Large-Scale Needfinding Methods, Quality Metrics, and Need Prioritization in User-Centered Design

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Cory R. Schaffhausen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Prof. Timothy M. Kowalewski

August, 2015

Acknowledgements

Timothy Kowalewski, PhD was instrumental in guiding the project as my advisor. His dedicated and consistent mentoring from the overall research goals to minutia of academic life and willingness to commit time and funding when needed was immeasurably valuable.

A sincere thanks go to the members of my advising committee for the patience and commitment to seeing this through to be a successful project: William Durfee, PhD, Arthur Erdman, PhD, and Kathleen Harder, PhD.

The Statistical Consulting Service at the University of Minnesota, and in particular Lindsey Dietz and Felipe Acosta, who helped with the analysis of this series of experiments.

Thanks go to Ted Pedersen, PhD at the University of Minnesota Duluth for valuable early suggestions and direction relating to candidate Natural Language Processing algorithms. Lushan Han at UMBC and Mladan Karan at the University of Zagreb for assistance during algorithm testing.

William Durfee, PhD for his valuable input on study design and research goals. And to Rob Sweet, MD and David Hananel for critical assistance and support with the simulation manikin case study.

Dedication

To my parents and their sacrifices that kept me moving towards the best possible education. To my wife, Allyson, who was always supportive of my rollercoaster career plans, and to little Miles and Sylvia who learned that good table manners meant patiently listening to what progress in "science" was done that day. To countless engineers, designers, and educators who demonstrated empathy and instilled an appreciation for the people who use the products we create.

Abstract

This thesis describes front-end, user-centered design methods to generate and prioritize unmet needs of large, diverse groups. Understanding user needs and preferences is increasingly recognized as a critical component of early stage product development. The large-scale needfinding methods in this series of studies attempt to overcome shortcomings with existing methods, particularly in environments with limited user access. The thesis is presented in four main parts, each with differing objectives. Part 1 focuses on need quantity and includes three studies to evaluate three specific types of stimuli to help users describe higher quantities of needs. Part 2 focuses on uniqueness and describes an automated method to effectively process large quantities of content commonly generated in open innovation practices, including the needs-based data produced in Part 1. Part 3 focuses on quality and describes methods to rapidly prioritize a large set of needs to identify a small subset for further consideration. Previous analytic methods have been used for small quantities (often fewer than 75 statements). Part 4 includes a case study relating to a target application area of medical technology.

Study participants in part 1 were trained on need statements and then asked to enter as many need statements and optional background stories as possible. One or more stimulus types were presented, including prompts (a type of thought exercise), shared needs, and shared context images. The topics used were general household areas including cooking, cleaning, and trip planning. In part 2, a series of studies explored automated duplication detection using state-of-the art natural language processing (NLP) algorithms. The Semantic Textual Similarity (STS) algorithms had been specifically developed to compare sentence-length text passages and were used to rate the semantic similarity of pairs of text sentences describing unmet needs. Additional participants

were recruited in part 3 to rate need statements using an online interface and a simplified quality metric appropriate to initially screen and prioritize lists exceeding 500 statements for a single topic or product area. In part 4, the methods described in parts 1 and 2 were adapted for use as an email accessible web application delivered to a group of professionals in the medical education field. The topic for the study was a needs assessment for a next generation of medical simulation manikins.

Across the series of studies, a number of hypotheses relating to need quantity and quality were tested and secondary research questions were explored. The novel methods were demonstrated as effective to rapidly general lists of unmet needs from large groups. A final quantity study collected 1735 needs statements and 1246 stories from 402 individuals in 24 hours. The Part 1 (Quantity) results show that users can articulate a large number of needs unaided, and users consistently increased need quantity after viewing a stimulus. Part 2 (Uniqueness) results identify top modern STS algorithms for needfinding. These predicted similarity with Pearson correlations of up to .85 when trained using need-based training data. Part 3 (Quality) results and individual hypothesis tests provide additional key contributions. Increasing the number of participants contributing needs can increase the quantity of unique needs as well as the number of high-quality needs. Increasing the number of needs contributed per person increases the number of high-quality needs. Increasing levels of self-rated expertise will not significantly increase the number of high-quality needs per person. Needs submitted first are not lower quality than needs submitted after a sustained period of time. Part 4 demonstrates feasibility of applying online needfinding methods to professional users and suggests that these methods can result in a set of overlapping needs compared to focus group data and can also identify unique needs not identified in focus groups.

The results contribute baseline studies to describe a systematic quantity focus as applied to finding needs and demonstrate how users can articulate quality needs given appropriate training and tools. Quality study results provide evidence of a benefit to balancing widespread short user interactions with longer, in-depth interactions. If the objective of a user research effort is to maximize the number of high-quality needs identified, the results in aggregate support the use of multiple approaches including 1) increase the user group size, 2) increase the quantity of needs suggested per person, 3) increase or maintain a diversity in levels of expertise in the user group.

Contents

Acknowledgements i

Dedication ii

Abstract iii

List of Tables xi

List of Figures xiii

1 Introduction 1

1.1 Overview of Multiple Studies 2

1.2 Contributions of This Research 7

2 Literature Review and Background 8

2.0.1 Needfinding and User Requirements 8

2.0.2 Types of Need or Problem Statements 10

2.0.3 Application Example: Heath Care and Medical Devices 11

2.1 Part 1 Background: Collecting User Need Statements 12

2.1.1 Lessons from Ideation for Needfinding 12

2.1.2 Ideation and Needfinding Compared 14

2.1.3 Users Articulating Needs 15

2.1.4 Rationale for Large Quantities of Needs 17

2.1.5 Rationale for Our Stimuli Design 18

2.2 Part 2 Background: Assessing Uniqueness of Need Statements 19

2.2.1 Large Data Sets in Ideation 19
2.2.2 Natural Language Processing Background 20
2.2.3 SemEval Algorithm Competition Background 21
2.2.4 Algorithm 1 Background: TakeLab-simple 22
2.2.5 Algorithm 2 Background: UMBC-PairingWords 22
2.3 Part 3 Background: Assessing Quality of Need Statements Submitted by Users 22
2.3.1 Kano Model 23
2.3.2 New Product Design and Development Texts 23
2.3.3 Importance and Satisfaction of Outcomes 24
2.3.4 Differences from Previous Work 25

3 Part 1: Collecting Large Quantities of User Need Statements 27
3.1 Needfinding Topics 28
3.2 Part 1 Methods Overview (Quantity) 28
3.2.1 Quantity Study 1 Methods 33
3.2.2 Quantity Study 2 Methods 34
3.2.3 Quantity Study 3 Methods 35
3.2.4 Control Stimulus 36
3.2.5 Stimulus 1: Narrative prompts 36
3.2.6 Stimulus 2: Shared Needs and Stories 37
3.2.7 Stimulus 3: Shared Images 37
3.2.8 Stimulus instructions 37
3.3 Part 1 Results (Quantity) 40
3.3.1 Quantity Study 1 Results 41
3.3.2 Quantity Study 2 Results 42
3.3.3 Quantity Study 3 Results 42
3.3.4 Aggregated Observations for All Three Studies 44
3.3.5 Aggregated Observations for All Three Studies: Unpublished . . 46
3.4 Part 1 Discussion (Quantity) 51
3.4.1 Fast, Large-Scale Need Collection is Feasible 51
3.4.2 Collecting Needs Does Not Require In-Depth Research 51

3.4.3 Effects of Incentives and Stimuli 51
3.4.4 User Expertise and Experience are Not Interchangeable 52
3.4.5 Collecting data on Amazon Mechanical Turk 53
3.4.6 High Volume of Needs without Stimulus 54
3.4.7 User interface, quantities, and rates 54
3.4.8 Limitations and Future Work 55

4 Part 2: Assessing Uniqueness of Need Statements 57
4.1 Part 2 Methods Overview (Uniqueness) 58
4.1.1 Need Statement Preprocessing 59
4.1.2 Need Statement Training Sets 59
4.1.3 Similarity Cutoff Scores 60
4.1.4 Algorithm 1 Analysis: TakeLab-simple 62
4.1.5 Algorithm 2 Analysis: UMBC-PairingWords 62
4.1.6 Uniqueness Study: Identifying Unique and Redundant Entries . 63
4.1.7 Analysis of Crowd Size Permutations 63
4.2 Part 2 Results (Uniqueness) 66
4.2.1 Performance of Algorithms 66
4.2.2 Uniqueness Study: Summary of Potential Duplicates 67
4.2.3 False Negatives and False Positives 67
4.2.4 Uniqueness Study: Unique Statements and Crowd Size 69
4.3 Part 2 Discussion (Uniqueness) 71
4.3.1 STS Can Detect Duplications and Uniqueness in Needs-Based Open Innovation Data 71
4.3.2 Reduced Resources for Automated Methods 72
4.3.3 Potential for Future Increases in Accuracy 72
4.3.4 Evidence Against Fraud or Malicious Use 73
4.3.5 Limitations and Future Work 73

5 Part 3: Assessing the Quality of Need Statements Submitted by Users 75
5.1 Part 3 Methods Overview (Quality) 76
5.1.1 Need Statement Data 77
5.1.2 Quality Rating Data Collection 77

5.1.3 Need Statement Quality Rating Phases 78
5.1.4 Quality Metric 79
5.1.5 Analysis Methods for Effects of User Characteristics 79
5.1.6 Analysis Methods for Effects of Need Statement Characteristics . 81
5.2 Part 3 Results (Quality) 84
5.2.1 Need Quality Distribution 85
5.2.2 Need Quality for Varying Group Sizes 85
5.2.3 High Quality and High Quantity 87
5.2.4 High Quality and User Expertise 87
5.2.5 Need Rater and Need Submitter Experience 88
5.2.6 Highest-Rated Need Statements 89
5.2.7 Need Quality and Sequence 91
5.2.8 Quality of Duplicate Statements 93
5.2.9 Need Statements With and Without Detailed Stories 97
5.2.10 Quality of Need Statements after Viewing a Stimulus 98
5.2.11 Need Statement Quality and Statement Uniqueness 98
5.3 Part 3 Discussion (Quality) 100
5.3.1 Higher Need Statement Quantity Leads to Higher Quality 100
5.3.2 Expertise Does Not Predict User-Rated Quality 101
5.3.3 High-Volume Quality Rating is Feasible 101
5.3.4 The First Needs to Come to Mind Are Not Lower Quality than Later Needs 102
5.3.5 Algorithmically-Rated Unique Need Statements Are Not Higher Quality than Those with Many Similar Variants 103
5.3.6 Omitting Detailed Context When Rating Need Statement Quality May Not Effect Ratings 104
5.3.7 Each Type of Stimulus May Result in Quality Need Statements . 105
5.3.8 Limitations and Future Work 105

6 Part 4: Medical Simulation Manikin Case Study 107
6.1 Medical Simulation Manikin Background 108
6.2 Part 4: Case Study Methods 110

6.3 Part 4: Case Study Results . . . 113
6.4 Part 4: Case Study Discussion . . . 115
6.4.1 Users Articulate Similar as Well as Unique Needs Compared to Focus Groups . . . 115
6.4.2 Lengthy Instructions Contributed to Low Completion Rate . . . 115

7 Conclusions **117**

References **121**

Appendix A. Part 1: Quantity **132**
A.1 Part 1: Quantity Study 1 Additional Material . . . 133
A.1.1 Amazon Mechanical Turk Interface for Study 1 . . . 133
A.1.2 Zoho Creator App for Study 1 . . . 133
A.1.3 Narrative Prompt Options . . . 140
A.2 Part 1: Quantity Study 2 Additional Material . . . 153
A.2.1 Amazon Mechanical Turk Interface for Study 2 . . . 153
A.2.2 Zoho Creator App for Study 2 . . . 154
A.3 Part 1: Quantity Study 3 Additional Material . . . 159
A.3.1 Amazon Mechanical Turk Interface for Study 3 . . . 159
A.3.2 Zoho Creator App for Study 3 . . . 159

Appendix B. Part 2: Uniqueness **162**
B.1 Pilot 1 Additional Material . . . 163
B.1.1 Amazon Mechanical Turk Interface . . . 163
B.1.2 Zoho Creator App for Pilot 1a . . . 164
B.1.3 Zoho Creator App for Pilot 1b . . . 166

Appendix C. Images Collection Pilot **168**
C.1 Pilot 2 Additional Material . . . 169
C.1.1 Amazon Mechanical Turk Interface for Pilot 2 . . . 169
C.1.2 Zoho Creator App for Pilot 2 . . . 170

Appendix D. Part 3: Quality **172**
D.1 Quality Study Additional Material 173
D.1.1 Amazon Mechanical Turk Interface for Quality Study 173
D.1.2 Zoho Creator App for Quality Study 173

List of Tables

1.1 Overview of the Main Parts of the Thesis and Individual Studies 3
1.2 Overview of Hypotheses Tested and Secondary Research Questions . . . 4
1.3 Overview of Published Work from This Dissertation 5
2.1 Comparison of Ideation and Needfinding in Current Practice 14
2.2 External Factors Contributing to Concept and Need Viability 17
3.1 Summary of Study Objectives 29
3.2 Quantity Study 2 Target Sample Sizes for Selected Group Means 34
3.3 Examples of each stimulus type 38
3.4 Amazon Mechanical Turk Data Summary for Quantity Studies 40
3.5 Summary of Quantity Study Participants 40
3.6 Summary of Need and Story Results for Quantity Studies 41
3.7 Quantity Study 3 Participant Preferences Selecting Help 43
4.1 Summary of Need Statement Data Collection 66
4.2 Algorithm Performance Compared to Human Ratings 66
4.3 Summary of Potential Duplicates at Cutoff = 4 67
4.4 Highest Similarity Sentences and Scores 67
4.5 False Positive Examples: Predicted Similarity is Too High 68
4.6 False Negative Examples: Predicted Similarity is Too Low 69
5.1 Summary of Need Statement and Topics 77
5.2 Overview of Exclusion Criteria for Phases in Quality Study 79
5.3 Summary of Need Statement Data Sets 84
5.4 Summary of Need Statement Quality Ratings 84
5.5 Summary of Need Statement Data Sets for Hypotheses 5-7 85

