics of the participant responses to this question can be found in Table 6 below.

Table 6: Descriptive statistics of self-reported task difficulty (by condition)

	N	Mean	Std. Dev.	Std. Error	95% Confidence Interval for Mean		MIN	MAX
					Lower Bound	Upper Bound		
A	49	3.59	1.74	.249	3.09	4.09	1	7
B	96	2.93	1.60	.163	2.60	3.25	1	7
C	201	3.00	1.66	.117	2.77	3.23	1	7
Total	346	3.06	1.67	.090	2.89	3.24	1	7

The mean responses appeared to be somewhat different, so an ANOVA (one-way) was performed. The Levene test for equality of variance was not violated, F(2,343) = 0.422, *p* = .656, so equal variances could be assumed. While the ANOVA's combined between-groups results were not significant at the = .05 level, the p-value was much closer to significance than the other cognitive load analyses performed earlier, F(2,343) = 2.698, *p* = .053. Thus, a series of contrasts were again employed, but these contrasts yielded significant differences between the perceived task difficulty of Condition A and Condition C – *t*(343) = -2.244, *p* = .025. Additionally, there were significant differences between the perceived task difficulty of Condition A and Condition B – *t*(343) = -2.287, *p* = .023. The difference between the perceived task difficulty between Conditions B and C was not significant. The summary of results for this analysis is shown in Table 7.

Table 7: Results of ANOVA (one-way) contrasts on task difficulty (by condition)

POQ10_1

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups		(Combined)	16.275	2	8.137	2.968	.053
	Linear Term	Unweighted	13.799	1	13.799	5.034	.025
		Weighted	8.159	1	8.159	2.976	.085
		Deviation	8.116	1	8.116	2.960	.086
Within Groups			940.326	343	2.741		
Total			956.601	345			

Contrast Coefficients

Contrast	EXP_COND_NUM			
	A	B	C	
1	-1	0	1	(A vs. C)
2	0	-1	1	(B vs. C)
3	-1	1	0	(A vs. B)

Contrast Tests

Contrast			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)	
POQ10_1	Assume equal variances	1	-.59	.264	-2.244	343	.025	(A vs. C)
		2	.07	.205	.355	343	.723	(B vs. C)
		3	-.66	.291	-2.287	343	.023	(A vs. B)
	Does not assume equal variances	1	-.59	.275	-2.151	70.780	.035	
		2	.07	.201	.363	193.896	.717	
		3	-.66	.298	-2.233	89.662	.028	

So, where the experiment's NASA-TLX measures of cognitive load failed to be statistically significant enough to support the initial hypothesis, the self-reported measure of task difficulty obtained contradict those results – the Condition A participants found their task to be more difficult than the participants in Condition B and Condition C. However, the mean perceived difficulty between Condition B and Condition C are not statistically different from each other.

This finding prompted the analysis of the other manipulation checks, to further evaluate the NASA-TLX tool (as it was implemented) and the findings it yielded.

4.3.3 Comparing self-reported comfort levels across treatments

The second manipulation check that attempts to identify potential differences in cognitive load experienced by the participants of this experiment is a self-reported evaluation of a participant's change in their level of comfort (after performing the task). This was accomplished by asking each participant to gauge their level of agreement to the phrase "I am comfortable and relaxed right now" – in both the pre-experiment survey (question 5.10) and the post-experiment survey (question 12.1). Each question featured a seven-point Likert scale of response options (where 1 represented "strongly disagree" and 7 represented "strongly agree"). The descriptive statistics of the participant responses to this question can be found in Table 8.

Table 8: Descriptive statistics of pre- and post-experiment comfort levels (by condition)

		N	Mean	Std. Dev.	Variance	MIN	MAX
A	PEQ5_10	50	6.06	.899	.792	4	7
	POQ12_1	49	4.88	1.550	2.401	2	7
	Valid N (listwise)	49					
B	PEQ5_10	100	6.20	.734	.539	4	7
	POQ12_1	96	5.35	1.056	1.115	2	7
	Valid N (listwise)	96					
C	PEQ5_10	202	6.25	.897	.804	1	7
	POQ12_1	201	5.62	1.216	1.478	1	7
	Valid N (listwise)	201					

In this case, support for the initial hypothesis would be manifest in negative changes in the mean reports of comfort level across the three treatment conditions – the largest decrease should appear in Condition A, and the smallest decrease should be seen in the results for Condition C.

A glance at the descriptive results reveals that this appears to be the case. Mean comfort levels in Condition A decreased by nearly 1.2 units (6.06 – 4.88 = 1.18) after the sort, while Condition B's participants' mean comfort fell by 0.8 units (6.20 – 5.35 = 0.85) afterward, and Condition C's mean comfort fell only 0.6 units (6.25 – 5.62 = 0.63) as a result of the sorting activity.

To test the significance of these findings, a series of paired-samples t-tests was performed. The results indicated that, for all conditions, the differences were statistically significant (see Table 9).

Table 9: Results of paired-samples t-test on self-reported comfort levels (by condition)

Paired Samples Statistics

EXP_COND			Mean	N	Std. Dev.	Std. Error Mean
A	Pair 1	PEQ5_10	6.06	49	.899	.128
		POQ12_1	4.88	49	1.550	.221
B	Pair 1	PEQ5_10	6.20	96	.734	.075
		POQ12_1	5.35	96	1.056	.108
C	Pair 1	PEQ5_10	6.25	201	.899	.063
		POQ12_1	5.62	201	1.216	.086

Paired Samples Correlations

EXP_COND			N	Correlation	Sig.
A	Pair 1	PEQ5_10 & POQ12_1	49	.304	.033
B	Pair 1	PEQ5_10 & POQ12_1	96	.180	.079
C	Pair 1	PEQ5_10 & POQ12_1	201	.349	.000

Paired Samples Test

EXP_COND			Paired Differences					t	df	Sig. (2-tailed)
			Mean	Std. Dev.	Std. Error Mean	95% CI of the Difference				
						Lower	Upper			
A	Pair 1	PEQ5_10 - POQ12_1	1.184	1.537	.220	.742	1.625	5.392	48	.000
B	Pair 1	PEQ5_10 - POQ12_1	.844	1.173	.120	.606	1.081	7.049	95	.000
C	Pair 1	PEQ5_10 - POQ12_1	.632	1.234	.087	.460	.804	7.257	200	.000

In Condition A, the initial comfort level (*M* = 6.06, *SD* = 0.90) of the participants was significantly higher than their post-experiment comfort level (*M* = 4.88, *SD* = 1.55), *t*(48) = 5.39, *p* < .001.

In Condition B, the initial comfort level (*M* = 6.20, *SD* = 0.73) of the participants was significantly higher than their post-experiment comfort level (*M* = 5.35, *SD* = 1.06), *t*(95) = 7.05, *p* < .001.

In Condition C, the initial comfort level (*M* = 6.25, *SD* = 0.90) of the participants was significantly higher than their post-experiment comfort level (*M* = 5.62, *SD* = 1.22), *t*(200) = 7.257, *p* < .001.

4.3.4 Comparing self-reported fatigue levels across treatments

The final manipulation check that attempts to identify potential differences in cognitive load experienced by the participants of this experiment is a self-reported evaluation of a participant's change in their level of mental fatigue (after performing the task). This was accomplished by asking each participant to gauge their level of agreement to the phrase "I am mentally fatigued right now" – in both the pre-experiment survey (question 5.11) and the post-experiment survey (question 11.1). Each question featured a seven-point Likert scale of response options (where 1 represented "strongly disagree" and 7 represented "strongly agree"). The descriptive statistics of the participant responses to this question can be found in Table 10.

Table 10: Descriptive statistics of pre- and post-experiment fatigue levels (by condition)

		N	Mean	Std. Dev.	Variance	MIN	MAX
A	PEQ5_11	50	2.96	1.591	2.531	1	6
	POQ11_1	49	3.98	1.677	2.812	1	7
	Valid N (listwise)	49					
B	PEQ5_11	100	3.05	1.698	2.882	1	7
	POQ11_1	96	3.13	1.578	2.489	1	6
	Valid N (listwise)	96					
C	PEQ5_11	202	2.84	1.561	2.436	1	7
	POQ11_1	201	2.92	1.555	2.418	1	7
	Valid N (listwise)	201					

In this situation, support for the initial hypothesis would be manifest in positive changes in the mean reports of mental fatigue across the three treatment conditions – with the largest increase occurring in the participants of Condition A.

The descriptive results again reveal this to be the case. The mean post-experiment fatigue levels in Condition A were not only higher than the means of the other treatment conditions (M_A = 3.98, M_B =

3.13, M_C = 2.92), but they increased by 1 full unit (3.98 – 2.96 = 1.02) over the initial fatigue levels after the sort.

However, the changes in mean fatigue level reported by the participants of both Condition B and Condition C were only slightly higher as a result of the task. The participants of both Condition B (3.13 – 3.05 = 0.08) and Condition C (2.92 – 2.84 = 0.08) reported slightly increased mental fatigue levels (0.08 units) after performing the sorting task.

To test the significance of these findings, more paired-samples t-tests was performed (see Table 11). The results indicated that mean reported mental fatigue levels for Condition A were once again statistically significant – the initial mental fatigue level (*M* = 2.96, *SD* = 1.58) of the participants was significantly lower than their post-experiment mental fatigue level (*M* = 3.98, *SD* = 1.68), *t*(48) = -5.43, *p* < .001. As expected, the t-tests of the fatigue values for the other treatment conditions were not significant.

The fact that Condition A's results were significant once again lends support to the initial hypothesis. So, while the primary instrument for measuring cognitive load in this study failed to generate significant results, the manipulation checks all hint at the presence of symptoms associated with increased cognitive load, particularly in Condition A.

Thus, the results of this study yielded were mixed in their support of the initial hypothesis. Several potential explanations for this statistical discrepancy will be discussed in the next sections of this dissertation.

Table 11: Results of paired-samples t-test on self-reported mental fatigue (by condition)

Paired Samples Statistics

EXP_COND_NUM			Mean	N	Std. Dev.	Std. Error Mean
A	Pair 1	PEQ5_11	2.96	49	1.581	.226
		POQ11_1	3.98	49	1.677	.240
B	Pair 1	PEQ5_11	3.05	96	1.688	.172
		POQ11_1	3.13	96	1.578	.161
C	Pair 1	PEQ5_11	2.84	201	1.565	.110
		POQ11_1	2.92	201	1.555	.110

Paired Samples Correlations

EXP_COND_NUM			N	Correlation	Sig.
A	Pair 1	PEQ5_11 & POQ11_1	49	.676	.000
B	Pair 1	PEQ5_11 & POQ11_1	96	.547	.000
C	Pair 1	PEQ5_11 & POQ11_1	201	.514	.000

Paired Samples Test

EXP_COND_NUM			Paired Differences					t	df	Sig. (2-tailed)
			Mean	Std. Dev.	Std. Error Mean	95% CI of the Difference				
						Lower	Upper			
A	Pair 1	PEQ5_11 - POQ11_1	-1.020	1.315	.188	-1.398	-.643	-5.433	48	.000
B	Pair 1	PEQ5_11 - POQ11_1	-.073	1.558	.159	-.389	.243	-.459	95	.648
C	Pair 1	PEQ5_11 - POQ11_1	-.080	1.537	.108	-.293	.134	-.734	200	.464

4.4 Comparing task completion times across treatments

The second statistical analysis that was performed attempted to provide an answer to the following research question:

RQ_2: In a collaborative setting, would a group be able to sort a set of ideas faster, if it were broken up into smaller subsets and sorted individually by group members working in parallel, rather than working serially?

The corresponding hypothesis to this question stated:

H_2: The speed of sorting smaller subsets of a larger pool of ideas is significantly faster than sorting the entire set.

It is important to note once again that, due to the design of the experiment, the completion time for a particular subgroup is calculated by finding the maximum subgroup member's sort task completion time. In other words, for the participants in Condition B and Condition C, the subgroup's completion time is equal to the longest time of any of its members, since the study is attempting to replicate parallel processing in a distributed setting. All time figures reported are in seconds.

The analysis of this measure begins by examining the descriptive statistics of the completion times, which can be found in Table 12.

Table 12: Descriptive statistics of sort task completion time (by condition)

	N	Mean	Std. Dev.	Std. Error	95% Confidence Interval for Mean		MIN	MAX
					Lower Bound	Upper Bound		
A	50	1442	438	62.0	1317	1567	878	2758
B	50	1226	387	54.8	1116	1336	669	2631
C	75	1168	430	49.7	1069	1267	569	3171
Total	175	1263	434	32.8	1198	1328	569	3171

As expected, the various treatment conditions feature what appear to be clearly different mean times, with Condition A being the longest, and Condition C being the shortest (M_A = 1442, M_B = 1226, M_C = 1168). To test the statistical significance of these differences in completion times, a one-way ANOVA was conducted to explain the variance. Because the Levene test for equality of variance was not violated in the analysis, F(2,172) = 0.123, *p* = .884, we can assume that the variance is equal. The result of the ANOVA confirms that there are indeed significant differences in completion time between the three treatment conditions F(2, 172) = 6.626, *p* = .002. However, since the means of Condition B and Condition C are reasonably close, another set of ANOVA contrasts was performed to verify the significance of all three conditions. The details of those results can be found in Table 13below.