5.6 Examples of Highest Rated Need Statements from Overall Population and Selected Population Segments 91
5.7 Examples of Highest Rated Need Statements Overall and Individual Metrics 92
5.8 Examples of Low Rated Need Statements with 10 or More Ratings . . . 92
5.9 Examples of Quality Ratings for STS Duplicate Statements 97
6.1 Data from Web-based Needfinding Methods Compared to Focus Groups Using Automated Algorithm 113
6.2 Unique Needs from Web-based Needfinding Methods (Not Similar to Those Identified in Focus Group Data) 114
7.1 Overview of Hypotheses, Research Questions, and Results 118
A.1 Topic Area 134
A.2 Topic Area Examples of Better 136
A.3 Prompt ID Numbers Assigned in Each Matrix Cell 141
A.4 Prompt Text Specific to Each Topic Area 142

List of Figures

1.1 Scope of Research Within the New Product Development Process 2
2.1 Multiple Approaches to User Research 16
3.1 Screen Capture of Quantity Studies 1-3 Needs Statement Quiz 31
3.2 Screen Capture of Quantity Studies 1 and 2 to Differentiate Needs Before and After Help 32
3.3 Summary Schematic of Quantity Study 1 and Study 2 33
3.4 Study 3 User Interface for Entering Needs 35
3.5 Summary Outline of Prompt Matrix 36
3.6 Quantity Study 2 Comparison of Stimulus Types 43
3.7 Quantity Study 3 Needs Submitted for Each Help Type 44
3.8 Study 3 Diminishing Returns with Increasing Help (lines at 90%, 95%, and 98% are shown) 45
3.9 Needs Submitted by Each Expertise Group (group sizes, n, are shown) . 46
3.10 Needs Submitted by Each Experience Group (group sizes, n, are shown) 47
3.11 Distribution of Needs Submitted per Person Across Expertise Groups . 48
3.12 Rates of Need Entries for Quantity Studies 2 and 3 (line at 90% shown) 49
3.13 Needs Submitted by Each Age Group 50
3.14 Needs Submitted by Each Gender 50
4.1 Overview of Study Work Flow, Data, and Analyses 58
4.2 Screen Capture of Sentence Pair Similarity Data Collection 60
4.3 Screen Capture of Sentence Pair Equivalency Data Collection 61
4.4 Schematic of Group Size Permutation Analysis 64
4.5 Accuracy of Predictions, Points Shown at Values for Cutoff = 3 - 4, by .1 68
4.6 Sorted Mean Equivalence Scores 69

4.7 Quantities of Unique Statements with Increasing Group Sizes 70
4.8 Quantities of Unique Statements at Cutoff Score = 4 70
5.1 Process for Analysis of One Group Size Permutation 80
5.2 Stacked Distribution of Need Statement Quality (All Phases) 86
5.3 High-Quality Needs for Increasing Group Sizes 86
5.4 High-Quality Needs (Cutoff Score = 7.5) for All Topics and Group Sizes 87
5.5 Top Quartile Needs for Users with Increasing Total Need Counts 88
5.6 Top Quartile Needs for All Topics and Expertise Groups 89
5.7 Top Quartile Needs for All Topics and Experience Groups 90
5.8 Mean Ratings for Differences in Submitter and Rater Experience 90
5.9 Quality of Need Statements for the Sequence of Needs per User 93
5.10 Count of Top Quartile Need Statements for the Sequence of Needs per User 94
5.11 Quality of Need Statements for the Sequence of Needs per User (Only Users Submitting 7 Needs) 94
5.12 Count of Top Quartile Need Statements for the Sequence of Needs per User (Only Users Submitting 7 Needs) 95
5.13 Distribution of Variation in Quality for STS Duplicates 96
5.14 Different in Quality for Duplicate Needs for Varying Similarity 96
5.15 Distribution of Variation in Quality for Omitted-Story Duplicates 97
5.16 Quality of Need Statement for Users Viewing Different Stimulus Types . 98
5.17 Quality of Need Statements and Algorithmically-Rated Uniqueness . . . 99
6.1 iStan: Commercial Medical Simulation Manikin 109

Chapter 1

Introduction

The success of a new product is often determined by the degree it satisfies customer needs and preferences. However, obtaining this mix of technical, personal, and emotional content from diverse user groups is challenging, resource intensive, and often results in an incomplete understanding of a group of users. This can be particularly evident in areas with complex, often conflicting stakeholder needs ranging from energy efficiency to water and food security. Specific areas such as health care and medical devices face additional barriers such as limited access to users and user environments.

The studies described herein introduce the design and validation of a large-scale needfinding method to better enable users to articulate their own needs. Unlike the scale of existing methods, these new methods are appropriate for needfinding output exceeding 500 need statements. This method attempts to overcome existing barriers to understanding the needs of diverse user groups. This work considers the process of needfinding to be characterized generally as a divergent process of generating needs followed by a convergent process of determining which are suitable for additional consideration. The process is viewed as analogous to ideation phases; however, the work here considers only the steps as shown in Fig. 1.1 which are unshaded and outlined in dashed lines. The portions include the divergent generation of needs and the early subjective screening phases, but not later detailed evaluations and final selections based on more complex, and often external, factors.

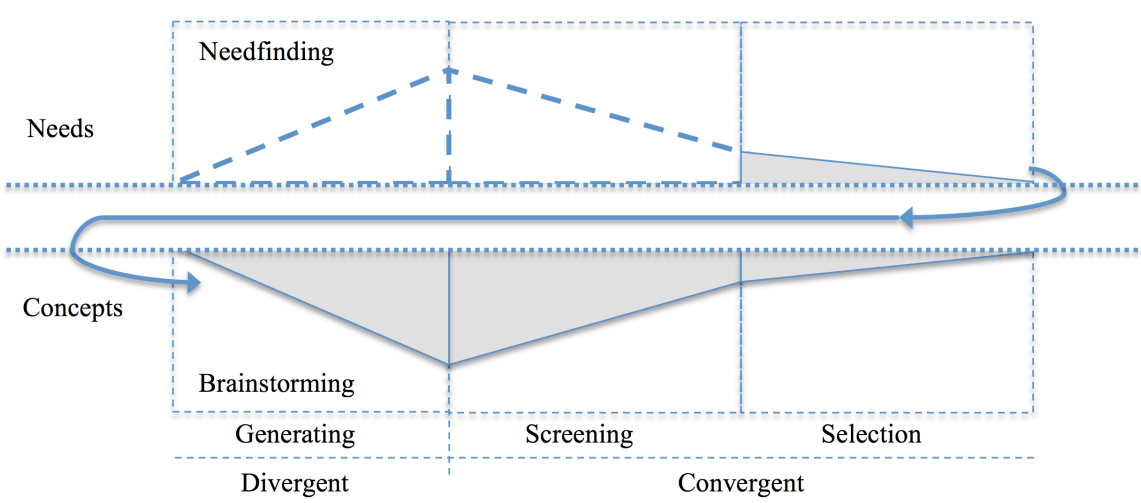

Figure 1.1: Scope of Research Within the New Product Development Process

1.1 Overview of Multiple Studies

The research is presented in four parts, each with a distinct focus and each including one or more systematic studies. Each study in the series builds upon the others and the cumulative work seeks to address five primary hypotheses. In addition, a number of secondary hypothesis and research questions are addressed that pertain to individual studies. Table 1.1 provides an overview of each part included in this thesis and individual studies within each part. Table 1.2 lists the individual hypotheses tested and secondary research questions for each part. The studies in this thesis were reviewed by the University of Minnesota IRB and were categorized as except (Study No. 1309E42581).

The body of the thesis is organized by presenting each part as a chapter. Parts 1-3 (Quantity, Uniqueness, and Quality) describe thesis work that is already published as outlined in Table 1.3. Permission to reprint copyright content has been granted for all publications [Confirmation Pending].

A brief synopsis of each chapter is provided below. Parts 1-3 all describe validation studies using three general consumer topics of: preparing food and cooking, doing housecleaning and household chores, and planning a trip. The topics were selected

Table 1.1: Overview of the Main Parts of the Thesis and Individual Studies

Part	**Studies**	**Overview of Tasks**
Part 1	Quantity Study 1 Quantity Study 2 Quantity Study 3	Evaluated several methodological elements of collecting needs from large groups, with a motivation to consider proven elements of effective ideation. Elements included a focus on quantity, specific stimulus methods to increase quantity, as well as validation of the user interface design. Tested the effects on quantity from user characteristics.
Part 2	Algorithm Evaluation Uniqueness Study	Evaluated elements of the method in order to manage large submitted data sets. The studies used an input of raw lists of submitted needs and output a subset of unique statements with a low likelihood of including duplicates. Tested rates of duplication for varying group sizes.
Part 3	Quality Study	Evaluated elements of the method to rapidly prioritize large quantities of need statements. Assessed the quality of all unique need statements from quantity study 3 based on user ratings of established quality metrics. Tested the effects on quality from user and need statement characteristics.
Part 4	Medical Technology Case Study	Applied methods previously validated using general consumer topics to a specialized area of medical simulation manikins.

Table 1.2: Overview of Hypotheses Tested and Secondary Research Questions (Bold Indicates Key Contribution)

Part	**Hypothesis/Research Question**
Part 1 (Quantity)	Does any specific type of stimulus help a user articulate a higher quantity of needs? Do levels of expertise or experience affect the quantity of needs a user can articulate?
Part 2 (Uniqueness)	Can automated machine learning algorithms detect duplication among textual need statements? **H1:** Increasing the number of participants contributing found needs increases the quantity of unique needs.
Part 3 (Quality)	**H2:** Increasing the number of participants submitting needs increases the number of high-quality needs as judged by users. **H3:** Increasing the quantity of needs contributed per person increases the number of high-quality needs as judged by users. **H4:** Increasing levels of self-rated user expertise will not significantly increase the number of high-quality needs per person. **H5:** Needs submitted first would be less likely to be high quality than needs submitted after a sustained period of time. H6: Semantically similar need statements would be rated as equivalent in quality. H7: Need statements would be rated as higher quality if a detailed description of the need context was available. Does the type of stimulus seen before entering a need affect need quality? Does the uniqueness of a need statement affect the need quality?
Part 4 (Case Study)	Will online needfinding result in similar and/or unique needs compared to focus groups?

Table 1.3: Overview of Published Work from This Dissertation

Part	**Citation**
Part 1 (Quantity)	*Large-Scale Needfinding: Methods of Increasing User-Generated Needs from Large Populations*, 2015, J. Mechanical Design [1]
Part 2 (Uniqueness)	*Large Scale Needs-Based Open Innovation Via Automated Semantic Textual Similarity Analysis*, 2015, In Proceedings of International Design Engineering Technical Conference & Computers and Information in Engineering Conference [2]
Part 3 (Quality)	*Assessing Quality of User-Submitted Need Statements from Large-Scale Needfinding: Effects of Expertise and Group Size*, 2015, J. Mechanical Design [3] [In Review] *Assessing Quality of Unmet User Needs: Effects of Need Statement Characteristics*, 2015, Design Studies [4] [In Review]
Part 4 (Case Study)	Not yet published. Planned ACS-AEI abstract.

to allow rapid evaluations of many methodological variables using sufficient statistical power to test hypotheses. The same prolonged sequence of studies would not have been feasible using specialized target groups with significantly higher resources required to recruit participants. After completing the consumer product studies, the user interface and methods were adapted for specialized users and implemented in a case study in Part 4.

The body of the thesis is organized as follows:

- **Chapter 2: Literature Review and Background for All Parts**

  Chapter 2 includes relevant literature and is divided into sections pertaining to each main part.

- **Chapter 3: Methods, Results, and Discussion of Part 1, Quantity**

  Chapter 3 presents three studies with a focus on need statement quantity. Quantity Study 1 tested one type of stimulus to increase the quantity of needs submitted by users. Quantity Study 2 compared three types of stimuli to increase the quantity of needs submitted by users. Quantity Study 3 presents adapted methods and

user interface for a complete case study.

- **Chapter 4: Methods, Results, and Discussion of Part 2, Uniqueness**

  Chapter 4 presents two studies related to analyzing how many of the needs collected in Part 1 are unique and how many are redundant or duplicate. The first study selected two state-of-the-art natural language processing algorithms and compared results of each algorithm to human ratings using a variety of algorithm training approaches. The second study used the highest performing algorithm and assessed the potential redundancy from the data of Quantity Study 3 of part 1.

- **Chapter 5: Methods, Results, and Discussion of Part 3, Quality**

  Chapter 5 includes one study to collect human ratings of all unique need statements from Quantity Study 3. The quality ratings are analyzed to assess the effects of a number of user and need statement characteristics on quality. To address the primary hypotheses, several analyses were performed on the complete data set or on subsets of need statements, such as a list of known duplicates.

- **Chapter 6: Methods, results, and discussion of Part 4, Case Study**

  This includes a case study for a needs assessment performed using the above methods. The topic of the case study was chosen to be an area particularly suited to the novel needfinding methods given the potential barriers to user access. The topic related to medical simulation manikins, which are physical, simulated patients with high-fidelity simulated physiologic responses and the ability to program a wide variety of illness and injury scenarios. These manikins are commonly used for nurse, first responder, military medic, and physician training activities. This study was performed in parallel with a Department of Defense research study to outline the requirements of a next-generation technology platform.

- **Chapter 7: Conclusions**

  The final chapter summarizes the conclusions for all previous parts.

1.2 Contributions of This Research

This thesis works contributes a number of novel needfinding methods applicable to early stage user-centered design. In addition, the results test a number of hypotheses relevant to effectively recruiting and studying users to understand unmet needs. The new methods described in this thesis differ from existing methods by providing a systematic emphasis on increasing the quantity of needs generated, capturing input from large, diverse groups rather than in-depth methods applied to small groups, and exclusively targeting unmet needs rather than product benefits or inventions. The new methods demonstrate the use of specific visual and textual stimuli to increase the quantity of needs a user can articulate and also the use of state-of-the-art machine learning algorithms to detect duplication from large data sets. The use of simplified quality metrics and crowd-based ratings demonstrates preliminary screening methods to prioritize lists of unique need statements.

Several key contributions are evident from testing multiple hypotheses (see details of hypotheses in Table 1.2). Increasing the number of participants contributing needs can increase the quantity of unique needs as well as the number of high-quality needs. Increasing the number of needs contributed per person increases the number of high-quality needs. Increasing levels of self-rated expertise will not significantly increase the number of high-quality needs per person. Needs submitted first are not lower quality than needs submitted after a sustained period of time.