Table 13: Results of ANOVA (one-way) contrasts on sort completion time (by condition)

ANOVA

SORT_TIME

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		2349880	2	1174940	6.626	.002
	Linear Term	Unweighted	2256447	1	2256447	12.726	.000
		Weighted	2131982	1	2131982	12.024	.001
		Deviation	217898	1	217898	1.229	.269
Within Groups			30497661	172	177312		
Total			32847541	174			

Contrast Coefficients

Contrast	EXP_COND_NUM			
	A	B	C	
1	-1	0	1	(A vs. C)
2	0	-1	1	(B vs. C)
3	-1	1	0	(A vs. B)

Contrast Tests

Contrast			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)	
SORT_TIME	Assume equal variances	1	-274.25	76.879	-3.567	172	.000	(A vs. C)
		2	-58.55	76.879	-.762	172	.447	(B vs. C)
		3	-215.70	84.217	-2.561	172	.011	(A vs. B)
	Does not assume equal variances	1	-274.25	79.482	-3.451	103.746	.001	
		2	-58.55	73.998	-.791	112.334	.430	
		3	-215.70	82.804	-2.605	96.545	.011	

As these ANOVA contrasts showed, the differences between the sort task completion times across the three treatment conditions was only significant when another treatment condition was compared to Condition A.

According to the results, the completion time for Condition A was significantly longer than the time for Condition B – $t(172) = -3.567$, $p < .001$ – and Condition A was significantly longer than Condition B – $t(172) = -2.561$, $p = .011$. The ANOVA contrast between Condition B and Condition C showed that the completion times for those treatments was not significantly different – F(172) = -0.762, $p = 0.447$.

Thus, these results partially support the second hypothesis – sorting smaller subsets is faster than sorting the entire set, but the time differences between the two "partial subset" treatments in this experiment (Conditions B and C) were not significant.

4.5 Comparing sort effectiveness (NCE) across treatments

The final statistical analysis that was performed attempted to provide an answer to the following research question:

> *RQ_3: If a group's members sorted smaller, distinct subset of ideas in parallel, would this adversely affect the effectiveness of the overall result?*

The corresponding hypothesis to this question stated:

> *H_3: Sorting smaller subsets of a larger pool of ideas is as least as effective as sorting the entire set.*

As discussed earlier, the effectiveness of a subgroup's sorted results is measured by the calculation of a normalized clustering error (NCE) score for their work product, when compared to a "gold standard" result. This provides an objective measure of the sort quality and effectiveness – where an NCE score of 0 indicates a perfectly matched sort to the "gold standard" (i.e., a high quality sort), and an NCE score of 1 indicates a poor quality sort that is completely dissimilar to the ideal metric. Table 14 shows the descriptive statistics of the subgroups' NCE scores, by condition.

The mean NCE scores of the various sorted results shows distinct differences between the three treatment conditions – on average, the participants in Condition A created higher-quality sorts than the participants in Condition B, and Condition C's participants produced the lowest-quality sorts in this study, in terms of the NCE scores ($M_A = 0.748$, $M_B = 0.796$, $M_C = 0.812$).

Table 14: Descriptive statistics of the NCE values (by treatment condition)

	N	Mean	Std. Dev.	Std. Error	95% Confidence Interval for Mean		MIN	MAX
					Lower Bound	Upper Bound		
A	49	0.748	0.082	0.012	0.725	0.772	0.538	0.867
B	49	0.796	0.037	0.005	0.785	0.807	0.701	0.873
C	75	0.812	0.050	0.006	0.800	0.823	0.583	0.926
Total	173	0.789	0.064	0.005	0.780	0.799	0.538	0.926

To test the statistical significance of these differences in NCE scores, a one-way ANOVA was attempted to explain the variance, but because the Levene test for equality of variance was violated in this analysis, F(2,170) = 14.4, $p < .001$, we cannot assume that the variance is equal. So, a series of ANOVA contrasts were again employed, and these contrasts identified significant differences between the NCE measures of sorting effectiveness between the three treatment conditions.

According to the results, the participants in Condition A were able to produce significantly higher-quality sorts than the participants in Condition B – $t(67.1) = 3.70$, $p < .001$ – and Condition C – $t(71.1) = 4.86$, $p < .001$. Additionally, the sorted product of Condition B's participants was also significantly better than the people in Condition C – $t(119.5) = 2.021$, $p = .046$. The summary results are shown in Table 15: Results of ANOVA (one-way) contrasts on sort effectiveness (by condition).

On the basis of this analysis, the initial hypothesis (H_3) would be rejected. The results appear to indicate that the quality of a sorted list of items is diminished when a member is given smaller subset of ideas to process. Thus, the group is more effective when its members are forced to sort longer lists of items.

Table 15: Results of ANOVA (one-way) contrasts on sort effectiveness (by condition)

SORT_NCE

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		.123	2	.061	18.172	.000
	Linear Term	Unweighted	.120	1	.120	35.453	.000
		Weighted	.114	1	.114	33.742	.000
		Deviation	.009	1	.009	2.602	.109
Within Groups			.574	170	.003		
Total			.696	172			

Contrast Coefficients

Contrast	EXP_COND_NUM			
	A	B	C	
1	-1	0	1	(A vs. C)
2	0	-1	1	(B vs. C)
3	-1	1	0	(A vs. B)

Contrast Tests

Contrast			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)	
SORT_NCE	Assume equal variances	1	.0635	.0106	5.954	170	.000	
		2	.0158	.0106	1.486	170	.139	
		3	.0476	.0117	4.063	170	.000	
	Does not assume equal variances	1	.0635	.0130	4.866	71.172	.000	(A vs. C)
		2	.0158	.0078	2.021	119.545	.046	(B vs. C)
		3	.0476	.0128	3.701	67.101	.000	(A vs. B)

5 Discussion

This study represents an attempt to investigate the effects of cognitive load experienced by the members of a collaborative group working on a sorting task, with respect to the quality of their effort and the time required. The goal was to identify methods and strategies that help maintain higher satisfaction levels throughout the convergence process, thus enabling groups to make better decisions faster.

The experiment conducted in this study was specifically designed to simulate a “distributed parallel sort” which would expose the participants to conditions that imposed variant levels of cognitive demand as a result of the sorting task, achieved by controlling the number of items to be sorted.

The primary method for measuring the cognitive load experienced by a participant that was explored in this study was the raw (unweighted) variant of NASA’s TLX instrument, and NCE was used to determine the effectiveness of the participants’ sorted results, as measured against a “gold standard.” Table 16 below summarizes the statistical results of the empirical data gathered in the experiment.

Table 16: Summary of Hypothesis Support

	Hypothesis	Experimental Results	
		Measure	**Support?**
H_1	An individual’s perception of the cognitive load associated with sorting sets of ideas is positively related to the number of items in the set to be sorted.	NASA-TLX	No
		Difficulty	Partial (A vs. B and A vs. C)
		Comfort	Yes
		Fatigue	Partial (A vs. B and A vs. C)
H_2	The speed of sorting smaller subsets of a larger pool of ideas is significantly faster than sorting the entire set.	Sort time	Partial (A vs. B and A vs. C)
H_3	Sorting smaller subsets of a larger pool of ideas is as least as effective as sorting the entire set.	NCE	No

The results confirm that, to some degree, the three treatment conditions utilized did indeed vary the levels of cognitive load experienced by the participants. However, the particular implementation of the NASA-TLX instrument was not very effective in measuring or amplifying these differences. The three manipulation check variables, individually serving as a proxy, were much more effective at highlighting the symptoms of increased cognitive load in this environment, particularly in relation to Condition A (where participants sorted an entire list of 110 items).

Thus, the first key contribution of this study is that the objective measurement of cognitive load in collaboration environments can be accomplished with simple self-report data, but identifying the optimal method for more accurately measuring that load will require more scientific investigation. The next section will discuss this conundrum in more detail.

The second contribution of this study is that designers (or facilitators) of collaborative sessions are faced with a trade-off between time and effectiveness when they are planning activities to achieve convergence: The results of this study imply that individuals who sorted the full list of items were more effective (i.e., generated higher-quality results) than those who sorted pieces (either one-half or one-third) of the full list together as a group. However, more time is required for an individual to complete a full sort. While this trade-off is nothing new and applies to virtually any complex task that can be divided amongst other individuals, the counter-intuitive implication here is that more people do not necessarily improve performance in collaborative convergence – even though the opposite is true in collaborative divergence (brainstorming, for example). Clearly, a "flex point" must exist in convergence-based activities at which both time and effectiveness are maximized – but identifying that optimum value will likely be dependent upon a variety of contextual factors which must be successfully evaluated by the designer/facilitator of the collaborative session.

Both of these contributions (and their practical implications) will be discussed in greater detail later in the "future research" section.

5.1 Limitations

The unweighted NASA-TLX scores were insufficient in their raw format, to highlight any discrete differences in the resultant task-based cognitive load that may have been experienced by participants as a result of the experimental task. There are many potential reasons that might explain this failure, but there are four explanations that deserve more consideration at this point. They are, in order of importance:

- The relationship between the cognitive load experienced by an individual performing a sorting task and the number of items to be sorted may be non-linear – those who sort a full list of items also have a full set of contextual information at their disposal to aid in their sort, while the lack of that information (which could be observed in anyone sorting a partial list of items) may impose a confounding, higher level of cognitive load and that load could possibly increase as the number of items decreases;
- The task of having the participants perform a pairwise comparison of the components in the NASA-TLX is more critical to the instrument's effectiveness in this experimental environment than initially believed;
- The computer-based delivery method is not as effective in measuring the cognitive load of this type of task as the pencil-and-paper delivery method, as some researchers have observed in other research contexts before; and
- The NASA-TLX tool may not be as appropriate or effective as other measurement strategies in this context.

It is the opinion of the author that the primary limitation of this research is fundamentally the unknown true relationship between cognitive load and convergence task design. A better measurement instrument is required to more accurately discern the levels of cognitive load a particular task imposes upon a collaborative participant, but evaluating those instruments may prove to be problematic, if the experimental designs are not carefully created and delivered consistently. Perhaps the design or the context of the task used in this experiment confounded the results, and the NASA-TLX tool measured the cognitive load correctly?

The implications of this notion will be discussed in the next section.

5.2 Future research

The principal ramification of this exploratory research on the future academic study of collaborative convergence is highlighted in the failure of its central measurement instrument -- a well-regarded, heavily-researched method for measuring cognitive load. Despite its scientific acceptance and use, the NASA-TLX failed to yield statistically significant results. Meanwhile, the simple manipulation checks employed in the experiment accomplished what the more complicated tool could not, and provided significant results.

Perhaps the biggest contribution of this work lies in the lessons learned as a result of this analytical shortcoming. With that in mind, the remainder of this discussion will be dedicated to describing a suggested research road map that could be followed to enable more extensive inquiry of the relationship between cognitive load and group effectiveness in collaborative convergence tasks.

5.2.1 Measuring cognitive load in collaboration environments

Accurately measuring the cognitive demands imposed upon an individual by the tasks in a collaborative group context is absolutely critical to the future success of this research. Although the NASA-TLX instrument (as implemented in this particular experiment) was ultimately ineffective in this study, there are other ways to implement the tool that may offer improved results.

Thus, the next step in this research stream is to exhaust all of the opportunities that the tool affords, beginning with adding the pairwise comparison of the component measures to the existing experimental design. Perhaps the findings of this experiment would have been different, had those questions been asked of the participants.

Additional investigation into the NASA-TLX delivery method might also be a prudent investment of time. It seems odd that an agency of NASA's caliber, technologically-speaking, would resort to a pencil-and-paper version of a survey tool, but it appears that they used that manual approach for many years. Regardless of how the delivery method is tested – quantitatively in the lab, or using qualitative means like personal interviews with NASA research personnel – another important step in this research is to determine the optimal implementation method of the tool for use in the collaborative environment. Additionally, an ancillary research project could be to evaluate the use of an electronic "rank-order" tool (like the feature in GroupSystems' ThinkTank software) in implementing the instrument, rather than asking the prescribed 15 questions usually associated with the weighted instrument.

However, if all of the variations of the NASA-TLX prove to be ineffective in the collaborative space, then other measurement techniques have to be considered. In today's society, invasive means of measurement are strictly a last resort, but those could have tremendous power and add significantly to this topic of interest in the distant future. Yet given the moderate success of the simple manipulation check variables in this study in detecting changes in cognitive load, those alternatives are not a priority as of this point in time – there are certainly better non-invasive questions to ask (or symptoms of increased cognitive load to monitor for) that could yield similar desired results.

5.2.2 Designing experimental treatments for convergence activities

As the process of measuring cognitive load in collaborative environments is improved and made more reliable, the research path should evolve to focus on the design of the collaborative convergence activities themselves. Empirical tests of the optimal strategies for original list fragmentation are likely the next most important area of scientific inquiry. For example, the experimental conditions featured in this study were comprised of groups that sorted 100%, 50% or 33% of the original list items – but what other factors might influence the ideal quantity of a distributed parallel sort design? Are there ordinal minima and maxima to the number of items that are independent of a particular environment or context?

It is logical to assume that there is an optimal amount of information that a group member would need to perform an effective, high-quality sort. Other collaboration researchers have found that certain ranges seem to provide optimal results in their prior efforts, but it is likely that there could have moderating variables present in their experiments that affected their results. For example, one could posit that if a set of feedback items was generally unfamiliar to a particular individual, would the inclusion of additional contextual information help them generalize a set of information to a scenario with which they were familiar? In other words, an expert in business process reengineering doesn't always know the operational details of a particular environment that they have no first-hand experience with, but if they were given enough information to be able to glean enough about that environment's function and identify its problems, they could provide invaluable feedback and guidance to the group. But at what point does the unfamiliar become familiar? Does that point change with age and/or experience?

The experiment in this study featured a data set that addressed the challenges observed by the students of a university's business school – and this was an appropriate set of familiar information, given that the participants of this experiment were students in that same business school (albeit years removed from the original group that generated those items). Would the results of this study have been markedly different if they had been presented with a list of aerospace engineering requirements or another foreign concept to them? This is a valid question because it directly influences the quality of their sorting performance.

Thus, future experiments in this research stream must carefully consider the contextual environment of their design, as it could yield dramatically different results. As future experimental designs are crafted in this path, the researcher might want to test the elasticity of familiar vs. unfamiliar topics in terms of cognitive load, sorting time, and work product efficiency.