The methods described were demonstrated as effective in use for general consumer products and services when recruiting users from the general population. In addition, early work applying these methods to a specialized medical area have shown promise. The medical case study contributed a list of needs submitted by medical simulation stakeholders and also an improved understanding of beneficial user interface modifications to utilize the method with a more specialized user population of technical professionals.

Chapter 2

Literature Review and Background

This chapter provides background information for relevant topics, primarily through a thorough literature review. Throughout the literature review, gaps in research to date are described and discussed as a motivation for further work.

2.0.1 Needfinding and User Requirements

Research shows a strong consensus on increasing user involvement in order to create more successful products [5, 6, 7, 8]. Recently, techniques have been borrowed from social sciences - such as anthropology and sociology - and have been repurposed to supplement information from more traditional methods such as user surveys or focus groups [6, 9].

Product design and development literature has identified numerous techniques to identify and understand user needs, such as needfinding [10, 11] (also labeled needs finding [12] or problem finding [13, 14]), user research, market research, or ethnographic research. These often are also grouped within the umbrella of user-centered design [15] or as voice of the customer (VOC) [16] and are described in disciplines ranging from product development [17] to business management [18, 19].

Needfinding is an element of user-centered design used to inform early development phases [17]. The objective of studying the user is to understand what unmet needs

exist and how these needs can inform the requirements of new products [10, 11]. One dominant theme in needfinding is to go straight to the group of users itself. Often product failures can be traced to a faulty over-reliance on input from company managers or designers rather than information directly validated with users [18]. Validating these assumptions has often required prolonged engagement to develop a deep understanding of the users' actual behavior, because actions can differ from what is said. This engagement also facilitates empathy for users, and empathy is critical for recognizing the needs and differing perspectives of users [18, 20, 21, 22]. Direct observation can have a particularly lasting influence on empathy in the observer [23], yet information on user needs can come from many sources. Direct statements from users is also one source.

The engagement with users typically occurs with observations and in-depth interviews, which might be described as methods for ethnographic research [24] or qualitative research [25]. Observational studies focus on what is done rather than what is said. They do not require a user's conscious awareness of a need in order to capture it [25, 24, 10]. Qualitative interviewing is a form of interviewing that relies on open-ended questions to allow for depth and completeness in answers where there appears to be the opportunity to uncover important insights. The questions are carefully directed by the interviewer to allow the subject to give a thorough report from the subject's point of view [26, 27].

In some cases, the process for identifying needs is intentionally divergent, to identify a large pool of potential needs. Griffin and Hauser (1993) use consumer products data to develop a function for the increasing proportion of total needs with increasing user group size. They suggest a range of 20-30 one hour interviews with different individuals with data reviewed by multiple (up to 7) analysts in order to identify approximately 90-95% of possible needs [28]. While this study does not include user observations and is a single study, it remains a commonly cited baseline. A filtering, or convergent, process follows and may be largely data driven, for example based on market size, development costs, etc, or may be similarly influenced by personal factors such as individual interests and motivations [12]. Bayus (2008) provides a thorough literature review on how this phase feeds into subsequent phases such as ideation for solutions [29].

Using an understanding of users to develop successful products remains challenging, in particular when developing radically new products [30, 31]. Within consumer products, purchasing decisions and shopping experiences may be significantly affected

by hedonic value [32, 33], and products can include complex emotional content [14] as well as symbolic meaning (e.g. evoking luxury or personal aspirations) [34]. In contrast, specialized areas such as medical device purchasing is increasingly institutionalized and data-driven, suggesting a greater importance of needs [12].

2.0.2 Types of Need or Problem Statements

Previous research on user needs quality might refer to needs, problems statements, or product requirements. For the same term, definitions often vary. Commonly, a user need is a statement created from interpretations of observations or verbatim user statements [17]. In this case, need statements are generally specific attributes expected for a new or incremental future product [17, 35] or product family [36]. Others suggest the identification of product affordances as a method for capturing user needs [37, 38]. The need can then be paired with a product requirement, indicating a quantitative metric to achieve in order to satisfy the customer [39, 40]. An example from automotive products could be a need to "accelerate quickly to merge onto highway traffic" and the requirement might be a 0-60 mph acceleration of under 10 seconds.

Ulwick uses "requirements" in a general sense (without a quantifiable metric) and points out that companies discuss requirements and include "needs, wants, solutions, benefits, ideas, outcomes, and specifications, and they often use these terms synonymously" [41, p. 17]. He assumes the most valuable customer input is task related, such as jobs-to-be-done or desired outcomes of using a product [42, 41], which is consistent with a focus on problems rather than desires [43].

A broad sense of the word "needs" is assumed for this thesis, and it is influenced by formal needfinding methods. Needfinding seeks to understand a richer breadth of user information and context than a list of product attributes [10, 11]. Ma et al. also take a broad approach, presenting short storyboards to online users. However, these story boards combine an example need with a potential solution in order to collect needfinding validation data; therefore, this data inherently combines needs and solutions [44]. The objective of the present needfinding methods is to consider only needs, agnostic of solutions, and to be mindful that statements seemingly reflecting needs can include embedded solutions [12, 10]. The need statements included in this study were collected with an explicit instruction to not include embedded solutions (e.g. a new feature or

invention). For this thesis a need or need statement reflect problems or omissions in products or services that contributes to a poor user experience or an unsatisfactory outcome.

2.0.3 Application Example: Heath Care and Medical Devices

Large-scale needfinding can benefit a variety of application areas. As an example, we herein discuss medical device development. In addition to complex stakeholder groups, this area faces additional challenges such as restricted access to user groups. Traditional needfinding methods such as in-depth or immersive observations face greater barriers to accessibility compared to consumer devices [45]. In addition, the highly regulated nature of medical devices increases the cost and time of development projects [46], and therefore, increases the risk to companies if the product is unsuccessful.

Researchers commonly noted that known, formalized methods are not consistently used within the constraints of actual industry development projects [47, 48, 49, 50]. Money et al. document a series of in-depth interviews with industry management and describe deficiencies such as primarily consulting physicians or surgeons in spite of identifying the end user as a patient or nurse. Even when device users were professionals, interviewees expressed a preference for informal methods of capturing user requirement input from "a small number of esteemed medical experts" [47, (p. 11)]. A different survey found that while "informal expert review" was among the top five methods used, it was ranked as one of the least effective [51]. A third survey, extending beyond health care, supported these findings. Here, methods such as ethnography were ranked as the most effective, but saw limited use in practice. [16]. In addition, previous studies have highlighted the improved outcomes resulting from understanding stakeholder needs from diverse groups [52].

A wide range of barriers has been observed which prevent both the use of formal methods to assess user requirements and the implementation of findings [47, 45, 53]. Barriers include cultural beliefs of management [54], limited resources for time intensive methods, lack of expertise of methods [47, 45], lack of accessibility of users, [45], and uncertain value of qualitative results [47, 53].

New methods could potentially overcome several of these barriers through remote interactions and shorter time commitments for individual users. Given limitations for

access to health care professionals, testing a sequence of new methods initially with professional groups was not feasible. Validation work engaged alternative groups of users to prepare for future case study work in target application areas, such as health care and medical devices.

2.1 Part 1 Background: Collecting User Need Statements

Needfinding shares several core challenges with ideation. Decades of research have largely addressed these in the domain of ideation, and the further study of remaining challenges in needfinding may benefit from lessons of ideation research.

2.1.1 Lessons from Ideation for Needfinding

Ideation has been studied in the context of product design [55] and problem solving in general [56]. It is a divergent process used to generate a large pool of ideas, typically focused on solving a specific problem. In this sense, needfinding could be considered similar to ideation, although to identify many needs, rather than solutions.

Brainstorming is arguably the most prominent technique used today for ideation and dates to work by Osborn (1953) [57]. His work describes the brainstorming process as an interactive group activity with a goal of generating a large number of ideas in a short amount of time. Some key procedures are to focus on quantity, encourage building off of the ideas of others, and to withhold criticism of other ideas or members. He also hypothesized that simply generating more ideas will lead to more good ideas [57].

In the years since, a trend has emerged from the general body of research supporting Osborn's hypothesis, namely that there is a correlation between quantity and quality of ideas during brainstorming. There are a large number of constructs suggested in literature for evaluating the quality of ideas [56, 55], yet the correlation between quantity and quality has been affirmed both for cumulative group quantity [58, 59, 60] and also individuals within a group [55].

Evidence also suggests that the content of the instructions used to begin the session can impact the results. Paulus et al. (2011) replicated previous results showing that instructions to "generate as many ideas as possible" improve quantity and quality results relative to control groups and groups given instructions to focus on "high quality ideas"

[60, p. 41].

A number of studies have also suggested flaws in Osborn's brainstorming hypotheses; however, critical findings typically related to one topic that is not incorporated into the present research. Osborn suggested that interactive small-group brainstorming methods could be used to increase the creative output relative to cumulative individual output. In spite of a common perception of productivity [61], a meta-analysis of over 20 studies does not support this hypothesis for superior performance in interactive groups [62]. Commonly cited mechanisms limiting group brainstorms are production blocking (e.g. interrupting each other) [58, 63], evaluation apprehension (fear of having ideas criticized) [64], and social comparisons of productivity [65, 66].

However, ideation research extends beyond interactive group methods and remains a relevant resource, in particular for individual and asynchronous methods to increase ideation productivity and quality. This research shows clear evidence that the specific techniques of ideation or brainstorming have a significant effect on the result. In particular, studies have evaluated a variety of brainstorming methods to improve outcomes including several software interfaces to mediate group interactions [67, 68, 69], and the results support the benefits of a software interface for achieving high quality and quantity in ideation [70, 71, 72]. In some cases, electronic brainstorming has been shown to be superior to similar, non-software methods [73]. Furthermore, the ability to use software to present targeted, diverse examples can improve productivity [74] and is additional rationale for similar technology applied to finding needs.

As described in Section 2.0.1, needfinding methods suggest going straight to the source, and this often means targeting expert users [18]. On the other hand, it is relatively common to brainstorm on a solution to a problem without being an expert on the problem. Evidence from crowdsourcing ideation platforms such as InnoCentive® suggests that outsiders to a specialized field can make connections across disciplines and suggest innovative solutions that were not evident to experts [75]. A diversity of users contributing to creative tasks has been shown to be generally beneficial over a range of characteristics [76].

2.1.2 Ideation and Needfinding Compared

The previous sections present a summary of both needfinding and ideation. One apparent difference is the focus on short duration group methods for ideation, such as brainstorming or electronic brainstorming, and a focus on long duration individual methods for needfinding, such as observations and interviews. This is in spite of similarities to follow a divergent path to generating a large pool of options. A more complete list of similarities and differences is shown in Table 2.1. There are a number of possible explanations why these differences may exist; however, much remains untested theory at the present time because there is a lack of research specifically studying this comparison.

Table 2.1: Comparison of Ideation and Needfinding in Current Practice (Critical Difference in Bold)

Comparison Criteria	**Ideation**	**Needfinding**
Objective	Find a new idea or the best of many new ideas	Find an unmet need or the best of many unmet needs
Process	Divergent and then convergent	Divergent and then convergent
Output Types	Incremental improvements, combinations of ideas, radically new	Incremental, blue-sky, mixed [12]
Source	Designers, managers, developers (or co-creation with users)	Users
Quality Criteria	Subjective ratings such as novel, useful, and feasible [55]	Subjective ratings such as Importance and Satisfaction [42, 41]
Achieving Quality	**Increase quantity [57, 55, 58, 59]**	**In-depth understanding and careful interpretation [29, 12]**
Current Methods	Group brainstorming	In-depth interviews and observations
Duration	Short creative sessions	Prolonged and immersive

Table 2.1 points to a specific difference in the method used to achieve quality. In the case of ideation, this is often a set of tools to increase quantity, whereas, quantity is

not commonly correlated to quality in customer needs literature. There is insufficient data to know whether the same correlation will hold in needfinding and whether this relationship could be combined with in-depth understanding to improve the quality and validity of needs.

2.1.3 Users Articulating Needs

A common perception in needfinding literature is that users typically are not able to directly articulate their own needs when asked. This is often attributed to the formation of habits or dogma, as users adapt to shortcomings and no longer recognize them or are not easily able to see beyond the existing set of solutions [18, 12]. Researchers studying ideation have struggled with similar effects, often described as design fixation [77]. While cognitive mechanisms may not be identical, it is worth noting that the presence of fixation in ideation is seen as an obstacle to be overcome, not as a fundamental flaw of the method [78, 79]. The scrutiny in research shown in ideation has not been directed towards testing methods to overcome this fixation-type effect in needfinding.

Individual users may struggle to articulate needs when asked simple questions, and this limitation is a significant motivation for more in-depth interactions. Figure 2.1 shows a schematic to represent different approaches to this limitation. In the case of immersive methods, such as interviews used by Griffin and Hauser [28], the number of needs articulated from basic questions is small; therefore, in-depth interviews are used to guide the discussion, interpret comments and help the individual think of a significantly greater quantity. Large-scale needfinding represents a method to help users articulate their own needs using specific types of stimuli and collecting this data via a content-rich, interactive online application. The types of stimuli might include visual (e.g. images) or textual (e.g. examples of needs). In this approach, the quantity of needs articulated without help remains small. The quantity increases with the help of various stimuli, although the increase may not be to the same degree as with in-depth methods. However, an increasing portion of the total needs space can be filled by increasing the quantity of users. This is particularly useful if the available quantity of users is high, even if individual users have little time. This is typically the case in crowd-sourcing scenarios.

Later phases of product development have adapted methods from cognitive psychology to incorporate user verbalizations as data. Think aloud protocols originally suggested for cognitive studies [80] are now considered essential tools for product and software usability studies, in spite of concerns over inconsistencies in practice compared with the original theoretical underpinnings [81].

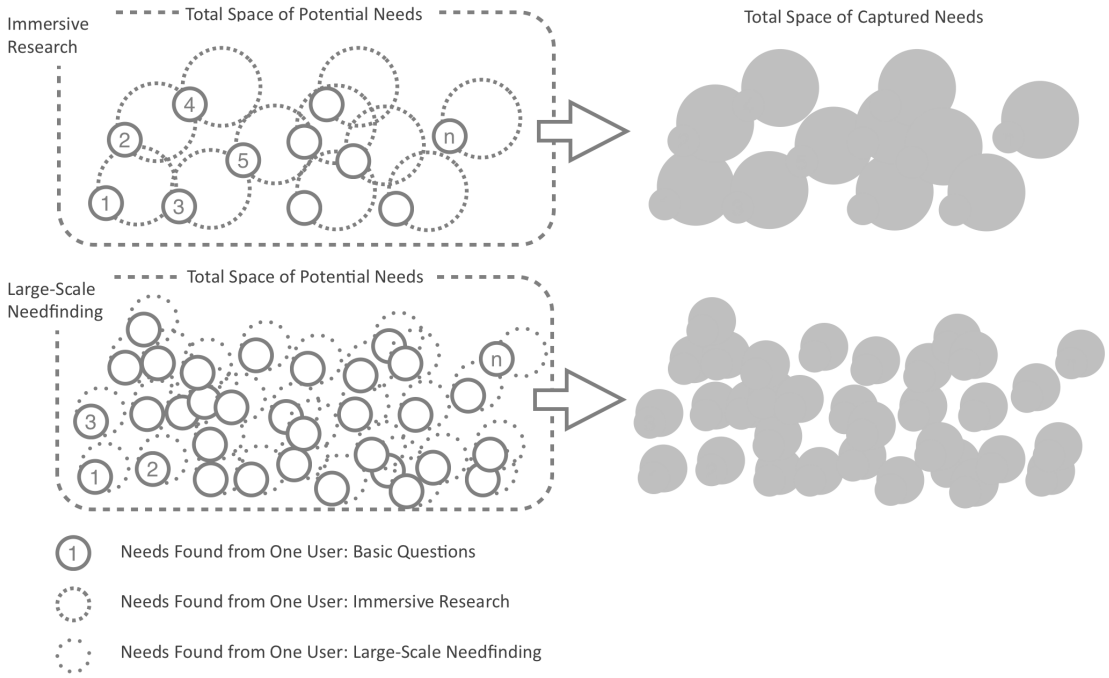

Figure 2.1: Multiple Approaches to User Research Intending to Maximally Cover the Total Needs Space

Many methods to understand user needs focus on improving a designer's ability to identify someone else's needs [82, 29]. Few methods have been rigorously tested to help a user better articulate his own needs, although successful methods such as empathy tools have been reported in a trial study [83]. The use of crowds appears often for ideation ("open innovation") and later phases of product development [75, 84]. Crowd data has been used for needfinding and user preference modeling; however, this work employs data mining of existing content such as blog posts and comments rather than direct solicitation of needs [85, 86]. A gap remains for needfinding applied both to crowds and to directly solicit needs from users. In spite of a lack of prior research

targeting crowd-submitted needs, Faste (2011) explicitly states that advances in online knowledge management "could be applied to crowdsourced needfinding research"[87, p. 5], and further observes "Perhaps one of the most important ways in which open-innovation can therefore be made to thrive is by enabling individuals to report their own needs." [87, p. 4]

As shown in Figure 1.1, the scope of the present work recognizes the complexity of factors contributing to the quality of a need and the difficulty in converging on a final selection for product development. The proposed research will focus entirely on quality aspects independent of external factors in Table 2.2. Within this narrower scope, the emotional content of needs remains an important consideration. Because of this complexity of understanding needs, successful needfinding may never be fully decoupled from a deep understanding of the user. Needfinding may not entirely follow the trajectory seen in ideation and brainstorming; however, testing where specific similarities do exist can inform future research and may improve some needfinding approaches.

Table 2.2: External Factors Contributing to Concept and Need Viability But Excluded from the Scope of the Present Work

Category	**Potential Factors**
Market	Large potential market, effective distribution
Legal	Patent, copyright, trademark protection
Reimbursement	If applicable: Feasible strategy for approval by public payor and insurance agencies (medical devices) [12]
Regulatory	If applicable: UL approval, OSHA safety requirements, FDA approval or clearance, etc.

2.1.4 Rationale for Large Quantities of Needs

As described in Section 2.1.1, a correlation does exist between high quantity and high quality in ideation, and this process shares a similarity with needfinding in that the desired outcome is a pool of potential candidates to pursue further. Prior to testing this correlation, a method to collect large quantities of needs would be a necessary first step. In addition, given the challenges identified in Section 2.0.1, uncovering a unique need

can be a rare event and a higher quantity of attempts would have a higher likelihood of a rare event occurring. Ultimately, what matters is the count of high quality needs, not necessarily the proportion. A final output of 10 high quality needs and 500 poor quality needs would be superior to 2 high quality needs and 50 poor quality.

2.1.5 Rationale for Our Stimuli Design

A majority of research using stimuli in creative tasks has focused on ideation, and stimuli are predominately visual (e.g. sketches, images) [77, 79]. However, textual stimuli have been effective for design tasks in architectural design [88]. Recent studies have begun to apply analogous methods to identifying needs. Participatory methods have been suggested a means to improve the designer's empathy of needs [29, 89], and preliminary studies support the use of empathy tools to aid users articulating needs [83].

Three types of stimuli were tested in these past studies (described in Part 1) and were selected as feasible for use in online applications and with text-based need statements. They include shared example need statements, example contextual images, and short narrative prompts. The shared needs stimulus was intended as analogous to ideation and brainstorming sessions where participants are primed with the ideas generated by others. Dugosh and Paulus (2005) have evaluated this method for ideation, which assumes that exposure to ideas from others will stimulate new ideas [59]. The positive effects of shared ideas in ideation have also been reported for increasing the number of ideas categories [90], increasing idea generation in electronic brainstorming sessions [91], and increasing combinations based on shared ideas [92]. Verbatim shared idea content is not required, in fact, subtle encouraging cues are also sufficient for increasing idea generation [93], even via electronic media.

The contextual image stimulus was intended to help provide context to the activity, as most participants might be at a computer far removed from an environment related to the topic. Availability of context is previously described as a key rationale for observational study. Retaining contextual information may potentially trigger useful insights [10, 6]. Visual examples have been used previously for priming and mitigating fixation in ideation, and the effects have been positive as well as negative [77, 94, 79]; however, this study presented images for a more general purpose. The images were assembled to be more than a set of visual examples of problems. They represented broader, general

context for the topic area.

The prompt stimulus was intended as a substitute for probing questions present in qualitative interviews. Some allowed participants to focus on particular elements of products, for example a specific faulty or broken product. Others allowed participants to focus on particular events, such as a recent emotional experience. The objective was to facilitate self-reflection or thoughts of empathy.

2.2 Part 2 Background: Assessing Uniqueness of Need Statements

The use of stimuli and the explicit focus on quantity, essentially modeled on brainstorming, are elements of a systematic approach to collect the greatest number of need statements as possible for a given topic. Once collected, these statements must first be processed to differentiate unique and redundant entries. A list of only unique needs will be more manageable and will facilitate an efficient assessment of need quality.

2.2.1 Large Data Sets in Ideation

Open innovation is a method for seeking ideas for innovation from external sources. The importance of this trend has been previously discussed and shown to be successful in several applications [95, 96, 97, 98]. However, open innovation processes result in large quantities of submitted content [99, 100], and the resources required to assess and filter redundancy and quality impede the use of open innovation in practice. Survey results of companies with open-innovation experience note complaints that reviewing and assessing externally-submitted ideas takes "an army of internal people" [16, p. 15]. A recent Cisco open innovation project required a team of six full-time employees working for three months in order to evaluate 1200 submissions [100].

Commercial idea management systems exist such as Ideascale (www.http://ideascale.com) and The IdeaWall (www.theideawall.com). Ideascale includes keyword search for predictive duplication merging. While exact methods are not listed for these commercial systems, existing keyword methods such as Lucene [101] have previously been used. However, existing systems must still rely on administrator oversight or knowledge of

users to identify and flag duplicates. Relationships between ideas may be more complex than only duplication; however, annotations of relationship hierarchies also remain largely manual [102]. Idea overflow and redundancy remain a challenge for managing open innovation data [99, 102]. Assessing similarity using pairwise comparisons of submissions analyzed with automated algorithms such as Natural Language Processing (NLP) may identify redundancy and reduce resources needed to manage data.

2.2.2 Natural Language Processing Background

NLP algorithms utilizing machine learning are increasingly studied and are rapidly improving [103, 104]. A goal of natural language understanding—a subtopic of NLP—is to comprehend the intended semantic content of text. This is of particular relevance to textual design processing such as needfinding or analyzing textual innovation content. Modern techniques have moved well beyond keyword analysis or parts-of-speech comparison to extracting the concepts in or semantic meaning of sentences, phrases, and passages. The increasing trend towards employing probabilistic machine learning techniques ensure that semantic content can be automatically extracted from otherwise prohibitively large corpora, for example, from phrases never seen in the original training set used to tune the algorithms. While still in its infancy, it is evident that the accuracy and speed of these semantic techniques continue to improve with increasing momentum [104]. Simultaneously, the need for arduous supervision and human input during training or use continues to decrease [104]. This suggests that certain NLP approaches can not only enable the automated semantic processing of textual design content but do so at large scales (e.g. large-scale needfinding).

Semantic Textual Similarity (STS) is a form of NLP analysis used to measure the similarity of two phrases or sentences. Two distinguishing elements of STS are graded and symmetric ratings. A graded rating refers to a sentence pair being "more" or "less" similar than another sentence pair. A symmetric rating indicates there is no directionality when comparing sentence A to sentence B (e.g. A to B is the same as B to A). STS also provides a framework to cohesively combine a number of different NLP components, such as word sense disambiguation and induction, lexical substitution, and semantic role labeling, into a single evaluation [105]. Because of these unique characteristics, STS may be a useful tool to perform pairwise comparisons and assess

redundancy in open innovation content. In addition to obvious applications in idea management, STS methods are appropriate for novel open innovation tasks such as evaluating large sets of user-submitted needs. Needs descriptions are inherently text-based and might represent complex, nuanced aspects of a problem, thus necessitating state-of-the-art analysis methods.

This work (described in Part 2) selected two STS algorithms based on performance and ease of use. Both had previously been a top-performer at international NLP competitions, allowed Python script querying, and were freely available. In the present work, two algorithms were trained and tested using multiple data sets and then used to rate similarity for need-based open innovation.

2.2.3 SemEval Algorithm Competition Background

SemEval is a series of evaluations coinciding with the *SEM conference (Joint Conference on Lexical and Computational Semantics). For each conference, the organizers supply standard annotated data sets in a wide variety of NLP tasks, including English and multilingual versions. As is typical in machine learning research, an algorithm is "tuned" to an application area by setting internal parameters during a training step that invokes a training data set. Next, the algorithm's performance is evaluated on a different data set, not included in the training set, which is referred to as the evaluation set or test set.

During SemEval, research teams submit algorithms to run the supplied training and test data, and the outcomes are ranked based on evaluation metrics, such as Pearson correlation to gold standard data (human ratings of similarity). The 2012 SemEval was the first to introduce a semantic textual similarity (STS) rating task, and a similar task was repeated in 2013. Each year had over 30 participating teams. In this task, the similarity of two text passages (e.g. sentences) is computed on a scale of 0 (different topics) to 5 (completely equivalent) [105, 106].

The STS task provided several different data sets. The 2012 task used training and test data derived from the same sets. The 2013 task used the same training data as 2012 but provided several new data sets for testing. One example of a provided data set is the MSR Video Paraphrase Corpus (MSRvid) set, originating from Microsoft Research. The data was collected from human participants who were describing a short video

clip. These descriptions were combined with descriptions of other similar and different video clips to produce a set of 1500 sentence pairs with a wide range of similarity. The MSRvid and other data sets included short sentences, often with common words, such as "A chef is slicing a vegetable" [105].

2.2.4 Algorithm 1 Background: TakeLab-simple

The TakeLab-simple system was one of two 2012 SemEval submission from the TakeLab research group (University of Zagreb, Croatia) for the 2012 STS task. The final mean ranking of the system was 2nd overall for 2012. The TakeLab group provided open-source files for TakeLab-simple following the conference. The TakeLab-simple algorithm combines a variety of tools into an aggregate similarity score for two text passages. These tools included knowledge-based word similarity using WordNet, corpus-based word similarity using Latent Semantic Analysis (LSA), and several others [105, 107]

2.2.5 Algorithm 2 Background: UMBC-PairingWords

The UMBC-PairingWords system was one of the three 2013 SemEval submissions from the UMBC Ebiquity research group (University of Maryland, Baltimore County and Johns Hopkins University). The final mean ranking of the system was 1st overall for 2013 in the CORE task. The UMBC group provides an online interface and Python-based code to query the existing system (http://swoogle.umbc.edu/SimService/index.html). The UMBC algorithm also combines a variety of tools for an aggregate similarity score [106, 108]

2.3 Part 3 Background: Assessing Quality of Need Statements Submitted by Users

When a large set of need statements has been condensed to only a subset of unique need statements, the value remains limited unless there is an indication of quality. Assessing the quality of *ideas* generated during later phases of development has been thoroughly studied and previously summarized [56]; however, the development of quality metrics for need statements is much more limited. Three commonly cited or particularly relevant

examples are described here in more detail.

2.3.1 Kano Model

The Kano model is a framework developed in the 1980's for classifying different types of user requirements [109]. A number of researchers have since expanded upon this framework and developed varying methods of collecting survey data with specific questions to determine the classification for individual requirements [35]. The model describes three types of desirable requirements or attributes: a basic requirement (also called a dissatisfier or must-be), a performance requirement (also called hybrid or one-dimensional), and an excitement attribute (also called satisfier or attractive). Two undesirable, and less common, requirements are indifferent and reverse [109, 35, 110, 111, 43].

After identifying the list of requirements, customers answer a pair of questions for each requirement. One asks what satisfaction results from the fulfillment of the requirement. The other asks what satisfaction results from the absence of the requirement. The relative rates of high satisfaction and dissatisfaction determine the classification. While the classification implies a degree of importance, the specific relative priorities may require additional computation, in particular when a trade-off must be made. Potential methods include analytical hierarchy process [112], Taguchi methods [110], Monte Carlo simulation [111] or as an element of quality function deployment or house of quality [43, 113]. Reports of these analytic methods often limit the quantities of statements (75 or fewer) [110, 114].