Another topic of interest might be to test the cognitive load differences that are associated with the various types of sorting strategies, such as "open sorts" (where the participant must create the sort categories on their own, as was the case in this experiment), "closed sorts" (where the participant is provided with a pre-defined set of labeled categories), or “hybrid sorts” (where the participant is given some suggested categories, but are free to add/edit/delete categories as they see fit).

One last area of inquiry would involve testing the sensitivity of cognitive load across the various synchronicities and CMC modalities. Is the cognitive load associated with a sort process different in a synchronous, face-to-face meeting than in a distributed, asynchronous text-only context? What if other communication backchannels (like text-based chat, shared audio, or video conferencing) were employed in the collaborative session? Could this affect satisfaction levels and the perception of cognitive load when attempting to complete convergence tasks?

Clearly, there are myriad opportunities available in this research space regarding the relationship of the impact of cognitive load and collaborative convergence activities. Currently, there are many more questions than answers. However, this research path could yield some very beneficial results to society in the future – as the desire to make better decisions faster is a universal desire in every intelligent being.

5.3 Practical recommendations

While the research in this area continues, there are a few “suggested practices” that could be employed by collaborative activity designers and facilitators to help mitigate the decline in group member satisfaction during convergence tasks, and sorting tasks in particular. These suggestions are the result of the author’s first-hand experience, after hundreds of hours facilitating collaborative GSS sessions with a wide variety of audiences and an even broader array of collaborative environments and topics.

5.3.1 The “sort-your-own” method of facilitating convergence

The first method proposed to help mitigate the drop in satisfaction levels and improve a group’s effectiveness during the idea organization phase involves a simple piece of guidance given by a facilitator to the group at the start of a sorting activity – sort your own ideas.

Once the group members have entered in a sufficient amount of feedback to complete the idea generation phase, they could be instructed to only move or process the ideas they personally contributed.

The rationale for providing this guidance directly addresses the top three explanations that have been suggested as the root cause in the drop in user satisfaction (as mentioned earlier in the introduction):

- Users don't like receiving critical/negative feedback regarding their ideas;
- Users don't like seeing their contributions "diluted," or "lumped in" with other ideas;
- Users are intimidated by the cognitive difficulty of sorting large sets of feedback data; and

First of all, by forcing the group members to move their own ideas, the initial cognitive difficulty of the sorting task is reduced immediately. There is no initial feeling of "task intimidation" that often arises suddenly when instructed to critically evaluate dozens or hundreds of comments all made by other anonymous members – all a member has to do is re-read their own submission and find or create a suitable category for it. This significantly reduces the amount of work for each participant because it is much faster for them to do – they've already read and fully understand their own comments.

Additionally, in this method, the group members are allowed to retain a sense of "ownership" of each of their ideas for a while longer, and delays (even if only for a few seconds) any indication of an evaluation on behalf of the group that might be perceived by the author as dismissive or insulting – thus, this should help maintain everyone's current level of satisfaction with the process. Furthermore, by enabling them to create their own category names, they can ensure that the spirit of their ideas (not just the letter or words) isn't diluted right away. The author of a particular comment entered in a feedback session might be the only one who really understands the background or motivation of that comment – and if another member moved it haphazardly (after only a cursory glance, without really understanding its spirit or importance) might cause the original author of the comment to feel somewhat slighted. An individual's level of satisfaction with the collaborative process will decrease slightly every time a comment of theirs is moved into a category that doesn't match the author's intent (or is treated in manner that is beneath the author's expectations).

This is yet another area that merits further research (although it would be relatively costly and time-consuming to do): Empirically test the "sort-your-own" method against other sorting strategies and compare the resultant levels of cognitive load, satisfaction, and effectiveness.

5.3.2 Technology-enabled sort aggregation to generate a "consensus sort"

Work is currently underway on a new technology-enabled approach to refining sort results that holds some promise in terms of addressing the trade-off mentioned earlier in this discussion section – it is a novel method that incorporates two computer science algorithms to approximate the best cases of both dimensions of that trade-off (sort quality and sort completion time).

If this proposed approach ultimately proves successful, it would allow collaborative activity designers to employ a distributed work paradigm (similar to the experiment's Condition C) on a full set of feedback using a limited number of ad-hoc subgroups working independently. Doing so would enable groups to produce a high-quality sort of a set of items (comparable to the results achieved in Condition A, the best) in a minimal amount of time (comparable to the results achieved in Condition C, the fastest).

The implementation of this approach aggregates multiple full sorts of a list of items (created by asking group members to sort a random small subset of the original ideas, in ad-hoc groups) to generate a "consensus sort." To generate an initial aggregated cluster of sort results, two steps are required. First, the individual full sorts are combined using the Cluster-based Similarity Partitioning Algorithm (CSPA) developed by Strehl and Ghosh (2002). The CSPA counts the total number of links between two brainstorming ideas and creates a final matrix showing these counts. The second step in the process is to partition the similarity matrix into discrete clusters. Using the METIS program (Karypis and Kumar, 1998), the matrix is split into discrete clusters and the resulting groups are re-populated with the actual corresponding items from the list, which becomes the "consensus sort."

Of course, this approach will require extensive research to validate its external generalizability and practicality, but the foundation of the applied algorithms appears to be fundamentally sound.

6 Conclusion

This research explored the relationship between the demands of a difficult task (sorting feedback in the “idea organization” phase of a collaborative GSS session) and the success or effectiveness of the sorted results. The experiment that was designed and conducted is an approximation of a “distributed parallel sort” with treatment conditions that varied the number of items to be sorted by each participant. A variant of the NASA Task Load Index (NASA-TLX) was employed in an attempt to objectively measure each participant’s self-reported levels of cognitive load resulting from each experimental condition.

The NASA-TLX instrument’s measurements were insufficient to produce significant findings, but statistical analysis of several manipulation checks (put in place to verify the symptoms of increased cognitive load) was able to confirm that sorting a very long list of 110 items imposed a significantly higher level of cognitive load than sorting a smaller list of 55 or 37 items.

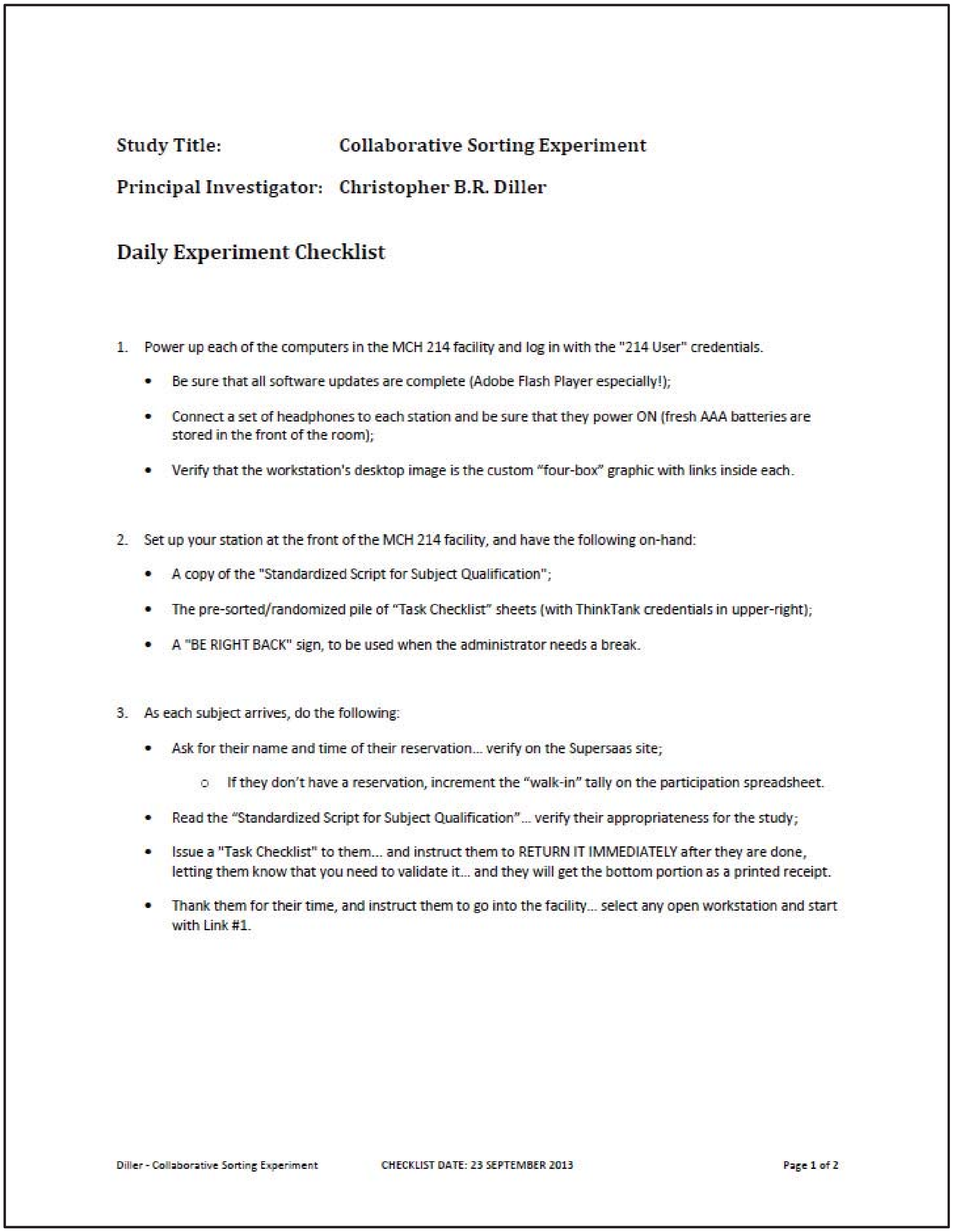
The results of the experiment also indicated that it is indeed faster to break up and distribute a long list of items for a sorting task, but that the resultant sorts are of lower quality than the work product of individuals forced to sort the entire list.

The work’s primary contribution to collaboration research is a roadmap for further study into the measurement of cognitive load in GSS environments. Despite the failings of the unweighted NASA-TLX data collected in this initial experiment, there are other weighted variants and delivery methods that may prove to be successful in the future. Additionally, this research describes several critical session design issues that must be considered as this research stream continues in its attempts to improve collaborative processes in general (regardless of synchronicity, proximity, or facilitation modality).

The research also highlights two practical facilitation methods for mitigating the decline in participant satisfaction often experienced by group members during the idea organization phase – the “sort-your-own” method for facilitating convergence tasks, and a technology-enabled method for generating a consensus result by using a distributed sort task. While both methods lack empirical support and require additional testing, they both have shown promise to date.

APPENDIX A – DAILY EXPERIMENT OPERATIONS GUIDE

To insure the consistency of the participant experience, the two-page document below was used:



Study Title: **Collaborative Sorting Experiment**

Principal Investigator: **Christopher B.R. Diller**

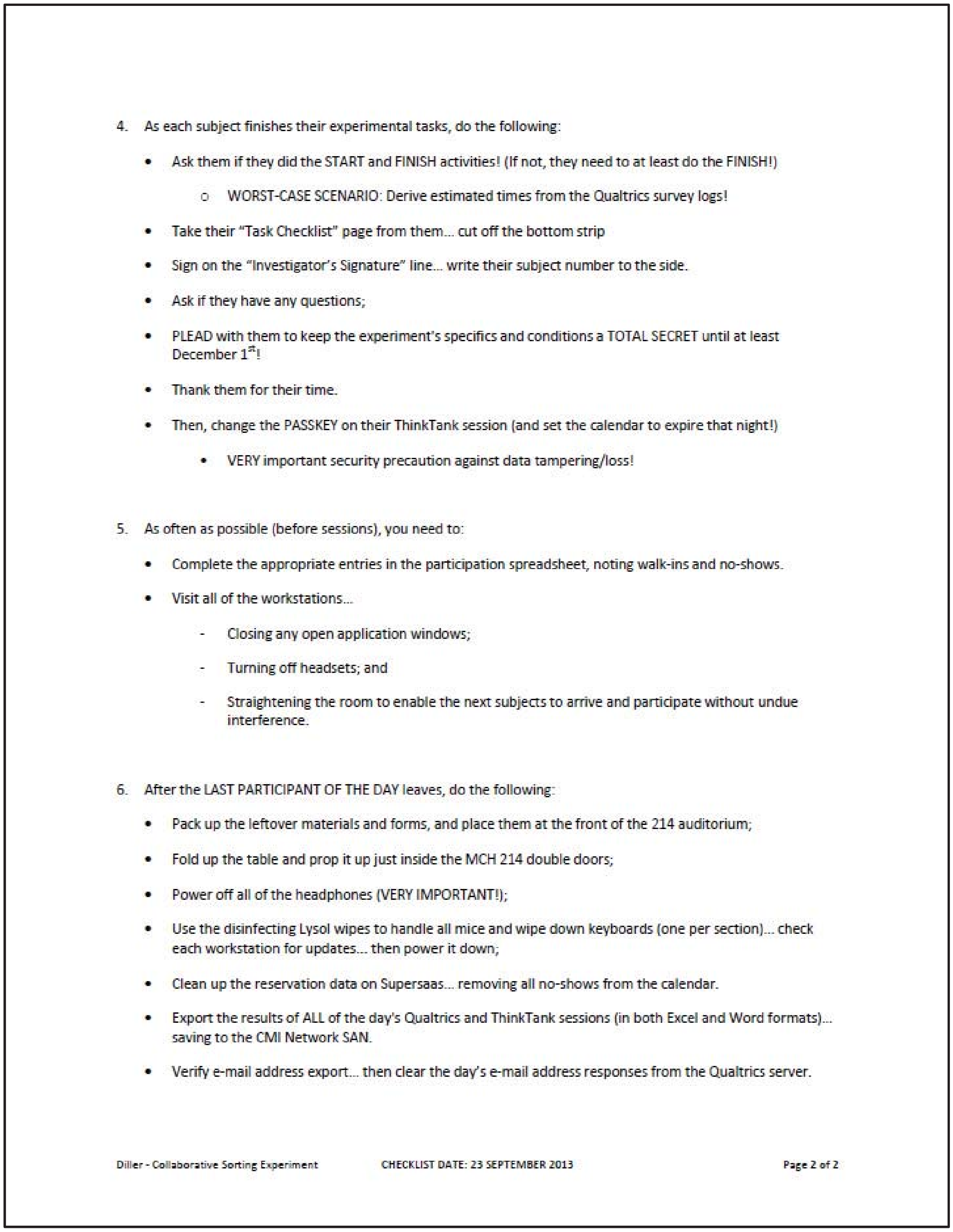
Daily Experiment Checklist

1. Power up each of the computers in the MCH 214 facility and log in with the "214 User" credentials.
   - Be sure that all software updates are complete (Adobe Flash Player especially!);
   - Connect a set of headphones to each station and be sure that they power ON (fresh AAA batteries are stored in the front of the room);
   - Verify that the workstation's desktop image is the custom "four-box" graphic with links inside each.

2. Set up your station at the front of the MCH 214 facility, and have the following on-hand:
   - A copy of the "Standardized Script for Subject Qualification";
   - The pre-sorted/randomized pile of "Task Checklist" sheets (with ThinkTank credentials in upper-right);
   - A "BE RIGHT BACK" sign, to be used when the administrator needs a break.

3. As each subject arrives, do the following:
   - Ask for their name and time of their reservation... verify on the Supersaas site;
     - If they don't have a reservation, increment the "walk-in" tally on the participation spreadsheet.
   - Read the "Standardized Script for Subject Qualification"... verify their appropriateness for the study;
   - Issue a "Task Checklist" to them... and instruct them to RETURN IT IMMEDIATELY after they are done, letting them know that you need to validate it... and they will get the bottom portion as a printed receipt.
   - Thank them for their time, and instruct them to go into the facility... select any open workstation and start with Link #1.

Diller - Collaborative Sorting Experiment CHECKLIST DATE: 23 SEPTEMBER 2013 Page 1 of 2



4. As each subject finishes their experimental tasks, do the following:
   - Ask them if they did the START and FINISH activities! (If not, they need to at least do the FINISH!)
     - WORST-CASE SCENARIO: Derive estimated times from the Qualtrics survey logs!
   - Take their "Task Checklist" page from them… cut off the bottom strip
   - Sign on the "Investigator's Signature" line… write their subject number to the side.
   - Ask if they have any questions;
   - PLEAD with them to keep the experiment's specifics and conditions a TOTAL SECRET until at least December 1st!
   - Thank them for their time.
   - Then, change the PASSKEY on their ThinkTank session (and set the calendar to expire that night!)
     - VERY important security precaution against data tampering/loss!

5. As often as possible (before sessions), you need to:
   - Complete the appropriate entries in the participation spreadsheet, noting walk-ins and no-shows.
   - Visit all of the workstations…
     - Closing any open application windows;
     - Turning off headsets; and
     - Straightening the room to enable the next subjects to arrive and participate without undue interference.

6. After the LAST PARTICIPANT OF THE DAY leaves, do the following:
   - Pack up the leftover materials and forms, and place them at the front of the 214 auditorium;
   - Fold up the table and prop it up just inside the MCH 214 double doors;
   - Power off all of the headphones (VERY IMPORTANT!);
   - Use the disinfecting Lysol wipes to handle all mice and wipe down keyboards (one per section)… check each workstation for updates… then power it down;
   - Clean up the reservation data on Supersaas… removing all no-shows from the calendar.
   - Export the results of ALL of the day's Qualtrics and ThinkTank sessions (in both Excel and Word formats)… saving to the CMI Network SAN.
   - Verify e-mail address export… then clear the day's e-mail address responses from the Qualtrics server.

APPENDIX B – SCRIPT FOR SUBJECT QUALIFICATION

Thank you for your interest in this research project. To keep things consistent, I'm reading from this script.

This research experiment is expected to last for one hour. If you decide to participate, you will be asked to complete a couple of questionnaires, watch a brief training video on how to use a software application, and then use that software to sort a list of items.

So, before we begin, I have to ask the following:

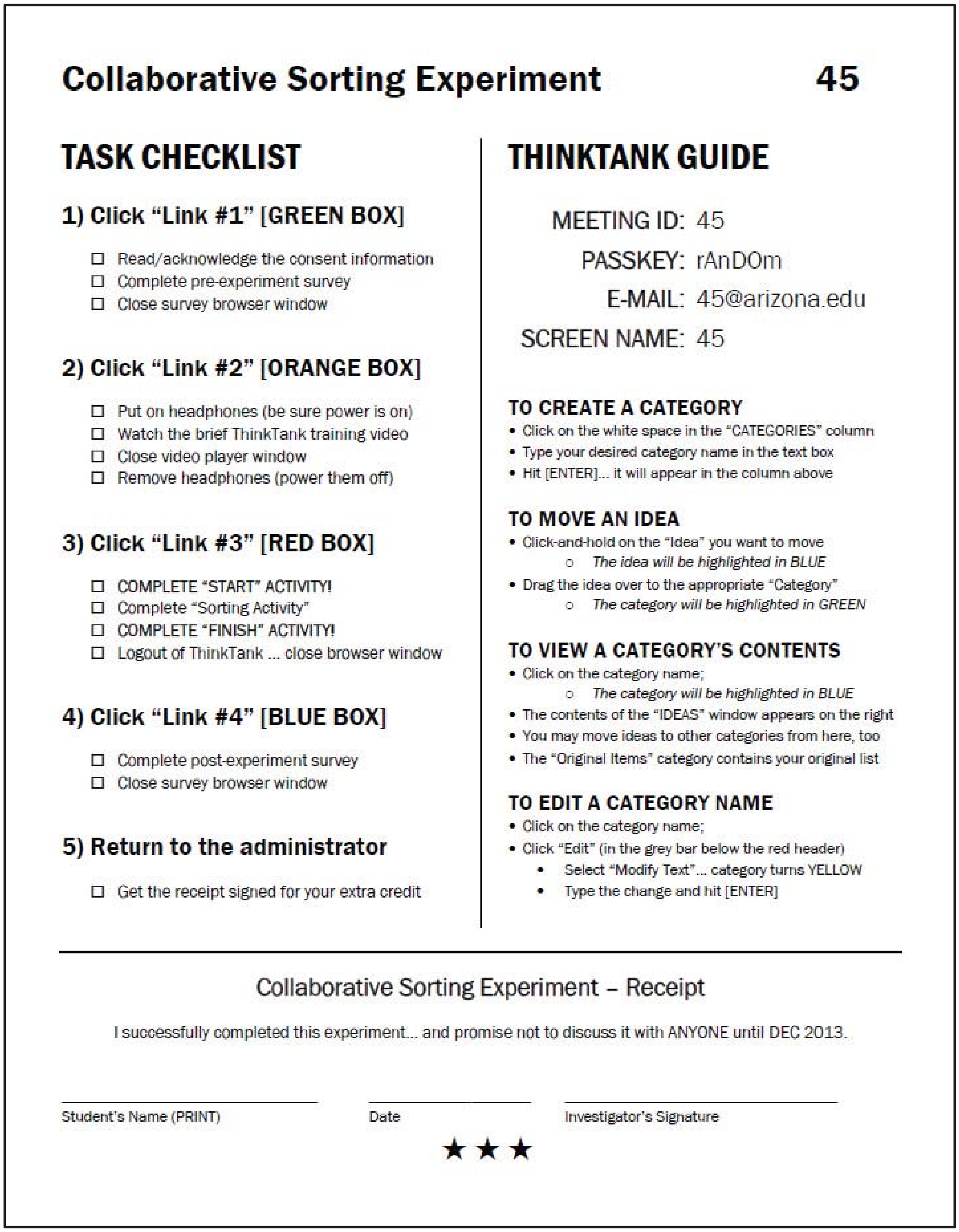
- Are you 18 years of age or older?
- Do you have the ability to read and write (or type) in English?
- Are you comfortable with the idea of performing the tasks that I have described?
- Do you have any visual impairment that might prohibit you from working on a computer?
- (OPTIONAL) I can accommodate some assistive technology requests, if you need me to.
- Do you have any cognitive impairment that might prevent you from completing these tasks?

[IF YES TO ALL] Thank you. Please take this piece of paper, it lists (step-by-step) everything you will need to do. Also, you will see that it has a number and a password that you will need for the tasks. Just take any available seat inside the auditorium… and follow the instructions on the paper. Do you have any other questions?

[IF NO TO ANY] Thank you for your interest, but the demands of this particular research project are not compatible with your situation. If you are here for extra credit for a course, there will be other opportunities for you throughout the term. I am sorry, but I hope you understand. Do you have any other questions?

APPENDIX C – PARTICIPANT TASK CHECKLIST

Upon arrival, each participant in the study was given a page of instructions, like the one below:



Collaborative Sorting Experiment 45

TASK CHECKLIST

1) Click "Link #1" [GREEN BOX]

- □ Read/acknowledge the consent information
- □ Complete pre-experiment survey
- □ Close survey browser window

2) Click "Link #2" [ORANGE BOX]

- □ Put on headphones (be sure power is on)
- □ Watch the brief ThinkTank training video
- □ Close video player window
- □ Remove headphones (power them off)

3) Click "Link #3" [RED BOX]

- □ COMPLETE "START" ACTIVITY!
- □ Complete "Sorting Activity"
- □ COMPLETE "FINISH" ACTIVITY!
- □ Logout of ThinkTank ... close browser window

4) Click "Link #4" [BLUE BOX]

- □ Complete post-experiment survey
- □ Close survey browser window

5) Return to the administrator

- □ Get the receipt signed for your extra credit

THINKTANK GUIDE

MEETING ID: 45
PASSKEY: rAnDOm
E-MAIL: 45@arizona.edu
SCREEN NAME: 45

TO CREATE A CATEGORY

- Click on the white space in the "CATEGORIES" column
- Type your desired category name in the text box
- Hit [ENTER]... it will appear in the column above

TO MOVE AN IDEA

- Click-and-hold on the "Idea" you want to move
  - *The idea will be highlighted in BLUE*
- Drag the idea over to the appropriate "Category"
  - *The category will be highlighted in GREEN*

TO VIEW A CATEGORY'S CONTENTS

- Click on the category name;
  - *The category will be highlighted in BLUE*
- The contents of the "IDEAS" window appears on the right
- You may move ideas to other categories from here, too
- The "Original Items" category contains your original list

TO EDIT A CATEGORY NAME

- Click on the category name;
- Click "Edit" (in the grey bar below the red header)
  - Select "Modify Text"... category turns YELLOW
  - Type the change and hit [ENTER]

Collaborative Sorting Experiment – Receipt

I successfully completed this experiment... and promise not to discuss it with ANYONE until DEC 2013.

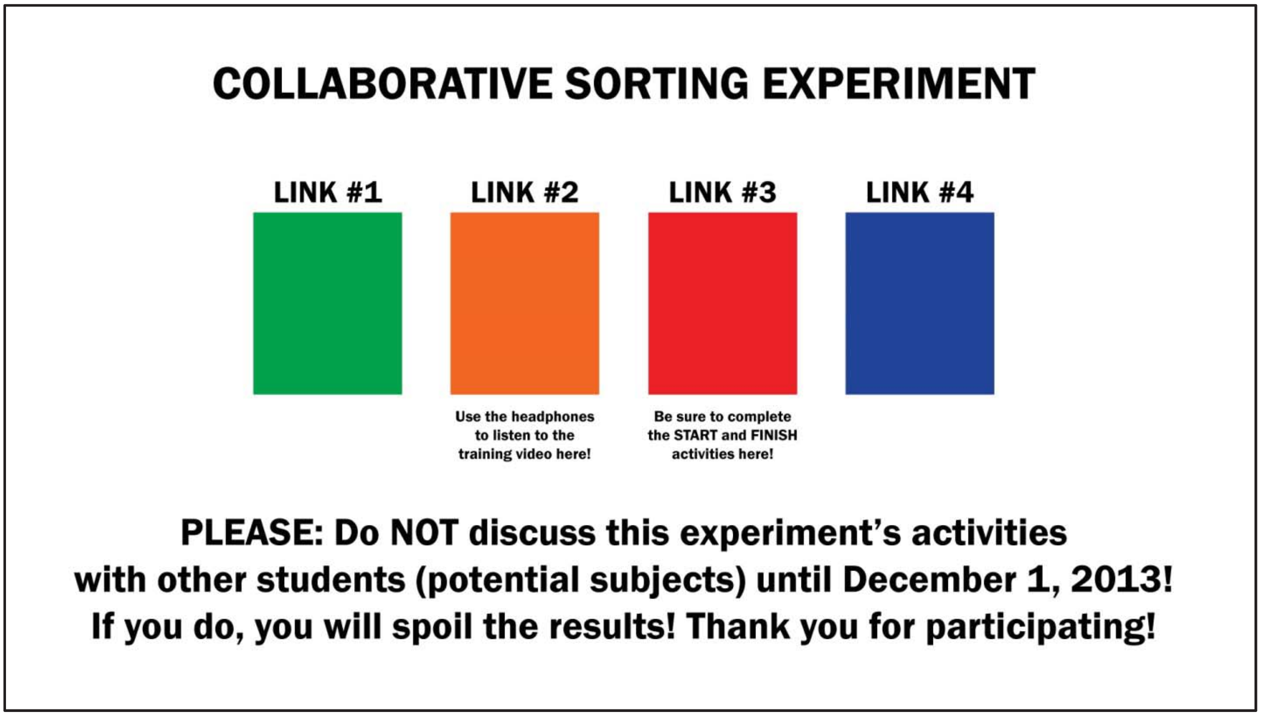
Student's Name (PRINT) ____________ Date ____________ Investigator's Signature ____________

★ ★ ★

APPENDIX D – PARTICIPANT WORKSTATION DESKTOP

The following image was used as the "desktop wallpaper" on all workstations utilized in the study. Desktop shortcuts (link icons) were positioned in the center of each of the four squares to make it clear which icon needed to be clicked at each step.