2.3.2 New Product Design and Development Texts

Ulrich and Eppinger suggest determining relative importance of features using survey data from customers. The authors differentiate between verbatim customer statements and translated customer needs, typically representing product features or "attributes" [17]. Features are arranged hierarchically, consistent with Voice of the Customer methods [28]. The set of features used can be a subset of the total with a preference for those where importance is non-obvious. For example, obvious critical features for a product to function can be omitted from the survey. The suggested survey uses two questions: a rating of importance 1 (Undesireable) to 5 (Critical), and a checkbox to indicate if

the feature is exciting or unexpected. The practical limit for prioritizing statements is suggested as about 50 [17]. While quantifying the excitement from a feature might imply the degree existing products satisfy the particular need, a more explicit question might be beneficial.

2.3.3 Importance and Satisfaction of Outcomes

Ulwick describes a simplified approach to quantify user preferences and applies the method to lists often exceeding 100 statements. A unique element of this method is a strict adherence to listing only the performance outcomes relevant to the job a specific product will perform [42]. The author states that an unfocused reliance on statements representing product solutions or benefits is a reason why Voice of the Customer methods continue to produce unpredictable results [41].

Rather than list "brakes" as a basic requirement of a vehicle, the performance outcome that impacts purchasing decisions might be "Minimize stopping distance on slick roads." The complete list of outcomes is developed during a series of in-depth interviews with individuals from a wide range of demographics and experience levels. Analysts interpret what is said in interviews and rephrase statements into discrete outcomes using the form "Minimize X" or "Maximize Y". Ideally, each rephrased statement is read back to the participant to validate the intended meaning in real time. The consistency in language is used to minimize variation in prioritizing [41].

Once the list of outcomes is complete, it is distributed to a large number of potential users (often between 180-600), and respondents rate each outcome on two criteria: How important is each outcome, and To what degree do existing solutions satisfy these outcomes? Average responses for each criteria are entered into a linear formula to rank outcomes with high importance and low current satisfaction. These outcomes are termed the "Opportunity" score and become priorities for future development [42, 41]. These metrics share similarities with those used in quality function deployment [43], but incorporate fewer additional weights and calculations to facilitate implementation on a larger set of statements. The formula given by Ulwick to calculate the Opportunity score is shown in Equation 2.1.

$$Opportunity = Importance + max[(Importance - Satisfaction), 0] \tag{2.1}$$

The current study used a quality criteria derived from the Opportunity calculation by Ulwick but with important changes. The Opportunity equation used by Ulwick [42] looses fidelity when the satisfaction is high and importance is low. In Ulwick's calculation, Satisfaction is subtracted from Importance, but cannot go below 0. Statements might be rated the same Opportunity but have different Satisfaction scores. This was justified as acceptable because the impact was limited to low Importance needs, thus not altering final priorities. However, this calculation might impact analysis of correlations performed in the present study.

2.3.4 Differences from Previous Work

Previous research has prioritized need statements using similar methods. However, as previously described, there are numerous variations in methods, such as varying definitions of need statements. In addition, previous work focuses with differing degrees on population overviews or segments of a population [113, 28]. Other methods are intended to inform requirements on specific products or a product category [17, 42]. Critical areas where current methods typically differ are summarized below.

1) Needs not solutions: The content of need statements in this study differed from existing similar user research methods. Primarily, the scope of statements emphasized problems experienced by users or desired outcomes, not necessarily product features or attributes. Features relevant to a particular solution were explicitly discouraged.

2) Population overview: The output of quality ratings were not necessarily used to target a specific population segment (e.g. "soccer moms"). The list of highest-rated statements in this study represented an overview or cross-section of problems commonly experienced. These problems could later be addressed through innovative new products or services using existing new product development and/or open innovation methods. The analysis of overall priorities may be combined with an assessment of population-segment preferences as both points of view might be valuable to prioritize new projects depending on the target market.

3) Not product specific: The list of top-rated statements did not necessarily represent an exhaustive list that should be implemented into a single product. Because of this, there was no need to specifically measure user preferences when a trade-off must be made. Subsets of high-rated statements relevant to a specific target product should be

further assessed to inform these trade-off decisions.

4) Quantity focus: A quality metric for need statements is a necessary first step to analyze what processes might improve the quality of need statements collected during early stage research. One approach to increase need quality is to systematically increase need quantity (as is common for ideas during ideation phases). Previous research might have pursued divergent user needs research, but without an explicit focus on quantity.

Chapter 3

Part 1: Collecting Large Quantities of User Need Statements

3.1 Needfinding Topics

In order to facilitate a high quantity of input from a general population, the needfinding topic areas were confined to common tasks that rely on products or services. The study did not screen for any particular degree of familiarity or expertise, rather, these data were collected along with demographic information to allow subsequent covariate analysis of any potential differences in outcomes that are correlated to level of expertise. The focus areas included both tasks relying on physical objects or products and tasks relying on services or software. The topics were selected to be familiar to a majority of individuals recruited from online communities and provided variation in the types and nature of products and services that might be discussed.

Physical object/Product:

1. Preparing food and cooking

   This includes any step you take to start with food on the shelf or in the refrigerator and end with a meal ready to eat.

2. Doing housecleaning and household chores

   This includes cleaning up messes, whether dirt or clutter, doing laundry, sorting mail, and other jobs around the house.

Service/Software:

1. Planning a trip

   This includes any travel beyond daily routines. Trips might be work-related, vacations, local, abroad, by yourself, or with others.

3.2 Part 1 Methods Overview (Quantity)

Three studies were completed to evaluate different aspects of the needfinding method. Study methods differed by necessity to address different objectives as shown in Table 3.1. Each study included an online user interface to collect open-ended need statements. This interface was combined with a method to display stimulus information to potentially help increase the quantity of needs a user could articulate.

Table 3.1: Summary of Study Objectives

Quantity Study 1	Test matrix of prompts to increase need quantity
Quantity Study 2	Compare control group, prompts, shared needs, and shared images to increase need quantity
Quantity Study 3	Test unstructured availability of three help types to evaluate a case study scenario

All three studies asked users to submit single-sentence statements describing problems or unmet needs relating to a single topic. After entering a need statement, a participant could enter a more elaborate story to describe relevant background information. Each participant was randomly assigned using a software-based random calculator to one of three topic groups: preparing food and cooking, doing housecleaning and household chores, and planning a trip.

All participants were recruited from Amazon Mechanical Turk (www.mturk.com/) (AMT). AMT is a site allowing a community of task requesters (analogous to employers) to recruit individuals from a community of online workers. The tasks are divided into discrete deliverables called HITs (Human Intelligence Tasks), and workers self-select to begin this task when viewing lists of available tasks. Workers are paid nominal amounts for each deliverable. Pay is generally proportional to task duration and falls within a broad range of approximately $.10 per minute. AMT is increasingly used as a source for research participants, and the user population has been previously characterized [115]. Data integrity from AMT workers can be maintained, in particular when targeting high reputation workers [116]. Participants had to meet basic requirements in order to be eligible. These included approval rates of 95% or higher for completed work, a history of at least 100 completed HITS, and a United States IP address location. Each study allowed repeat workers who had previously completed an earlier study. In this case, the worker would automatically be assigned to a different topic area than any previously seen. Complete details for the AMT interface for each study are provided as an Appendix in Sections A.1.1, A.2.1, and A.3.1.

Participants recruited from AMT were directed to a custom web application developed using Zoho Creator (https://creator.zoho.com/). Zoho Creator is a cloud-based custom database platform with integrated logic scripting and graphic user interface (GUI) development tools. AMT workers would accept the HIT and would see that the

objective was to describe problems with common products and services. Instructions were framed in a variety of ways that might be clear to a wide range of people. For the topic of cooking, instructions included: "We want to know what would make preparing food and cooking a better experience. Examples can be very broad, for example: more convenient, less effort, safer, easier to understand, cheaper, more consistent, or faster." (examples differed for each topic, see Appendix Section A.1.2) "You will type in descriptions of problems or unmet needs you face preparing food and cooking." "You want to describe these so someone could make improvements or offer solutions in the future. Try to think of as many as you can."

After reviewing consent information, participants completed a training exercise. Training began with brief instructions stating that inventions should not be included and to describe the problem in a complete sentence. Participants then took a quiz including five example statements and were required to identify which were not consistent with the instructions. The examples and quiz related to a new topic (reading books) to avoid providing example needs relevant to assigned topics. A screen capture of this portion is provided in Figure 3.1. The training was paid as a fixed amount of $0.65 for both pass and fail outcomes. Participants who failed were not able to continue.

Following the quiz, participants answered optional demographics questions including gender, age, and self-reported levels of expertise and experience (hours per week). When training was complete, participants began entering needs and stories. The final instructions, again for the topic of cooking, were "Don't worry about whether the benefit is worth the cost. We simply want lots of suggestions." Each entry was paid as an individual bonus. Bonus amounts varied for different studies, as shown in Table 3.4. Exact base and bonus payment amounts were displayed in the instructions. Complete details for the Zoho interface for each study are provided as an Appendix in Sections A.1.2, A.2.2, and A.3.2.

Quantity Studies 1 and 2 were designed to differentiate between needs readily available to the user and those that may have arisen as a result of viewing some type of stimulus information. These studies presented users with two options in the interface: "Enter Another" and "I'm Stuck" buttons. The display for these button choices is shown in Figure 3.2. The "I'm Stuck" button was described as the option if the participant was not sure what to say. The intention was to treat initially submitted needs as

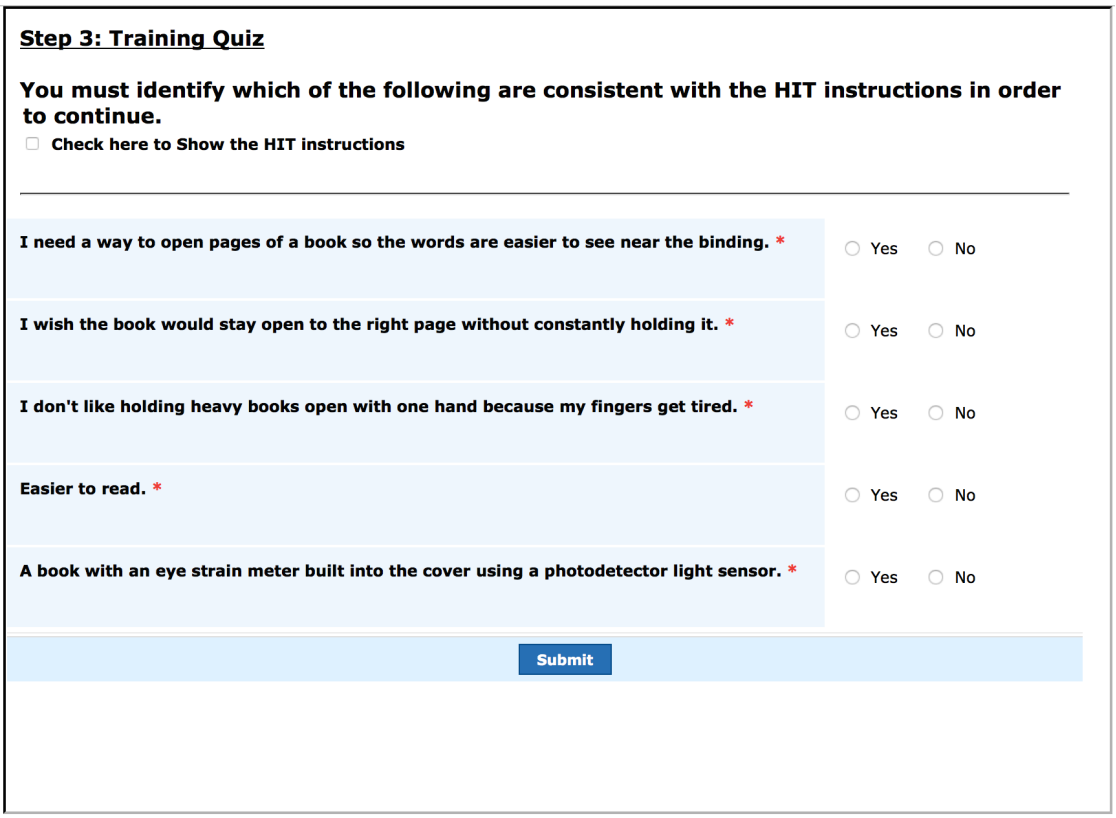

Figure 3.1: Screen Capture of Quantity Studies 1-3 Needs Statement Quiz

available to the user simply when asked, and needs submitted after pressing "I'm Stuck" as potentially generated through viewing the stimulus. The stimulus was described to users as "help", and the available help differed for each study. For this discussion, "stimulus" and "help" can be considered interchangeable. Table 3.1 describes the types of help available for each study.

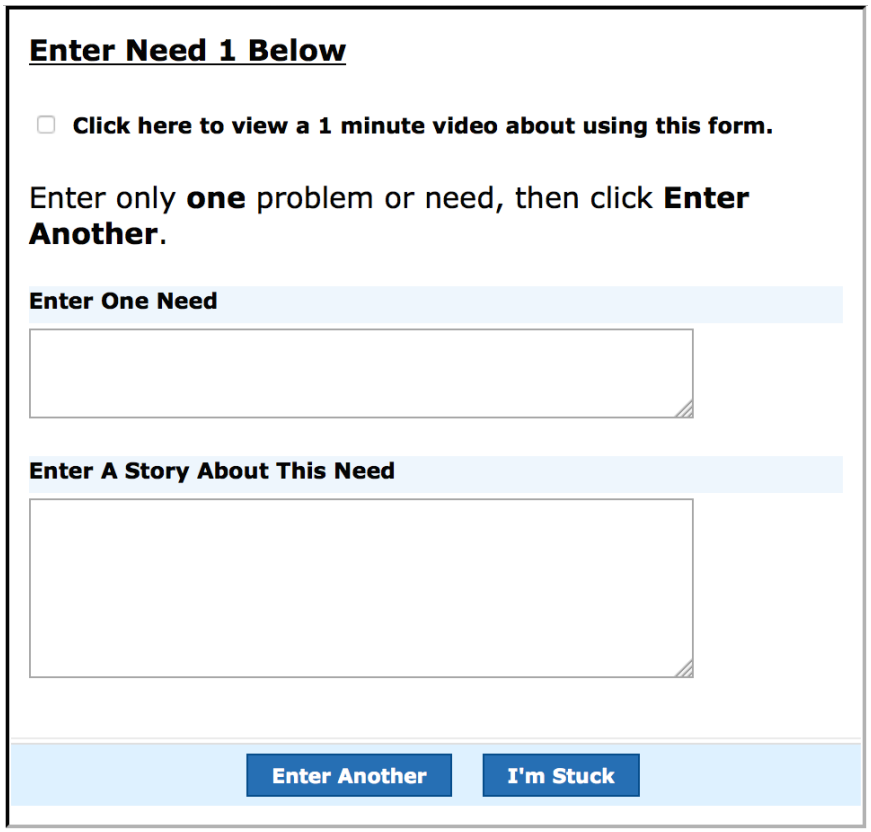

Figure 3.2: Screen Capture of Quantity Studies 1 and 2 to Differentiate Needs Before and After Help

After viewing a stimulus, the user returned to the interface to enter any new needs and stories. A general process schematic is shown in Fig. 3.3.