This was implemented in order to make the instructions more clear to the participant and make participant errors less likely to occur.



COLLABORATIVE SORTING EXPERIMENT

LINK #1

LINK #2

Use the headphones
to listen to the
training video here!

LINK #3

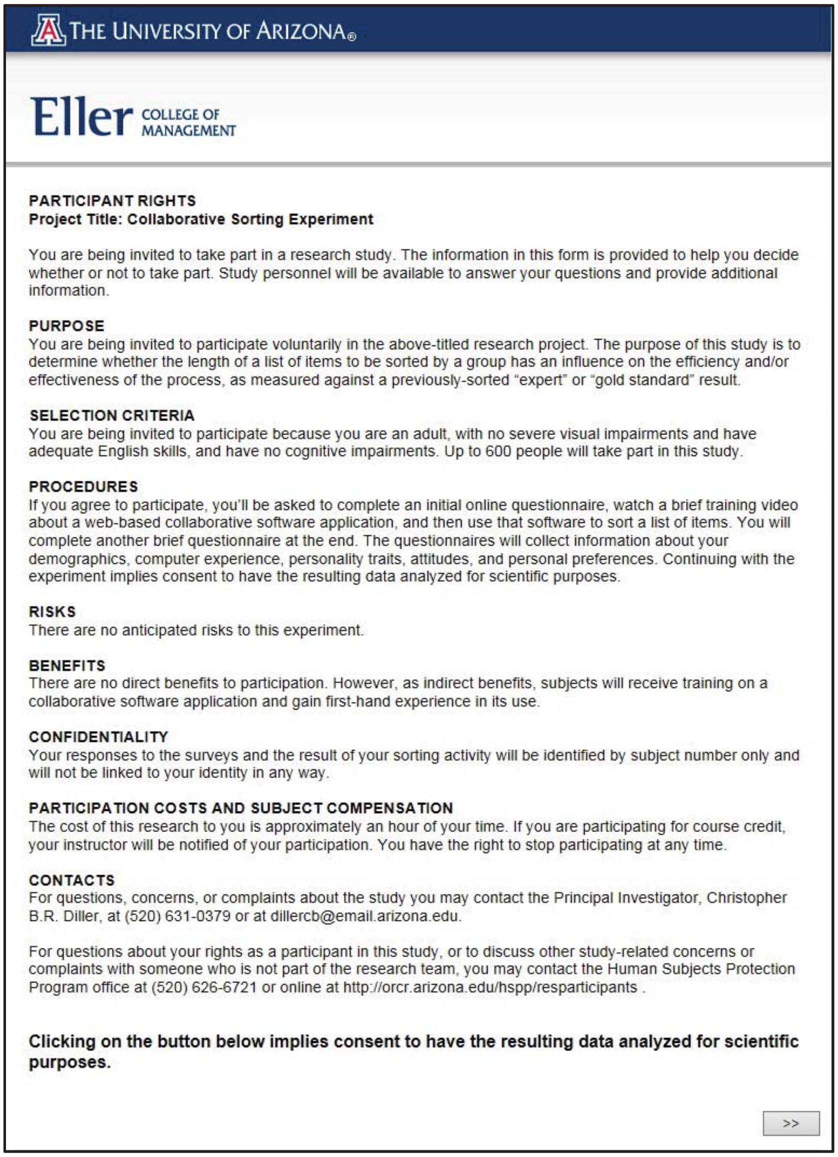
Be sure to complete
the START and FINISH
activities here!

LINK #4

PLEASE: Do NOT discuss this experiment's activities
with other students (potential subjects) until December 1, 2013!
If you do, you will spoil the results! Thank you for participating!

APPENDIX E – PRE-EXPERIMENT SURVEY QUESTIONS

The following screenshots show the online pre-experiment survey each participant was required to complete. Each image represents the successive screens displayed after a participant's prior responses were validated.



PARTICIPANT RIGHTS
Project Title: Collaborative Sorting Experiment

You are being invited to take part in a research study. The information in this form is provided to help you decide whether or not to take part. Study personnel will be available to answer your questions and provide additional information.

PURPOSE
You are being invited to participate voluntarily in the above-titled research project. The purpose of this study is to determine whether the length of a list of items to be sorted by a group has an influence on the efficiency and/or effectiveness of the process, as measured against a previously-sorted "expert" or "gold standard" result.

SELECTION CRITERIA
You are being invited to participate because you are an adult, with no severe visual impairments and have adequate English skills, and have no cognitive impairments. Up to 600 people will take part in this study.

PROCEDURES
If you agree to participate, you'll be asked to complete an initial online questionnaire, watch a brief training video about a web-based collaborative software application, and then use that software to sort a list of items. You will complete another brief questionnaire at the end. The questionnaires will collect information about your demographics, computer experience, personality traits, attitudes, and personal preferences. Continuing with the experiment implies consent to have the resulting data analyzed for scientific purposes.

RISKS
There are no anticipated risks to this experiment.

BENEFITS
There are no direct benefits to participation. However, as indirect benefits, subjects will receive training on a collaborative software application and gain first-hand experience in its use.

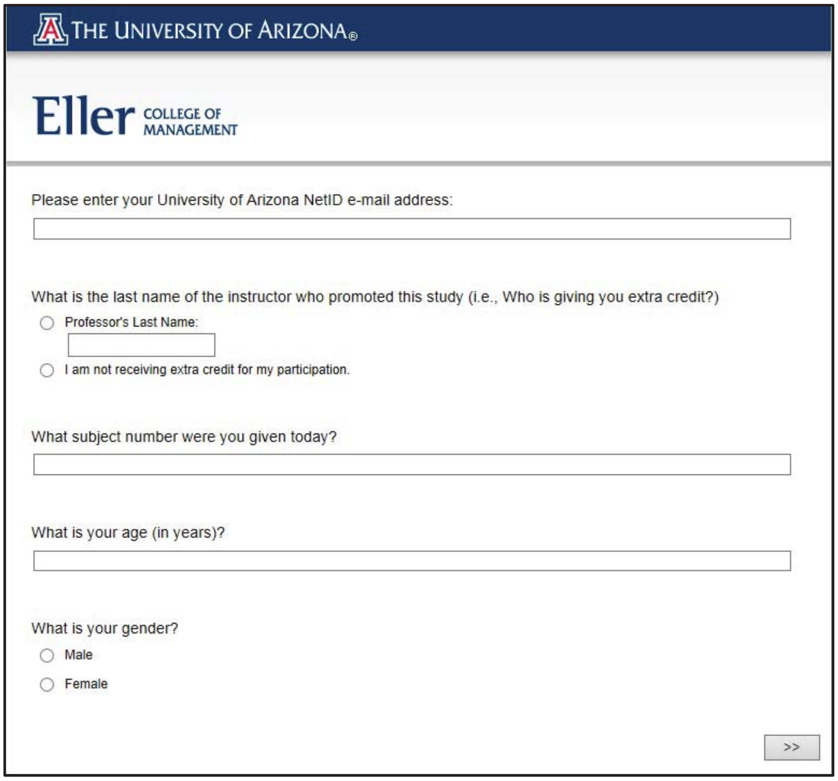
CONFIDENTIALITY
Your responses to the surveys and the result of your sorting activity will be identified by subject number only and will not be linked to your identity in any way.

PARTICIPATION COSTS AND SUBJECT COMPENSATION
The cost of this research to you is approximately an hour of your time. If you are participating for course credit, your instructor will be notified of your participation. You have the right to stop participating at any time.

CONTACTS
For questions, concerns, or complaints about the study you may contact the Principal Investigator, Christopher B.R. Diller, at (520) 631-0379 or at dillercb@email.arizona.edu.

For questions about your rights as a participant in this study, or to discuss other study-related concerns or complaints with someone who is not part of the research team, you may contact the Human Subjects Protection Program office at (520) 626-6721 or online at http://orcr.arizona.edu/hspp/resparticipants .

Clicking on the button below implies consent to have the resulting data analyzed for scientific purposes.



THE UNIVERSITY OF ARIZONA®

Eller COLLEGE OF MANAGEMENT

Please enter your University of Arizona NetID e-mail address:

What is the last name of the instructor who promoted this study (i.e., Who is giving you extra credit?)

○ Professor's Last Name:

○ I am not receiving extra credit for my participation.

What subject number were you given today?

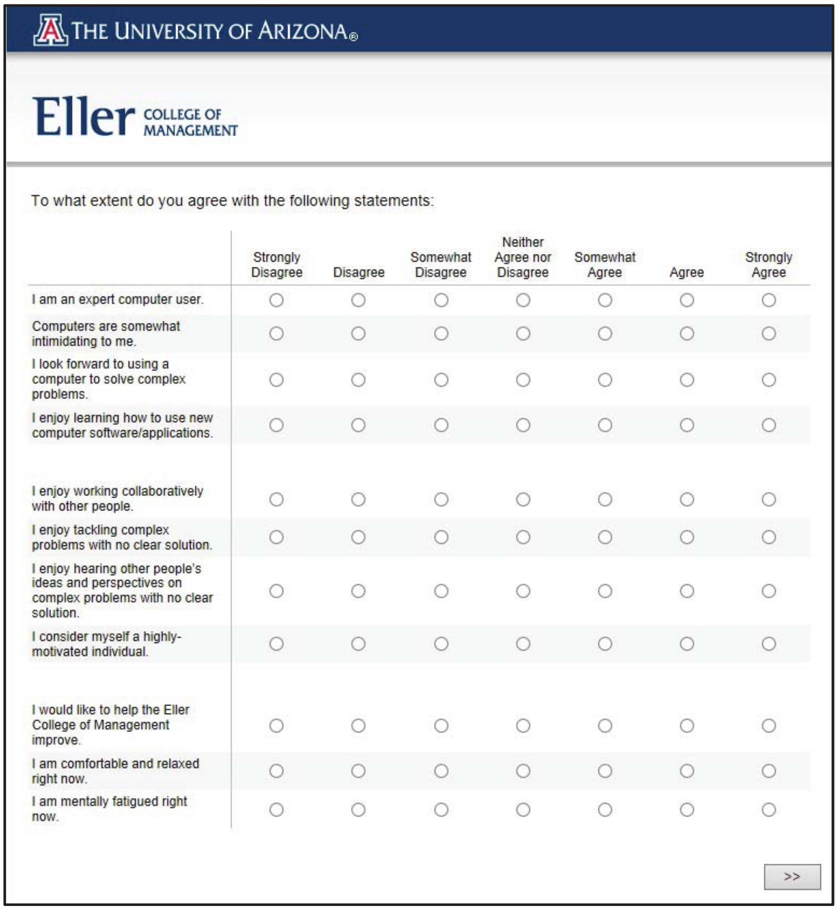
What is your age (in years)?

What is your gender?

○ Male

○ Female

>>



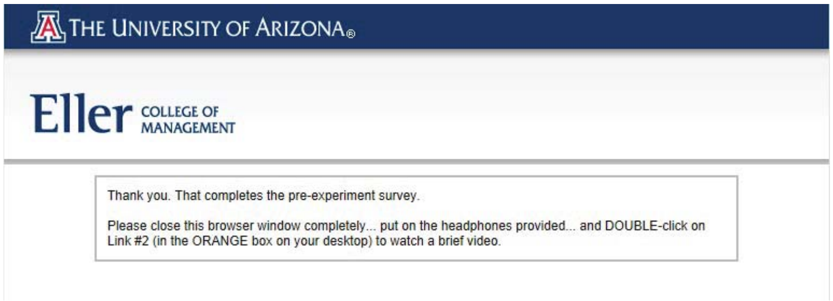
THE UNIVERSITY OF ARIZONA®

Eller COLLEGE OF MANAGEMENT

To what extent do you agree with the following statements:

	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
I am an expert computer user.	○	○	○	○	○	○	○
Computers are somewhat intimidating to me.	○	○	○	○	○	○	○
I look forward to using a computer to solve complex problems.	○	○	○	○	○	○	○
I enjoy learning how to use new computer software/applications.	○	○	○	○	○	○	○
I enjoy working collaboratively with other people.	○	○	○	○	○	○	○
I enjoy tackling complex problems with no clear solution.	○	○	○	○	○	○	○
I enjoy hearing other people's ideas and perspectives on complex problems with no clear solution.	○	○	○	○	○	○	○
I consider myself a highly-motivated individual.	○	○	○	○	○	○	○
I would like to help the Eller College of Management improve.	○	○	○	○	○	○	○
I am comfortable and relaxed right now.	○	○	○	○	○	○	○
I am mentally fatigued right now.	○	○	○	○	○	○	○

>>

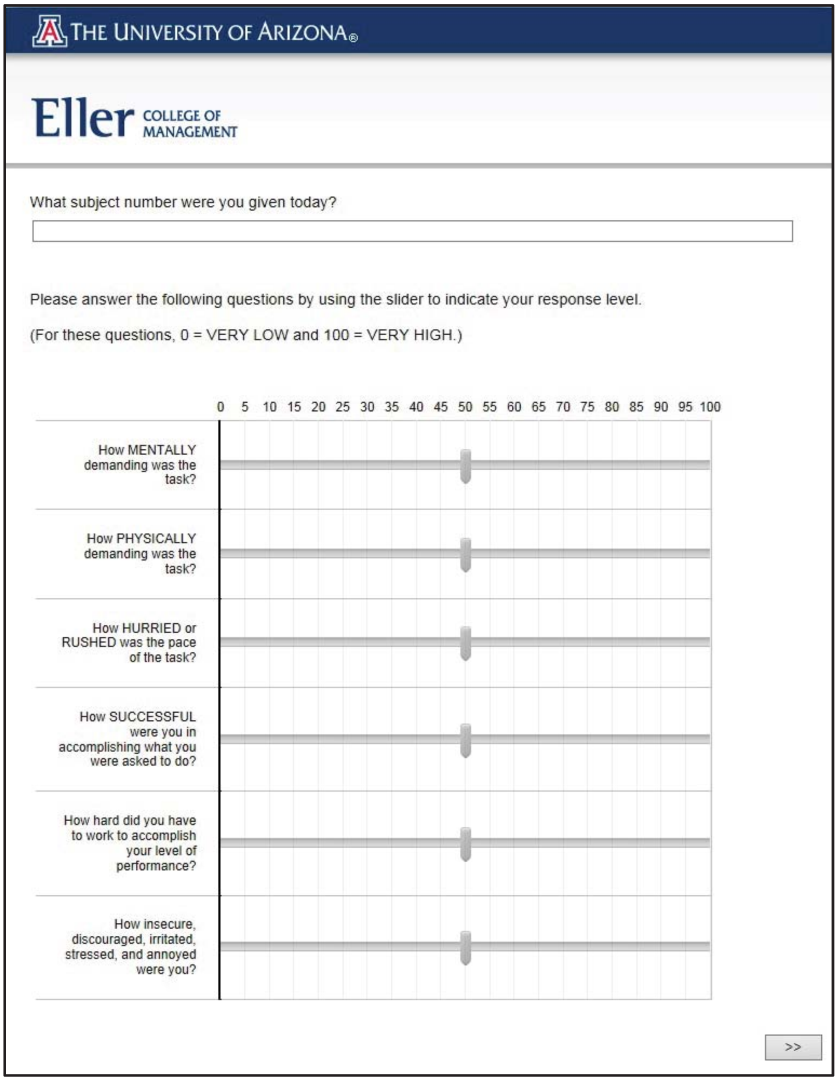


Thank you. That completes the pre-experiment survey.