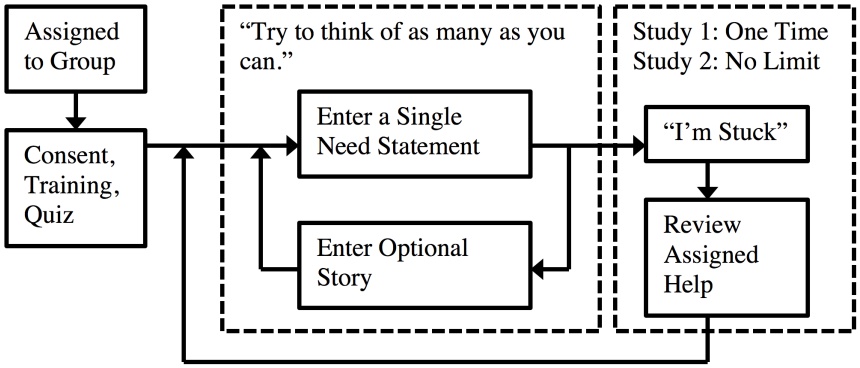

Figure 3.3: Summary Schematic of Quantity Study 1 and Study 2

3.2.1 Quantity Study 1 Methods

Quantity Study 1 tested the effectiveness of a single type of stimulus, a paragraph-length narrative prompt, described in more detail in Section 3.2.5. The effectiveness of a type of stimulus was measured using a count of needs submitted after selecting I'm Stuck and reviewing the prompt. Participants in Study 1 who clicked I'm Stuck twice were shown a message that only a single help was available and then the study ended.

Quantity Study 1 screened a total of 30 prompts (including a control). The study employed a sample size of 15 users per prompt. Quantity Study 1 data was analyzed to identify if any prompt or trait of prompts resulted in a lower mean of needs submitted after viewing. These prompts could be omitted in future studies. Prompt traits were analyzed by grouping prompts along rows or columns of the complete matrix described in Section 3.2.5. A likelihood-ratio test was used to determine the best fit model for count data comparing Poisson and negative binomial models. A regression analysis was used for the best fit model to test differences of groups (models tested differences of log(means)). A multiple comparison test (multcomp R package using "Tukey" parameter [117]) was used on the generalized linear model to test pairwise combinations of prompt matrix rows and columns.

The needs count data from Quantity Study 1 were used to calculate a sample size for groups in Quantity Study 2. For an initial approximation, the distribution was assumed to be a Poisson distribution to provide a more conservative estimate despite some evidence of over-dispersion. The approximate sample size would be dependent on

the assumed group means of needs per person, with a desired delta across group means of one need per person. Table 3.2 shows a range of assumed means with this delta value and the resulting range of sample sizes.

The assumed rate of failed training was 50% based on previous studies. However, given the uncertainty in the pass/fail rate and the potential for exiting the study prematurely, a conservative target sample size for Quantity Study 2 was 100 per group. This exceeds the calculated sample sizes shown in Table 3.2. In order to achieve this group size for passing and complete responses, Quantity Study 2 recruited 150 participants per treatment group.

Table 3.2: Quantity Study 2 Target Sample Sizes for Selected Group Means

Θ_1, Group 1 Mean [needs]	1	2	3	4
Θ_2, Group 2 Mean [needs]	2	3	4	5
$n^* = \frac{4}{(\sqrt{\Theta_1}-\sqrt{\Theta_2})^2}$, Sample Size	24	40	56	72

* for Poisson distribution

3.2.2 Quantity Study 2 Methods

In this study, three types of help were available and are listed in Table 3.1. Details for each type (as well as a control) are given in Sections 3.2.4-3.2.7. The first time a user selected "I'm Stuck", the randomized help was selected from the three types and control group. In Quantity Study 2, selecting "I'm Stuck" a second time allowed participants to begin to select any additional help at will. Quantity Study 2 analysis used the same metric for effectiveness of a stimulus, specifically, the number of needs submitted after viewing the stimulus. Only needs entered after viewing the first help but before viewing any subsequent help were included in this metric.

Quantity Study 2 data was analyzed to test for a significant effect of stimulus type (prompts, shared needs, images). Statistical tools were identical to Quantity Study 1, as described in Section 3.2.1.

3.2.3 Quantity Study 3 Methods

Quantity Study 3 differed from the process shown in Fig. 3.3 by providing options to view any of the available help from the beginning, omitting the "I'm Stuck" button. This study randomly assigned participants into topic groups, but did not assign participants into any test groups. Histograms, empirical cumulative distributions, and descriptive statistics were used to make basic observations such as the number of times participants would choose to view additional help and the resulting number of needs submitted. Quantity Study 3 used an interface relevant to a case study application where options to quit or receive ongoing help were readily available. Figure 3.4 is a sample image from Quantity Study 3. This screen represents a point where a participant in the travel group had selected to view the stimulus type of images, and a scrollable list was displayed. The right-hand portion of the screen was consistently used for entering needs. The target sample size for each topic was approximately 125 per topic group, resulting in a similar total size compared to Quantity Study 2.

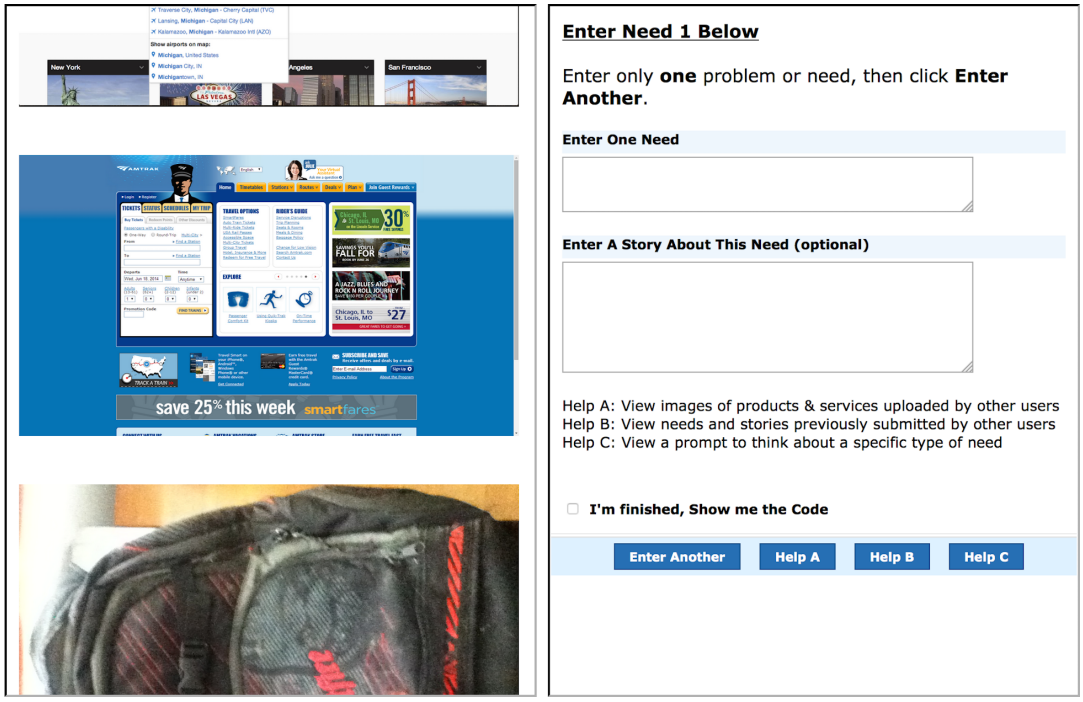

Figure 3.4: Study 3 User Interface for Entering Needs

3.2.4 Control Stimulus

Studies consisting of a control group used nominal additional bonuses, for example a "double bonus", to encourage continued participation and potentially limit the rate of quitting prior to reviewing the stimulus information. A control group would be offered only this additional bonus, and each treatment group would be offered the same additional bonus and also a display of stimulus information. This additional bonus was not considered a treatment, as it was consistent for all groups and effects of incentive were not tested.

3.2.5 Stimulus 1: Narrative prompts

The first type of stimulus was a prompt to ask users to think about a particular task from different perspectives. This focus may help identify a particular type of need. The prompts were arranged in a matrix to organize these perspectives based on similar traits. For example, one axis of the matrix related to differing content, such as a focus on a particular emotion (e.g. frustration) or type of communication (e.g. instruction manuals). The other axis of the matrix related to different subjects, such as a first-person view or a third-person view. Traits were derived from design empathy literature [18, 20, 21] and interviewing methodology [26, 27] and combined with new variations. Each cell of the matrix contained one or more combinations of these traits. Figure 3.5 shows an outline view of the matrix rows and columns.

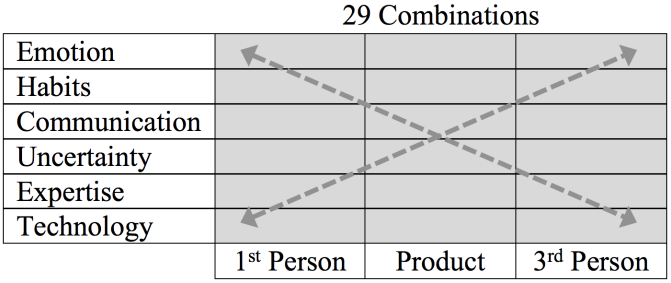

Figure 3.5: Summary Outline of Prompt Matrix

This type of stimulus included a total of 29 prompts combining a variety of traits described above. The same group of prompts was used for all studies in Part 1. Table

3.3 shows an example of a complete prompt. The complete details of traits, prompt matrix cells, and prompt content are provided as an Appendix in Section A.1.3.

3.2.6 Stimulus 2: Shared Needs and Stories

The second type of stimulus allowed a user to read the entries submitted by previous users. Quantity Study 1 was used to collect this pilot data. Needs submitted in Quantity Study 1 were reviewed, and incomplete sentences and inventions were omitted. Only needs with an accompanying user-generated story were shared in Quantity Studies 2 and 3. The complete group of needs was randomly ordered and grouped into batches of 10 needs and 10 corresponding stories. The total shared needs content included approximately 30 batches available for each topic. Table 3.3 shows an example of a need/story pair selected as one with a particularly vivid description.

3.2.7 Stimulus 3: Shared Images

The final stimulus was a display of a series of content-specific images submitted by previous users. An independent pilot was used to collect these images. This pilot is described as Pilot 2 and the full details are provided as an Appendix in Section C.1. The pilot was repeated twice, each time assigning a participant to one of the same topic groups used for Quantity Studies 1-3. Pilot participants were asked to avoid uploading images with identifiable information. In the first iteration, the pilot participant was asked to upload an image of a product or service used for or relevant to the topic. The second iteration asked the pilot participant to upload an image relating to something the person disliked about the topic. Images submitted in these pilots were reviewed and irrelevant images or images with faces or identifying information were omitted. The complete group of images was randomly ordered and grouped into batches of 10. The total shared images content included 10 batches for each topic. Table 3.3 shows an example of one shared image included in a batch for planning a trip.

3.2.8 Stimulus instructions

Users assigned to or requesting the narrative prompt stimulus were instructed to read the passage and see if thinking about the topic in this way resulted in any new needs. Users

Table 3.3: Examples of each stimulus type

Example narrative prompt for cooking	"Think of a time when you tried preparing food and cooking, and the result did not end up how you had hoped or wanted. You were expecting to get a certain result, but that isn't what happened. Can you identify any reasons why you didn't get the outcome you expected? What problem could be addressed to help get the outcome you wanted?" This prompt combined a first-person view with content relating to uncertainty.
Example shared need and story for cleaning	Need: "I wish there was an easier way to clean the back side of the toilet that is hard to reach." Story: *"The last time I cleaned the bathroom I got down on hands and knees as usual to clean the back part of the toilet. To my dismay I found that I had to hug the nasty toilet to even reach that part, and I have long arms, so I can only imagine how my wife gets back there to clean. I wish there was something to [sic] would make it easier to reach that part of the toilet without necessarily being hard on your wrists or hands or unnecessarily heavy."*
Example (cropped) image for planning a trip	

assigned to or requesting shared needs or images were instructed to review the complete batch as inspiration for a type of brainstorming activity and to think of new needs related to the shared information or anything new that comes to mind. Each participant was shown a random batch corresponding to the assigned topic. The participant never viewed repeated content (the same prompt or image), specifically when repeat help was available.

3.3 Part 1 Results (Quantity)

Summary data for each study is presented in Tables 3.4-3.6. Table 3.4 includes data relating to the AMT system and worker payments. In total, approximately 1730 workers were paid for completing portions of a study, and of these 1,135 participants entered need statements that were included in an analysis. Table 3.5 includes data relating to the Zoho survey database. In total, the 3 study surveys were accessed approximately 2300 times. The disparity in participant counts between the AMT data and the Zoho data is due to multiple exit points during the survey that were prior to completing training and getting an authorization code to be paid by AMT. For example, many participants agreed to be in the study, but quit after reading the instructions. Table 3.5 provides a list of participants who were excluded from analysis due to failing training or incomplete data.

Table 3.4: Summary of Amazon Mechanical Turk Data for Quantity Studies

	Study 1	**Study 2**	**Study 3**
Total HITs Submitted	530	600	601
Total Base Payments (USD, excluding Amazon fees)	$ 335	$ 390	$ 390
Total Bonus Payments (USD, excluding Amazon fees)	$ 228	$ 299	$ 273
Bonus for Needs (USD)	$.20 for 5	$.05 ea	$.05 ea
Bonus for Stories (USD)	$.10 ea	$.15 ea	$.15 ea
Study Duration (days)	20	2	1

Table 3.5: Summary of Quantity Study Participants

	Study 1	**Study 2**	**Study 3**
Granted consent and began the study	775	725	810
Excluded			
Quit during training	87	96	171
Did not pass training quiz or attempted to retake	276	264	219
Passed training but quit before need entry	44	0	18
Included in analysis	368	365	402
Repeat Workers (included in total)	N/A	4	57

Table 3.6 provides an overview of needs and stories submitted with each study. In

total, approximately 6000 need statements and 3750 stories were collected. Some data in Table 3.6 is not available for Quantity Study 1. Story entry length was inaccurate because a number of participants combined multiple needs into a single entry and the accompanying story may not have described all needs. Also, Quantity Study 1 did not record beginning and end times when participants were entering needs. Lastly, help was offered only a single time per user in Quantity Study 1.