Please close this browser window completely... put on the headphones provided... and DOUBLE-click on Link #2 (in the ORANGE box on your desktop) to watch a brief video.

APPENDIX F – POST-EXPERIMENT SURVEY QUESTIONS

The following screenshots show the online post-experiment survey each participant was required to complete. Each image represents the successive screens displayed after a participant's prior responses were validated.



What subject number were you given today?

Please answer the following questions by using the slider to indicate your response level.

(For these questions, 0 = VERY LOW and 100 = VERY HIGH.)

How MENTALLY demanding was the task?

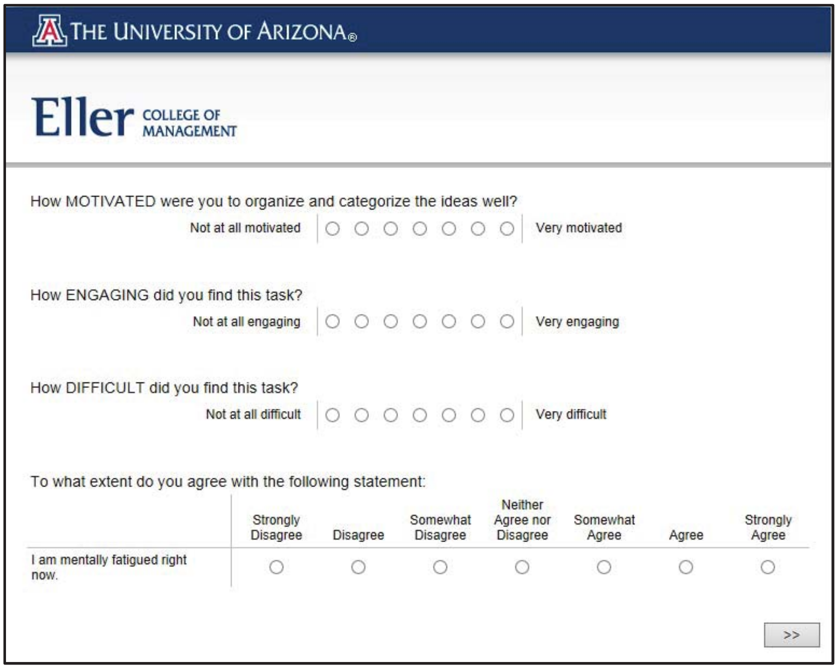
How PHYSICALLY demanding was the task?

How HURRIED or RUSHED was the pace of the task?

How SUCCESSFUL were you in accomplishing what you were asked to do?

How hard did you have to work to accomplish your level of performance?

How insecure, discouraged, irritated, stressed, and annoyed were you?



How MOTIVATED were you to organize and categorize the ideas well?

Not at all motivated ○ ○ ○ ○ ○ ○ ○ Very motivated

How ENGAGING did you find this task?

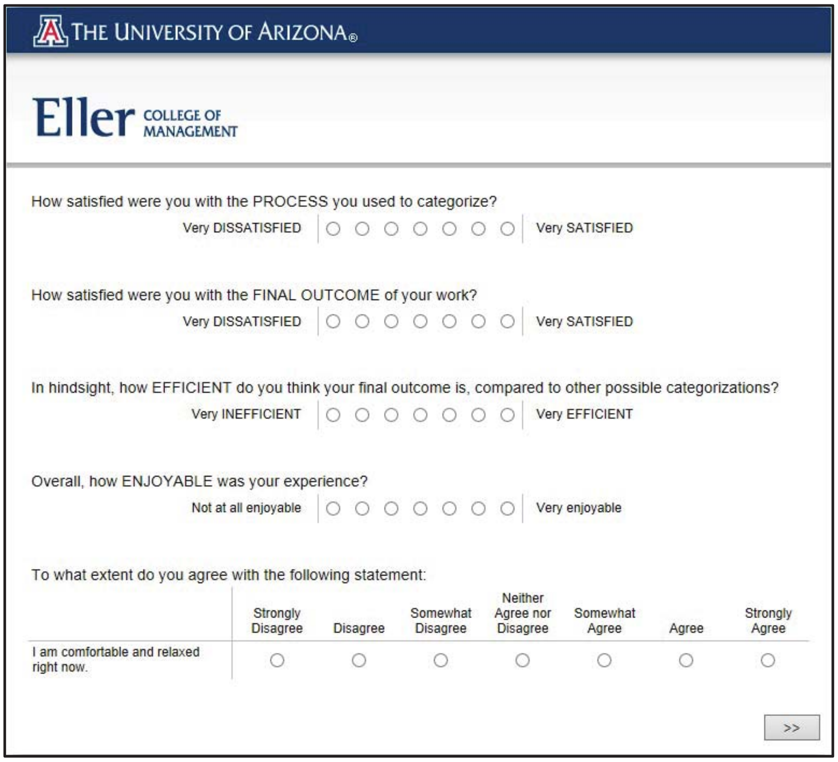
Not at all engaging ○ ○ ○ ○ ○ ○ ○ Very engaging

How DIFFICULT did you find this task?

Not at all difficult ○ ○ ○ ○ ○ ○ ○ Very difficult

To what extent do you agree with the following statement:

	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
I am mentally fatigued right now.	○	○	○	○	○	○	○



How satisfied were you with the PROCESS you used to categorize?

Very DISSATISFIED ○ ○ ○ ○ ○ ○ ○ Very SATISFIED

How satisfied were you with the FINAL OUTCOME of your work?

Very DISSATISFIED ○ ○ ○ ○ ○ ○ ○ Very SATISFIED

In hindsight, how EFFICIENT do you think your final outcome is, compared to other possible categorizations?

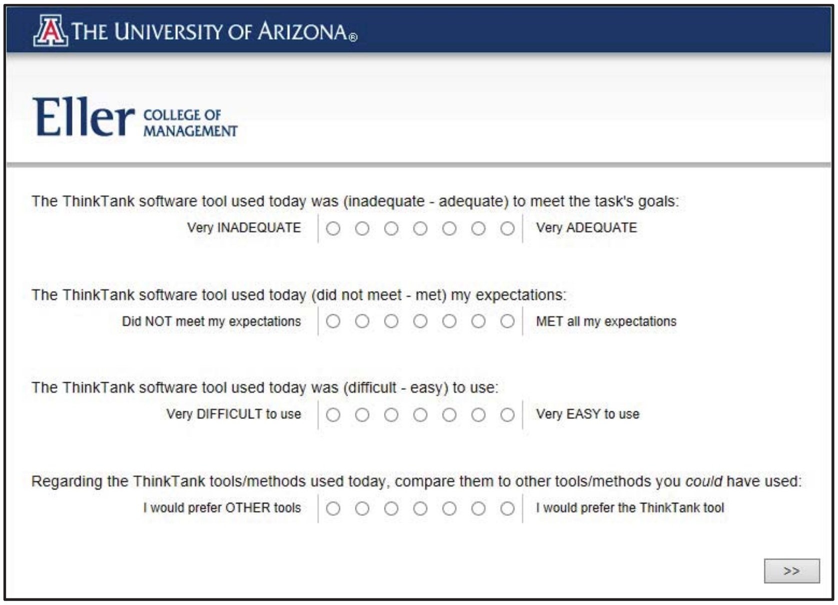
Very INEFFICIENT ○ ○ ○ ○ ○ ○ ○ Very EFFICIENT

Overall, how ENJOYABLE was your experience?

Not at all enjoyable ○ ○ ○ ○ ○ ○ ○ Very enjoyable

To what extent do you agree with the following statement:

	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
I am comfortable and relaxed right now.	○	○	○	○	○	○	○



THE UNIVERSITY OF ARIZONA®

Eller COLLEGE OF MANAGEMENT

The ThinkTank software tool used today was (inadequate - adequate) to meet the task's goals:

Very INADEQUATE ○ ○ ○ ○ ○ ○ ○ Very ADEQUATE

The ThinkTank software tool used today (did not meet - met) my expectations:

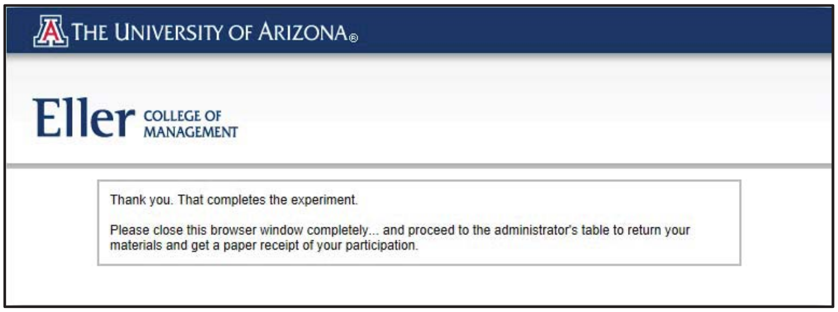
Did NOT meet my expectations ○ ○ ○ ○ ○ ○ ○ MET all my expectations

The ThinkTank software tool used today was (difficult - easy) to use:

Very DIFFICULT to use ○ ○ ○ ○ ○ ○ ○ Very EASY to use

Regarding the ThinkTank tools/methods used today, compare them to other tools/methods you *could* have used:

I would prefer OTHER tools ○ ○ ○ ○ ○ ○ ○ I would prefer the ThinkTank tool



THE UNIVERSITY OF ARIZONA®

Eller COLLEGE OF MANAGEMENT

Thank you. That completes the experiment.

Please close this browser window completely... and proceed to the administrator's table to return your materials and get a paper receipt of your participation.

APPENDIX G – TRANSCRIPT OF INSTRUCTIONAL VIDEO

Thank you for deciding to participate in this “Collaborative Sorting Experiment.”

Remember, your participation is ENTIRELY voluntary… you may quit at any time, without any adverse effects. But we hope that you will take the time to complete the tasks we have prepared for you today.

Now, I’m going to give you a “walk-through” of the GroupSystems’ ThinkTank software that you’ll be using for your next task. You might be interested to know that this software was originally developed here at The University of Arizona, and your efforts today will help us improve how groups can use software like this in the future.

To access the ThinkTank software, all you will need to do is to click on “Link #3” on your desktop, a browser will open and take you to the correct site. You’ll see a login screen that looks like this:

To log in to the activity prepared especially for you today, simply look at the paper that was given to you when you walked in… and enter the information for your session.

If you entered everything in correctly, you’ll see a screen that looks like this:

First off, we need you to record your START time… and all you have to do is to type the word START in the white box at the bottom of this screen. When you hit ENTER… it will appear near the top, along with a time stamp that shows when you started.

Once you’ve done that… DOUBLE-click on the item labeled “SORTING ACTIVITY” in the agenda column. You should now see a screen that looks like this:

Please note that the ideas shown in this demo are NOT what YOU will see.

Your task is to sort the “Original List” of ideas you will be given… like those shown on the far right.

But today we want you to create CATEGORIES that will allow you to group similar IDEAS together.

To create a category, simply click on the white space at the bottom of the CATEGORIES column… and then type in a name for your new category. When you hit ENTER, your category will appear in the list above.

Once you have created a category, you can drag the IDEAS into it… just click-and-hold on the idea… and drag it over to the category you want to put it in… Notice how the category name I'm moving my items to is turning green, to show me where I'm moving each idea to.

Continue creating categories and moving ideas into them until you have categorized ALL of the ideas in the ORIGINAL LIST.

Feel free to create as many categories as you need to effectively sort all these items… what is the best way to "make sense" of all the ideas that you were given?

At any point, if you want to review what you have moved to a particular category… just click on the category name. The category name will be highlighted in blue, and the ideas contained in it will appear on the far right.

If you want to CHANGE or EDIT a category name, just click on the category… then click on EDIT (located in the grey bar, just below the red header)… and "Modify Text"… a pop-up window will appear… and you can type whatever changes you want in there.

Most people find this software to be pretty easy to use… but in case you get stuck, or can't remember what was covered in this video, a transcript of the important points is included under the "Instructions" bar (located at the far left). Feel free to refer to it if you need to… or just raise your hands and ask for help.

But back to your task… Your goal here is to devise the most effective way to sort and organize the set of ideas that you are given… and help the other group make sense of the ideas they had.

When you are done… and your ORIGINAL ITEMS category is empty… DOUBLE-click on the FINISH activity in the AGENDA column (again, on the far left of your screen). Now, all you have to do is type the word "FINISHED" there… hit ENTER… and your finish time will be recorded.

After that, simply click on the word "logout" (in the red header… at the top-right of the screen)… and then close your browser.

When you open YOUR ThinkTank session, you are going to see ideas that were generated by OTHER students during an ACTUAL group meeting. And we want YOU to try to “make sense” of the ideas they typed in. Here is the scenario that they were given:

You are a part of a group that is attempting to determine the most effective actions to improve the University of Arizona's Eller College (formerly known as “BPA”). Let's brainstorm some ideas to help the college improve its effectiveness, rankings, and prestige. HOWEVER, keep in mind that the college has a limited BUDGET (that has been declining for years) and it has limited CAPACITY to add personnel/resources, so please be realistic and practical.

What you will see in your ThinkTank session are some of the results of their brainstorming. In this experiment, we want you to SORT the brainstorming feedback that they generated. Unfortunately, you can’t ADD ideas (or delete them, even)… we just want you to SORT what you are given. Be sure that EVERYTHING in the “Original List!” There should be NO items left in that category when you are done!

The last task in the experiment is to take the last survey… Link #4… which should only take a few moments to complete.

Thanks again for your participation! We sincerely appreciate your help!

You can take off your headphones now… you won’t need them again… close the video player… and click on Link #3 to begin.

APPENDIX H – COMPLETE LIST OF ITEMS TO BE SORTED

Below is the ENTIRE list of items that was utilized in the experiment. Depending upon the randomly-assigned treatment/condition, a participant will see either:

- All 110 items;
- A randomly-selected subset of 55 items; or
- A randomly-selected subset of 37 items.

ITEMS TO SORT:

1. Increase the admission standards
2. let teaching assistants only teach low level classes
3. hire fewer TA's--their lack of experience results in a poor learning experience for students
4. the bpa needs to realize that, especially in MIS, the teachers can make a lot more getting a job that uses their degree and expertise, so they need to offer higher salary or better benefit so they'll stay
5. If we want to keep our ranking in the MIS department high, they need to continue to cut the size of the classrooms so that students have teachers that actually know their names.
6. I think that we should have computers in the programming classrooms. This would improve MIS 121 and 301 which both sucked, in my opinion.
7. I like Dr. XYZ I just think she graded the first essay way too hard. Another example of someone else grading her stuff, though. I think teachers should grade their own stuff! Students need the feedback.
8. That is a good idea. If they had more people that taught fewer classes each, the teachers wouldn't get as burned out and might still be enthusiastic.
9. Yes, also true. Most teachers of big classes are good at that because they have been doing it for a while. But when a teacher that is used to 30 person classes now has a 60 person class, the quality of their teaching decreases because they don't know how to deal. Maybe UA should teach them.
10. If we have to do this for faculty research, maybe we shouldn't increase the level of faculty research.
11. I think that the University needs to increase student tuition while increasing faculty pay. I would have gladly paid more tuition to avoid having teachers like Dr. Thatcher.
12. less students * higher tuition = better learning experience without a loss of revenue
13. That's another thing. To teach high school, you need to have a degree in Education. However, at a University, a prof never needs to take even one teaching class. I would love for my profs to get some teaching education to maybe be able to make class more interesting and exciting. Instead, we get shoved into a room performing experiments like lab rats.
14. add more computer labs for BPA
15. Make graduating requirements stronger in order to ensure students have basic knowledge.
16. Offer more sections of classes to reduce overcrowding
17. Find equipment donors who can receive a tax write off for donating necessary equipment.
18. I haven't had Dr. XYZ but he appears to be very involved with the 441 experience. One solution for improving classroom quality is to attract instuctors who want to teach and enjoy the field.
19. Faculty should be given time to apply for contract funding. This funding could be used to supplement overcrowded classrooms.

20. Tucson is a nice retirement city. The university should recruit older employees who may be considering retirement and who would enjoying teaching while semi-retired. They could bring their experience to the classroom.
21. The amount of pay is the basis for a good quality of faculty, but the state of Arizona is not willing to give salaries comparable to what people can get outside of the university. This needs to be communicated to our law makers.
22. There should be more mentoring programs for students who feel they are not getting the attention they need in the oversized classroom.
23. Students need more experience with high tech equipment in order to be qualified for employment.
24. The U of A needs to hire more faculty somehow
25. Raise the money without raising tuition
26. The level of education will increase if there are better professors
27. I agree. The U of A needs to focus more on the necessary skills than silly classes such as 471
28. I would even appreciate feedback from the TA. I think that teachers are spending too much effort in being lazy.
29. It should be a requirement that the grader at least sit in the class every once in a while.
30. That basically means spending a lot of time to find ways to cut corners.
31. And I would gladly pay higher tuition to not have TA's teach my classes.
32. charge higher tuition
33. I agree with the Tucson/retirement city/older employees idea. I learn so much more from long-time employees than from life-long teachers.
34. I definitely agree that admission standards should be increased to keep out many of the potential students.
35. I know a lot of teacher pay students to grade their papers. I think feedback from the teacher is vital. they need smaller classes!
36. I know. It bugs me that a student or grader that isn't even in the class to hear the lectures grades the papers. I don't think they are qualified.
37. I think mentoring programs would be great. It would add value to the student's experience as well as the teachers. They would have a more personal experience with the students and might feel like there is more value to their jobs.
38. If salaries are not raised for good faculty, other benefits should be given to compensate for the low salary.
39. If the grader was in class, they would be more reluctant to tear into your paper without justification
40. If we were taught more things that applied to the real world, we would be more intrested in learning them.
41. In addition to basic knowledge received during the educational process, employers are also looking for practical skills which would allow employees to be productive immediately with out additional training.
42. Lobby congress for higher faculty salaries. This would draw in better professors.
43. Make the high level classes have fewer students
44. Many students are ill prepared after graduation, but usually it is because they don't care about the experience of learning. There should be an graduation exam before completion to determine if students know about the basics of their degree.
45. Out with TA's, in with retired business people, in with computers in the classrooms, in with higher tuition for higher salaries, out with bad teachers and bad classes.
46. pay faculty more
47. Pay the good professors more and let the bad ones quit
48. Require higher entry requirements
49. the in-state admission standards are low and we get all of the students that couldn't get accepted to ASU of even NAU! we should have higher standards than that
50. The U of A needs to realize who are the good profs and who are the bad ones. I would not mind a large class if I had someone that was interesting teaching me.
51. UA could hire more faculty if they offered them more money, and they could offer more money if admission standards were higher so tuition was higher.
52. We need to gain revenue to pay the teachers more
53. What does this have to do with 441? However, I would rather learn from retired business people than someone who has only taken teaching classes.
54. Yes, we need stuff that applies to the business world so that we can be prepared.

55. yes, we should increase admission requirements because UA's are the lowest in the state
56. #8299 is cool.
57. Amen my brother! It is these people that we need in the classroom. Not people who have learned everything from a book.
58. At this point, I think Universities are a lost cause. I learn better on my own. The only reason I still want to graduate is because it looks good on a resume. Otherwise, it was a waste of 5 years.
59. Being a professor is a very high stress job. Unfortunately, most students only think that teachers teach. However, the truth is that teachers are expected to perform an immense amount of research at the same time. Being a college prof is nothing like being a high school teacher.
60. but enjoying to teach only lasts so far. I would enjoy teaching but not under the current circumstances.
61. By the way, how do you "spend too much effort on being lazy?" That sounds like a contradiction to me...
62. Down with 471!
63. Either do I.
64. Employees expect that they are going to need to train students out of college. It would just be nice if the focus was more on skills than on all the touchy feely stuff in 471.
65. Especially if the grader is going to give everyone low grades. A few comments from the grader-from-hell would be nice.
66. Higher standards would only serve to bring in less tuition
67. How can we offer more sections when teachers are quitting?
68. how do you propose we raise money without raising tuition? please clarify.
69. How long will they want to keep this cushy job when they start having to grade 75 tests per class? Doesn't sound too cushy to me?
70. I believe that giving business students more in-depth computer training would benefit them much greater than requiring them to take courses such as ancient Ugandan history or "Transvestites in Modern Society."
71. Maybe self-paced courses with TA's available during certain hours for assistance are necessary. CBT is a wonderful educational tool and should be used to a fuller extent.
72. Tenure is an antiquated concept. The TCE (Teacher course evaluation) should bear much more weight in deciding the future of that professor. To alleviate any biases the teacher may have, they should be distributed and collected by an arbitrary 3rd party
73. That could work, but set a specified number of hours that need to be taught. BUt what about the technology? Most people don't have high speed internet connections at home to access streaming video lectures. THis would also be expensive to implement.
74. class sizes would have to be small then, because physical size constraints would make labs unavailable at convenient time for students. You would also need lab monitors during hours of operation.
75. Maybe we should have some sort of lottery based system to award benefits, but also performance based benefits.
76. It is much easier to learn in a small class. This is a proven fact. Making large classes like in Harvill 150 would reduce the quality of students.
77. Maybe an MCSE course would be more helpful.
78. If students are not prepared to contribute in the work force after graduation, the value of a U of A degree does not mean much, and employers realize this.
79. I'm out of good solutions. Sorry guys but count me out of any non-off-the-topic discussions.
80. I'm tired, I have a headache. . . I hated this assignment in 471 and I hate it now.
81. It is not so much the overcrowding that is bad, it is the poor quality professors
82. It's simple suppy and demand - the University will never keep instructors when they could make so much more elswhere.
83. Many teachers teach only a few classes a week. It is true that they then have to grade papers but I cannot think of a less stressful job.
84. Me too. Only ten minutes left. Hey--at least we don't have to answer the same question 100 times!
85. Nah, you have no idea how stressful a faculty position is. After seeing it in my family's life (you've met my mom, right Randy?) I would take a real-world job any day. Universities are soooo political its pathetic.
86. Okay, makes sense. Maybe I do that too :)
87. One benefit might me showing up to class for an hour and then having someone else grade your papers.
88. Or it might not.

89. have undergraduates use the graduate labs in the bpa
90. cancel some General Education classes and replace them with related to major coursess that will raise the technical skills of the BPA undergrad.
91. many grad students are not good teachers but still they get hierd to teach. There are many qualified undrgraduates that can be better teachers than other TA.
92. grant greater range of benefits to the professors family.
93. turn to major computer companies for financing computer labs should I remind you the name Pepsi???
94. there are many international students that come to study here and they contribue substantially to the academic ranking of the university but they dont get any assistant in finding internships by sending students to internships we could raise the level of the school in the eyes of companies that will want to sponser us
95. combain small sections into mass lectures. unite small classrooms that hold 60 students with ones that hold 120. by cutting sections we could use the money saved to hire the best professors avilable. It doesnt matter if a classroom is large. if you want to study for an - A - you will study !!!
96. Save money by cancelling the Friday evening and weekend classes
97. Have students help with research. This will help the professors and give the students an opportunity to gain experience
98. they could also make a semester-long internship program mandatory. The university would have to help students find these interneships
99. I think that the professors will stay in an environment where they are free to complete the research they desire. Maybe we could give them every 4th semester off purely for research. I think it's important that the students get taught by the professors as much as possible, so there should be a great emphasis on teaching during the other semesters
100. I think that the professors deserve regular pay increases (like any other profession), but I do not think that the university should focus on giving them fringe benefits. I don't want to bribe them to do their jobs
101. the TA's should be required to take a teaching course before being allowed in the classroom. They should also be evaluated and paid more money. Perhaps this would entice them to take their teaching responsibilities more seriously
102. Perhaps each professor should be required to write one grant proposal per year (or semester)
103. I don't like the idea of adding fringe benefits to get a professor to stay. If the college offers the professor the working environment that she/he wants, then they will stay without the extras
104. in order to let the students graduate on time, some of the general education requirements must be cut in order to fulfill the co-op/internship requirement
105. i agree that graduate students should not necessarily be forced to teach, but that means finding a lot more money to pay professors to teach all the classes. With the present money shortages, I'm afraid that isn't realistic
106. The faculty should not be required to participate on committees. This should be voluntary
107. I agree, benefits should be distributed based on performance. I do think that professors that have been here longer should get a few added benefits.
108. I competely agree that tenure should be done away with, but if we do, we will lost many of our senior professors and have a hard time getting new ones - unless all universities throw out tenure
109. A high-speed lab could be set up. I think the expense would be OK in the long run
110. many students like to have smaller classes so they get more individual attention. I agree that if an A is their only objective that it can be done in a larger class, but that is not the only thing many students are after (hopefully)

APPENDIX I – ANOVA CONTRASTS OUTPUT (RAW TLX)

Below are the complete results from the initial failed ANOVA analysis of the averaged Raw TLX scores, by treatment condition:

Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
A	49	41.1	11.96	1.71	37.7	44.5	13.3	63.3
B	49	38.8	9.60	1.37	36.0	41.5	23.8	74.2
C	75	39.0	7.49	0.87	37.3	40.8	22.5	63.3
Total	173	39.5	9.53	0.72	38.1	41.0	13.3	74.2

Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
6.999	2	170	.001

ANOVA

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		169.4	2	84.7	.932	.396
	Linear Term	Unweighted	126.3	1	126.3	1.391	.240
		Weighted	108.7	1	108.7	1.196	.276
		Deviation	60.7	1	60.7	.668	.415
Within Groups			15440.7	170	90.8		
Total			15610.0	172			

Robust Tests of Equality of Means

	Statistic[a]	df1	df2	Sig.
Welch	.679	2	91.9	.510
Brown-Forsythe	.845	2	124.1	.432

a. Asymptotically F distributed.