Table 3.6: Summary of Need and Story Results for Quantity Studies

	Study 1	**Study 2**	**Study 3**
Total Workers Submitting 1+ Needs	355	347	341
Total Workers Submitting 0 Needs	13	18	61
Total Needs Submitted	2441	1795	1735
Total Stories Submitted	1172	1332	1246
Average Need Length (characters)	84	73	74
Average Story Length (characters)	N/A	278	269
Min/Median/Max Needs per Person	0/6/68	0/4/34	0/3/28
Min/Median/Max Minutes to Enter Needs and Stories	N/A	2/16/116	1/11/172
Total Count of Help Views	N/A	483	549
Workers Viewing 0 Help	0	0	206

While the systematic assessment of unique and non-unique need submissions is described in Section 4.1, a preliminary review of data did not indicate malicious copying of other needs, particularly given opportunities to view shared needs. Complete need sets for Quantity Studies 2 and 3 were reviewed using standard software (R) to compute total sentences and total unique sentences. In addition, each complete need set was sorted alphabetically and manually reviewed for duplicate or near-duplicate entries (e.g. missing punctuation). Potentially copied sentences were less than 1% of totals in all sets.

3.3.1 Quantity Study 1 Results

A negative binomial regression analysis was used for testing difference of log(means) for Quantity Study 1 based on likelihood-ratio test results. The negative binomial model fit is preferred over Poisson due to the presence of count data with over-dispersion.

The results of a two-sided test indicated there were no individual prompts, rows, or columns with a significant difference lower than others. Likewise, pairwise comparisons

for both rows and columns did not reflect any significant differences. The lack of isolated lower performing types of prompt content gave no rationale to exclude any particular prompts in future studies.

Two individual prompts showed a significantly higher mean (p-values of less than 0.001 and 0.003), although the highest prompt mean corresponded to the most prolific individual (triple the count of the next highest individual). Only a single row and column mean showed a greater than marginal difference (p-value less than 0.01), and these corresponded to the row (p-value = 0.0005) and column (p-value = 0.004) of the most prolific individual. The median number of needs submitted after a prompt was the same for all prompts (1 need).

3.3.2 Quantity Study 2 Results

A negative binomial regression analysis was again used for testing difference of log(means) for Quantity Study 2 with the same rationale as in Quantity Study 1.

Due to a slightly higher than anticipated rate of exclusions for failed training (see Table 3.5), the final group sizes included a minimum of 90 per group, rather than the target of 100.

Figure 3.6 compares the needs submitted after viewing a stimulus for each type tested in Quantity Study 2.

The shared needs group resulted in a significantly higher mean of needs submitted compared to the control and prompt groups (p-values = 0.003 and less than 0.001, respectively). The shared images group was a marginally higher mean compared with the prompt group (p-value = 0.04).

3.3.3 Quantity Study 3 Results

Table 3.7 lists the number of times each type of stimulus was voluntarily requested during Quantity Study 3. Voluntary selections by all users reflects the number of times each type of help was selected for the entire study population. Voluntary selections by users viewing at least one of each help type show the number of times each type of help was selected by the subset of users who had the opportunity to see all three types and would have known what type of content is shown for each.

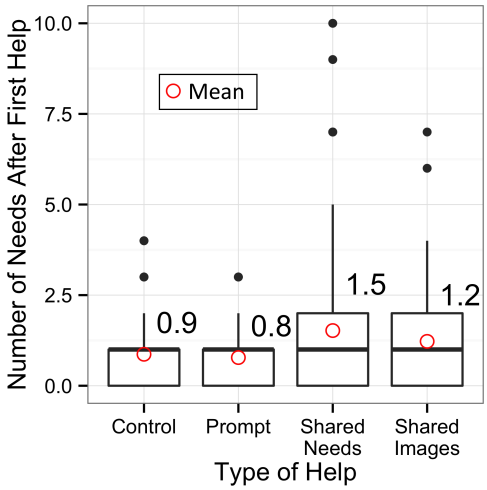

Figure 3.6: Quantity Study 2 Comparison of Stimulus Types

Table 3.7: Quantity Study 3 Participant Preferences Selecting Help

	Prompt	**Shared Needs**	**Shared Images**
Voluntary Selections by All Users	143	202	204
Voluntary Selections by Users Viewing at Least One of Each	114	132	130

Figure 3.7 shows how many needs were submitted after viewing each type of help for each request for help. The needs submitted at 0 help selections reflect those entered before viewing any help. For Quantity Study 3, the maximum number of help selections was not limited, and each additional selection shown resulted in additional needs entered.

The cumulative distribution of needs submitted after repeated requests for help is represented in Fig. 3.8. The figure includes needs submitted before viewing any help, indicating just under 50% of needs were submitted before viewing help. Observe a diminishing return for continuing help requests where 90% of all needs were attained after the first three helps, and 98% were attained after eight helps.

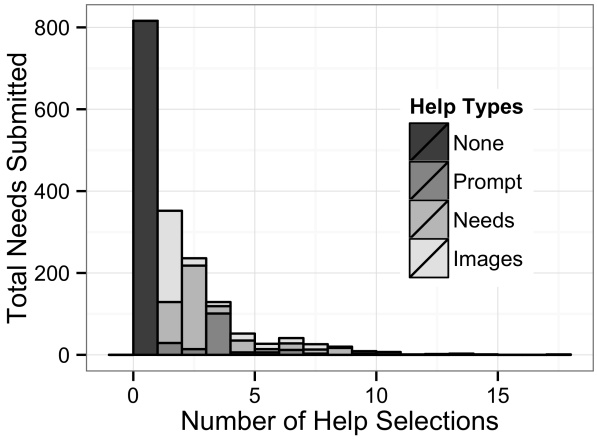

Figure 3.7: Quantity Study 3 Needs Submitted for Each Help Type

3.3.4 Aggregated Observations for All Three Studies

Data was aggregated only for descriptive statistics and measuring covariate effects. The analysis of the effects of topic (e.g. cooking or cleaning) and other covariates was performed on combined data. The relative contributions of level of expertise (self-rated), experience (self-rated hours per week), topic area, and study iteration were tested with likelihood-ratio tests for negative binomial regression models. The topic area was not significant (p-value = 0.42); therefore, data from all topic areas are aggregated for this analysis.

The level of expertise was not a significant variable (p-value = 0.13). Figure 3.9 shows the total number of needs submitted per person for each study and each expertise level. The group size for differing expertise groups varied from approximately five professionals per study to approximately 200 intermediates per study.

The level of experience (hours per week) was a significant variable (p-value = 0.003). Figure 3.10 shows the total number of needs submitted for each study and each level of experience. The group size for differing experience groups varied from approximately 30 individuals per study with 10+ hours per week experience to approximately 220 individuals per study with up to 5 hours per week. The “Up to 5 Hours” group mean was 0.9 higher than the “None” group, up to a maximum difference of 2.4 higher for

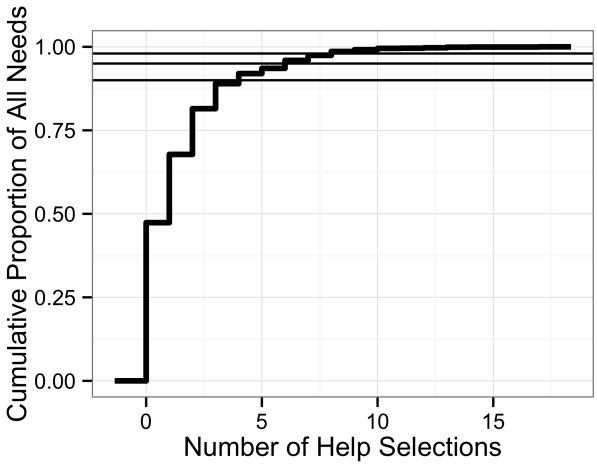

Figure 3.8: Study 3 Diminishing Returns with Increasing Help (lines at 90%, 95%, and 98% are shown)

"More than 10 Hours" compared to "None". A multiple comparison test (using only the significant model factors of study and experience) showed higher experience levels consistently resulting in higher need quantity. "5-10 Hours" and "More than 10 Hours" were significantly higher than "None" (p-values less than 0.001 and 0.001, respectively) and "5-10 Hours" and "More than 10 Hours" were significantly higher than "Up to 5 Hours" (p-values = 0.009 and 0.046, respectively).

Figure 3.11 shows the relative contributions of needs submitted for the complete study for each expertise level. The variation in group sizes is again evident, and descriptively, the shape of distributions for each group are similar.

Figure 3.12 shows the cumulative distribution of need submission over the duration of Quantity Studies 2 and 3. Quantity Study 1 was omitted as Table 3.4 shows that the duration of Quantity Study 1 was an order of magnitude greater than the other studies. Quantity Study 2 showed a distinct change in slope at approximately 20 hours, and after this point proceeded with a rate similar to Quantity Study 3. Quantity Study 3 reaches a point of 90% at approximately 8 hours, corresponding to an average rate of approximately 200 need statements per hour over this time interval.

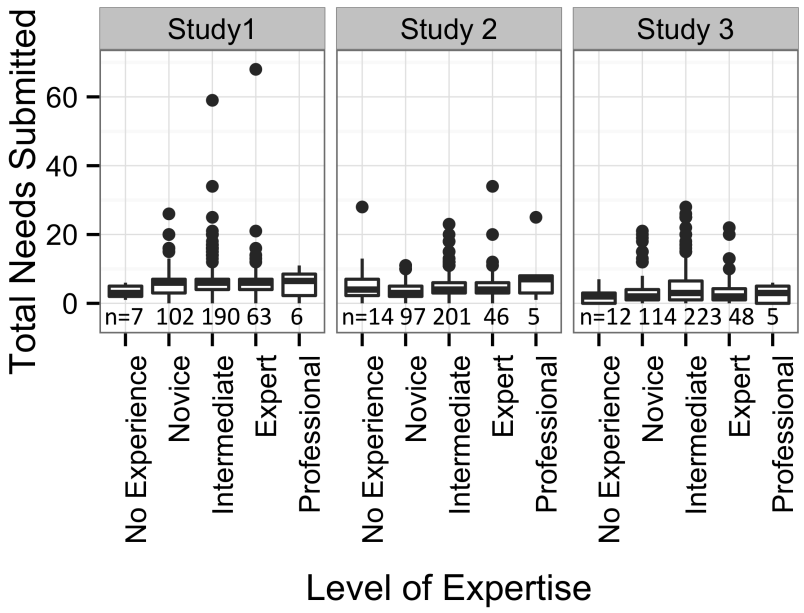

Figure 3.9: Needs Submitted by Each Expertise Group (group sizes, n, are shown)

3.3.5 Aggregated Observations for All Three Studies: Unpublished

Some analyses were conducted and not included in final publications as shown in Table 1.3. The age and gender of each participant was collected as part of the optional demographics information. After publication of expertise and experience results, a regression analysis was completed with age and gender variables included. Age and gender had been excluded from the initial combined analysis due to no evidence of age and gender effects from earlier study-by-study analyses.

The relative contributions of level of expertise (self-rated), experience (self-rated hours per week), topic area, study iteration, age, and gender were tested with likelihood-ratio tests for negative binomial regression models. The topic area was not significant (p-value = 0.41); therefore, data from all topic areas are aggregated for this analysis.

As before, the level of expertise was not a significant variable (p-value = 0.22). Age was also not a significant variable (p-value = 0.14). In a combined model with all studies, and including study iteration as a covariate, gender was a significant variable (p-value less than .001). As before, self-rated experience (hours per week) was a significant

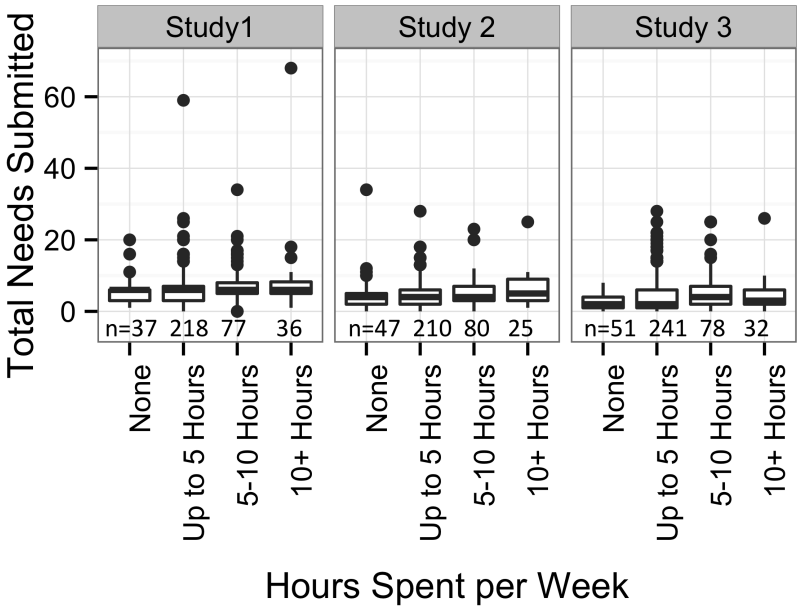

Figure 3.10: Needs Submitted by Each Experience Group (group sizes, n, are shown)

variable (p-value = 0.008).

Figure 3.13 shows the total number of needs submitted for each study and each age group. Figure 3.14 shows the total number of needs submitted for each study and each level of experience.