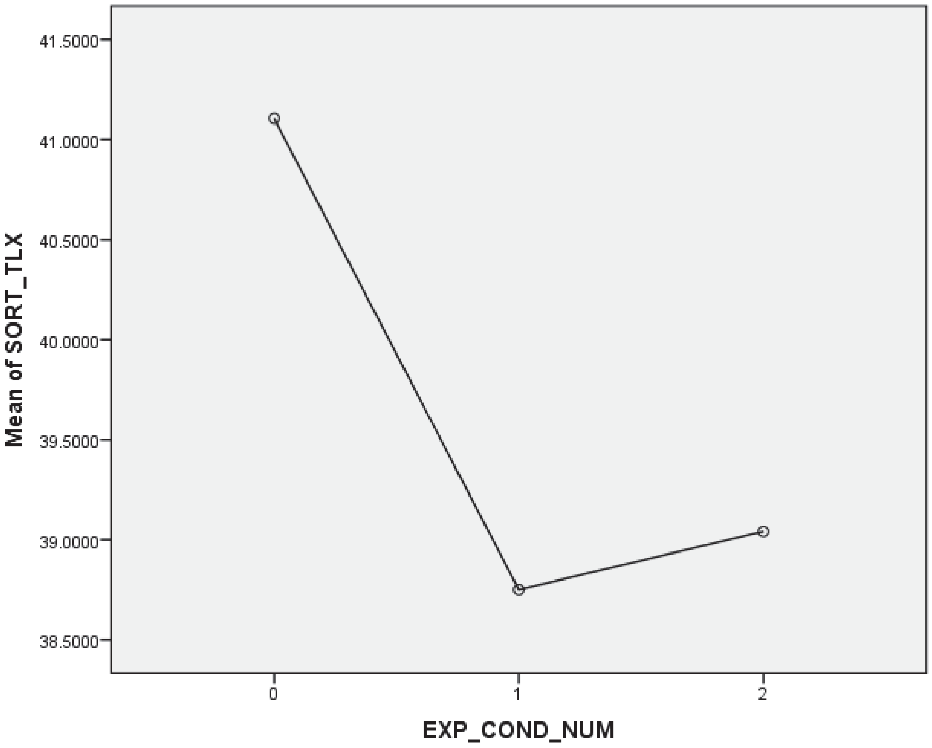
Contrast Coefficients

Contrast	EXP_COND_NUM		
	0	1	2
1	-1	0	1
2	0	-1	1
3	-1	1	0

Contrast Tests

Contrast			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
SORT_TLX	Assume equal variances	1	-2.065	1.751	-1.179	170	.240
		2	0.291	1.751	.166	170	.868
		3	-2.355	1.925	-1.223	170	.223
	Does not assume equal variances	1	-2.065	1.915	-1.078	72.7	.285
		2	0.291	1.621	.179	85.1	.858
		3	-2.355	2.191	-1.075	91.7	.285

Means Plots



POST-HOC - Multiple Comparisons

(I) EXP_COND_NUM			Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tamhane	0	1	2.36	2.19	.635	-2.97	7.68
		2	2.06	1.91	.634	-2.62	6.74
	1	0	-2.36	2.19	.635	-7.68	2.97
		2	-0.29	1.62	.997	-4.24	3.66
	2	0	-2.06	1.91	.634	-6.74	2.62
		1	0.29	1.62	.997	-3.66	4.24
Dunnett T3	0	1	2.36	2.19	.632	-2.97	7.68
		2	2.06	1.91	.630	-2.61	6.74
	1	0	-2.36	2.19	.632	-7.68	2.97
		2	-0.29	1.62	.997	-4.24	3.66
	2	0	-2.06	1.91	.630	-6.74	2.61
		1	0.29	1.62	.997	-3.66	4.24
Games-Howell	0	1	2.36	2.19	.532	-2.86	7.57
		2	2.06	1.91	.531	-2.52	6.65
	1	0	-2.36	2.19	.532	-7.57	2.86
		2	-0.29	1.62	.982	-4.16	3.58
	2	0	-2.06	1.91	.531	-6.65	2.52
		1	0.29	1.62	.982	-3.58	4.16
Dunnett C	0	1	2.36	2.19		-2.94	7.65
		2	2.06	1.91		-2.56	6.69
	1	0	-2.36	2.19		-7.65	2.94
		2	-0.29	1.62		-4.20	3.62
	2	0	-2.06	1.91		-6.69	2.56
		1	0.29	1.62		-3.62	4.20

APPENDIX J – ANOVA CONTRASTS OUTPUT (ADJ. TLX)

Below are the complete results from the second failed ANOVA analysis of the Adjusted TLX scores, by treatment condition:

Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		MIN	MAX
					Lower Bound	Upper Bound		
A	49	52.6	20.4	2.92	46.7	58.4	12.5	95.0
B	96	47.0	22.5	2.30	42.4	51.6	5.0	87.5
C	201	47.7	21.4	1.51	44.8	50.7	0.0	95.0
Total	346	48.2	21.6	1.16	45.9	50.5	0.0	95.0

Test of Homogeneity of Variances

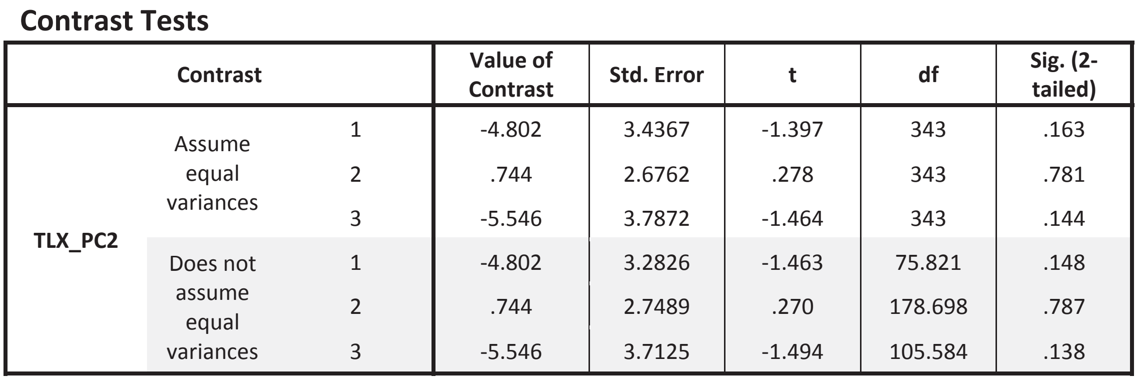
Levene Statistic	df1	df2	Sig.
.768	2	343	.465

ANOVA

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		1105.432	2	552.716	1.188	.306
	Linear Term	Unweighted	908.540	1	908.540	1.953	.163
		Weighted	515.483	1	515.483	1.108	.293
		Deviation	589.949	1	589.949	1.268	.261
Within Groups			159601.432	343	465.310		
Total			160706.864	345			

Contrast Coefficients

Contrast	EXP_COND_NUM		
	0	**1**	**2**
1	-1	0	1
2	0	-1	1
3	-1	1	0



Contrast Tests

Contrast			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
TLX_PC2	Assume equal variances	1	-4.802	3.4367	-1.397	343	.163
		2	.744	2.6762	.278	343	.781
		3	-5.546	3.7872	-1.464	343	.144
	Does not assume equal variances	1	-4.802	3.2826	-1.463	75.821	.148
		2	.744	2.7489	.270	178.698	.787
		3	-5.546	3.7125	-1.494	105.584	.138

REFERENCES

Anson, R., Bostrom, R., & Wynne, B. (1995). An experiment assessing group support system and facilitator effects on meeting outcomes. *Management Science*, *41*(2), 189-208.

Briggs, R.O., Crews, J.M., & Mittleman, D.D. (1998). Facilitating asynchronous distributed GSS meetings: Eight steps to success. *interaction, 11*(4), 2.

Briggs, R.O., de Vreede, G.J., Nunamaker Jr, J.F., & Tobey, D. (2001, January). ThinkLets: achieving predictable, repeatable patterns of group interaction with group support systems (GSS). In *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on* (pp. 9-pp). IEEE.

Briggs, R. O., de Vreede, G. J. & Nunamaker, J. F., Jr. (2003). Collaboration engineering with thinkLets to pursue sustained success with group support systems. *Journal of Management Information Systems 19*(4): 31-64.

Briggs, R. O., & Nunamaker Jr, J. F. (2006). Monograph of the HICSS-39 Symposium on Case and Field Studies of Collaboration.

Chen, H., Hsu, P., Orwig, R., Hoopes, L., & Nunamaker, J. F., Jr. . (1994). Automatic concept classification of text from electronic meetings. *Communications of the ACM, 37*(10), 56-73.

Cottrell, N.B. (1972). Social Facilitation. In C. McClintock (ed.), *Experimental Social Psychology* (pp. 185–236). New York: Holt, Rinehart & Winston.

de Vreede, G. J., Boonstra, J., & Niederman, F. (2002, January). What is effective GSS facilitation? A qualitative inquiry into participants' perceptions. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on* (pp. 616-627). IEEE.

de Vreede, G. J., Vogel, D., Kolfschoten, G., & Wien, J. (2003, January). Fifteen years of GSS in the field: A comparison across time and national boundaries. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference* on (pp. 9-pp). IEEE.

DeLeeuw, K.E., & Mayer, R.E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology 100* (1): 223-234.

Dennis, A., Haley, B., and Vandenberg, R. (1996). A meta-analysis of effectiveness, efficiency, and participant satisfaction in group support systems research. *ICIS 1996 Proceedings.* Paper 20.

Dennis, A. R., George, J. F., Jessup, L. M., Nunamaker, J. F., Jr., & Vogel, D. R. (1988). Information technology to support electronic meetings. *MIS Quarterly, 12*(4), 591-624.

Dennis, A. R., J. S. Valacich, et al. (1990). An experimental investigation of the effects of group size in an electronic meeting environment. *IEEE Transactions on Systems, Man and Cybernetics 20*(5): 1049-1057.

Gallupe, R. B., A. R. Dennis, et al. (1992). Electronic Brainstorming and Group Size. *Academy of Management Journal 35*(2): 350-369.

Grohowski, R., C. McGoff, et al. (1990). Implementing Electronic Meeting Systems at IBM: Lessons Learned and Success Factors. *MIS Quarterly 14*(4): 369-383.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam: North Holland.

Helquist, J. (2007). Participant-driven Group Support Systems: An approach to distributed, asynchronous collaborative systems.

Helquist, J. H., Deokar, A., Meservy, T., & Kruse, J. (2011). Dynamic collaboration: participant-driven agile processes for complex tasks. *ACM SIGMIS Database*, *42*(2), 95-115.

Helquist, J. H., Kruse, J., & Adkins, M. (2006a). Developing large scale participant-driven group support systems: An approach to facilitating large groups. *20 Years of Collaboration in the Military*, 11.

Helquist, J., Kruse, J., & Adkins, M. (2006b). Group support systems for very large groups: A peer review process to filter brainstorming input.

Helquist, J. H., Kruse, J., & Adkins, M. (2008, January). Participant-driven collaborative convergence. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual* (pp. 20-20). IEEE.

Hilmer, K. M. & A. R. Dennis (2000). Stimulating thinking: Cultivating better decisions with groupware through categorization. *Journal of Management Information Systems 17*(3): 93-114.

Karypis, G., & Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing, 20*(1), 359-392.

Lowry, P. B., Roberts, T. L., Romano, N. C., Cheney, P. D., & Hightower, R. T. (2006). The Impact of Group Size and Social Presence on Small-Group Communication Does Computer-Mediated Communication Make a Difference?. *Small Group Research*, *37*(6), 631-661.

Nunamaker, J. F., A. R. Dennis, et al. (1991). Electronic meeting systems to support group work. *Communications of the ACM 34*(7): 40-61.

Nunamaker, J. F., Jr., R. O. Briggs, et al. (1996). Lessons from a dozen years of group support systems research: A discussion of lab and field. *Journal of Management Information Systems 13*(3): 163.

Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist 38* (1): 63-71.

Paas, F., & Van Merriënboer, J. (1993). The efficiency of instructional conditions: An approach to combine mental-effort and performance measures. *Human Factors 35* (4): 737-743.

Roberts, T. L., Lowry, P. B., & Sweeney, P. D. (2006). An evaluation of the impact of social presence through Group size and the use of collaborative software on Group member. *Professional Communication, IEEE Transactions on*, *49*(1), 28-43.

Romano Jr, N.C., Nunamaker Jr, J.F., Briggs, R.O., & Mittleman, D. D. (1999). Distributed GSS facilitation and participation: Field action research. In *System Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference* on (pp. 12-pp). IEEE.

Rosenberg, M. J. (1965). When dissonance fails: On eliminating evaluation apprehension from attitude measurement. *Journal of Personality and Social Psychology, 1*(1), 28.

Roussinov, D. G., & Chen, H. (1999). Document clustering for electronic meetings: an experimental comparison of two techniques. *Decision Support Systems, 27*(1), 67-79.

Roussinov, D., & Zhao, J. L. (2003). Automatic discovery of similarity relationships through web mining. *Decision Support Systems*, *35*(1), 149-166.

Strehl, A., & Ghosh, J. (2003). Cluster ensembles---a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research, 3*, 583-617.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science 12* (2): 257-285.

Sweller, J., Van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review 10* (3): 251-296.

Tarmizi, R. A., & Sweller, J. (1988). Guidance during mathematical problem solving. *Journal of Educational Psychology*, 80(4), 424.

Valacich, J. S., Dennis, A. R., & Nunamaker, J. F. (1992). Group size and anonymity effects on computer-mediated idea generation. *Small Group Research*, *23*(1), 49-73.

Voorhies, D. J. & Scandura, J. M. (1977). 'Determination of memory load in information processing', in J. M.Scandura (ed.), *Problem Solving: A Structural/Process Approach*, Academic Press, (New York), pp. 299-316.