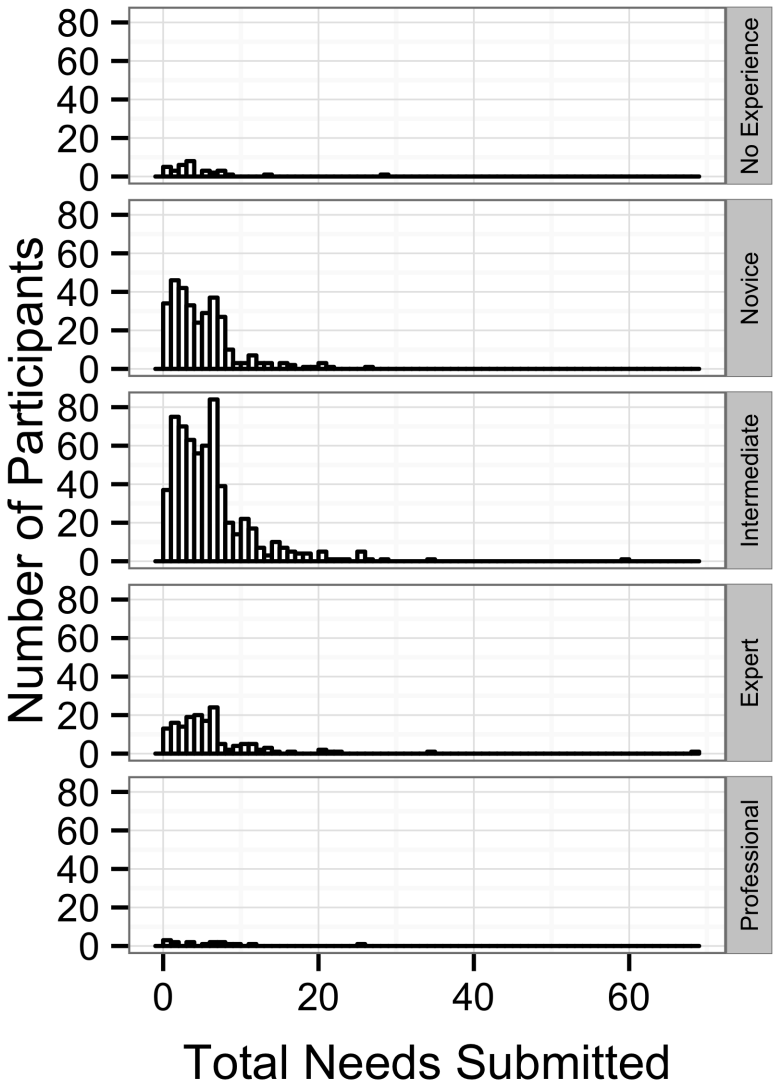

Figure 3.11: Distribution of Needs Submitted per Person Across Expertise Groups

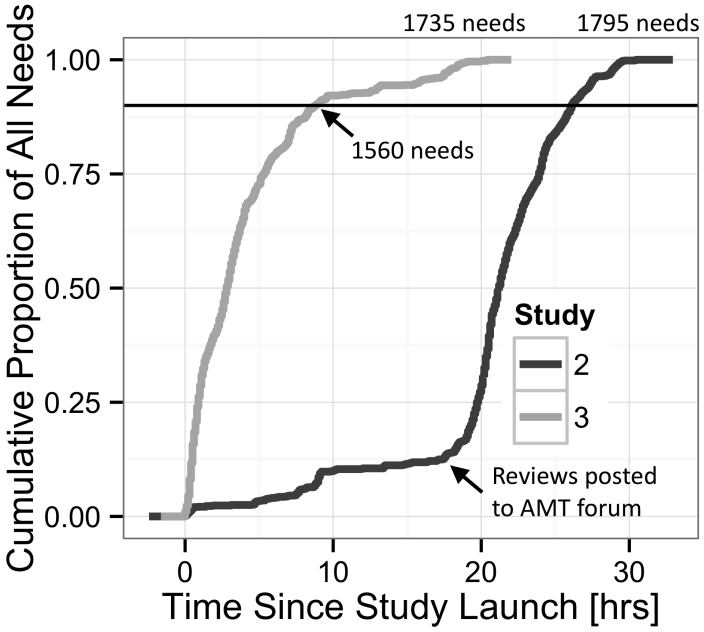

Figure 3.12: Rates of Need Entries for Quantity Studies 2 and 3 (line at 90% shown)

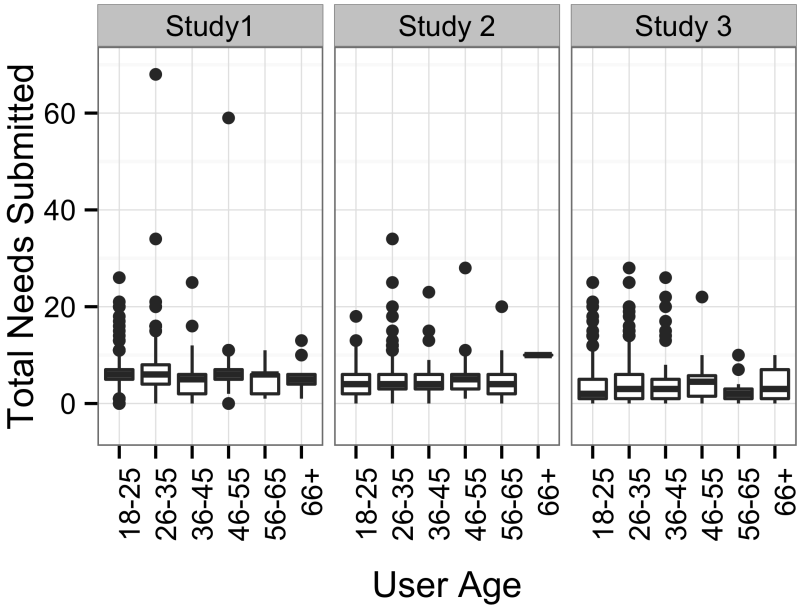

Figure 3.13: Needs Submitted by Each Age Group

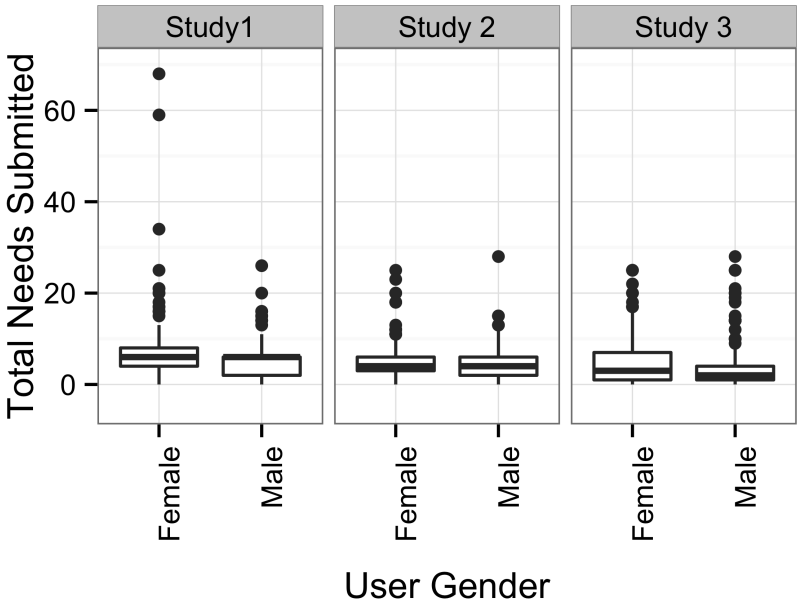

Figure 3.14: Needs Submitted by Each Gender

3.4 Part 1 Discussion (Quantity)

The goal of this study is to contribute a needfinding method allowing rapid collecting of needs from large groups. The results showed strong evidence supporting large group needfinding and spurred multiple important observations.

3.4.1 Fast, Large-Scale Need Collection is Feasible

Figure 3.12 shows that in the equivalent time of one day of traditional ethnographic observation, an alternate crowd-based method can collect 1500 need statements and 1100 stories. This is not sufficient to suggest this method is superior to existing ethnographic methods, only that there may be a higher rate of needs and that at a minimum, this source of input could complement data from interview (or observational) sources.

3.4.2 Collecting Needs Does Not Require In-Depth Research

These studies provide strong evidence that users will have the ability to articulate needs directly when the interaction is mediated by sufficient background and instructions, incentives, and stimuli. There is rationale to assume additional types of stimuli and incentive structures may further improve the outcome of directly soliciting needs from users. However, this does not suggest interviews or observations should be omitted when resources and user access permits. Data from traditional methods may continue to increase understanding and empathy at any phase of development and may also help identify, clarify, validate, and prioritize a set of needs.

3.4.3 Effects of Incentives and Stimuli

The results of these studies indicate that specific stimulus types can significantly impact the quantity of needs collected, and the incentive structure appears to influence user behavior. This was evident comparing different help types; however, Quantity Study 1 results did not suggest conclusive evidence that specific traits or prompt content were a significant factor. Additional study may be necessary to identify what specific content is most effective. While needs per person for different stimulus groups may vary by a relatively small difference of means (less than 1), this method consists of aggregating needs for hundreds of individuals, and the resulting effect of the combined group could

be a difference of several hundred needs. The outcome of the shared needs and shared images is a significantly higher quantity of needs; however, these types of stimuli require pilot data. Real-time sharing may reduce the dependence on pilot data for studies where all users begin at approximately the same time, but this might be less successful in an asynchronous method as used here. Although a direct financial incentive showed some positive effect as a control, some application areas will prohibit direct payment incentives, so a prompt stimulus may still be useful in absence of pilot data. This positive effect of stimuli is consistent with previously demonstrated improvements in user needs generation when providing users with empathy tools for extreme use scenarios [83].

The specific amount of bonus payments may have had an effect on user behavior. The bonus per story increased from $0.10 for Quantity Study 1 to $0.15 for Quantity Study 2 and Quantity Study 3. The change was motivated by a goal to increase what was viewed a as valuable source of additional information. The proportion of needs submitted with stories increased from approximately 50% to a minimum of 72% after this change.

3.4.4 User Expertise and Experience are Not Interchangeable

In spite of potential similarities between a user's expertise and experience, the former was not a significant variable and the latter was. One potential reason for the discrepancy could be a user's inaccuracy or bias in self-rating expertise. A sense of expertise may be influenced by multiple factors including past experience or comparisons to immediate peers. Specialized users might have expert status based on credentials rather than recent experience. In other words, an expert user may have formerly spent a significant time on the task, but no longer does. With this consideration, needfinding results might improve when prioritizing users with current experience over expert status.

However, this difference may not, in fact, warrant targeting only higher experience levels in practice. The increase from "Up to 5 Hours" (5.0 needs per person) to "More than 10 Hours" (6.5 needs per person) is 1.5, or approximately 30%. In this case, the aggregate effect discussed for stimulus groups may not be seen here because high experience groups generally were much smaller. The difference in mean should be considered in conjunction with other real-world factors such as overall cost as determined partly by recruiting costs. In a scenario where higher experience in users results in a 30%

cost increase, this method would collect a greater number of needs for a lower cost by recruiting available workers even if they are not highest in experience.

3.4.5 Collecting data on Amazon Mechanical Turk

Quantity Studies 1-3 were conducted as validation activities to prepare a needfinding method for use in a specific application area of clinical care delivery or training. However, the results do support the potential use for collecting large quantities of needs from the general public.

Table 3.4 clearly indicates that Quantity Study 3 benefited from a consistent and rapid rate of need entry; however, other studies had differing results. A likely explanation for the 20 day duration of Quantity Study 1 is found in the information-sharing infrastructure of crowd sourcing communities. There are a number of AMT worker forums where workers post reviews of completed HITs and rate the quality of the task and fairness of the requester regarding payments. Requesters who launch a first study have no reputation of task quality or fairness to aid in recruiting workers should they investigate a requester prior to starting. Each study was performed with a conscious effort to provide an experience worthy of positive AMT forum feedback, including setting clear expectations, a fair pay rate, and prompt payment processing. A slight increase in rate during Quantity Study 1 (not shown) and a significant increase in rate for Quantity Study 2 (see Fig. 3.12) corresponded in time to positive reviews posted to worker forums. It is likely this gradual accumulation of positive, public feedback contributed to an initial high rate of recruiting and need submission for Quantity Study 3.

Also of note, Table 3.6 shows that Quantity Study 1 actually finished with the highest count of needs regardless of the fact that help was most limited. One potential explanation again points to the influence of worker forums. Quantity Study 1 had little initial feedback posted to forums and quickly became one study in a sea of thousands of available tasks. Contrast this with Quantity Studies 2 and 3 where early positive feedback gave the study high visibility among a subset of the crowd who rely partially on this input to decide which tasks to complete. It is possible later studies were taken by crowd workers based on factors such as a reputation for prompt payment rather than the study content. Note that number of workers submitting 0 needs increased with each study as did the number of workers who quit during training (see Table 3.5).

An additional contributing factor could be the change in incentive structure, from a quota system in Quantity Study 1 to a piece rate system in Quantity Studies 2 and 3 (see Table 3.4). This would be consistent with improved AMT outcomes for quota approaches previously described [118]. There were several rationales for switching structures. One was to smooth the data to create a more uniform distribution because quota structures create bimodal or multimodal distributions. In addition, a payment for every 5 needs seemed to increase confusion and lead to workers entering 5 needs as a single entry. Lastly, the need quantities collected in Quantity Study 1 exceeded expectations, and the benefit of the quota system may not have outweighed the costs given a proficient crowd.

3.4.6 High Volume of Needs without Stimulus

Figure 3.8 reflects that users can readily articulate nearly 50% of the cumulative total need quantity with no official help, independent of evidence that certain types of stimulus can have a significant effect on the count of needs (see Figure 3.6). Here it should be noted that while this figure represents voluntarily selected help specific to the assigned topic area, it is not inclusive of all information that would be useful to workers. Not only did each worker review the instructions and training examples, but a short video summary of instructions was also available, and each worker then saw additional examples during the quiz. Nonetheless, this result shows that these controlled stimuli are beneficial but not required for large quantities of needs.

3.4.7 User interface, quantities, and rates

The data in Table 3.6 provides insight into the importance of user interface design and the potential influence on user behavior. In particular, each study recruited approximately the same number of workers. Quantity Studies 2 and 3 collected approximately the same number of needs and stories; however, the median duration each worker spent entering needs decreased 45%, from 16 to 11 minutes. One potential explanation is the effect of interface design. Quantity Study 2 was testing a specific treatment effect, and did not immediately present workers with a button to end the study while entering needs. This was intentionally withheld until after viewing the assigned help. With this

interface, 18 of 347 workers quit without entering any needs.

Quantity Study 3 was a modified interface with the rationale that readily available options would be appropriate in a case study application. Here the button to quit and buttons to access each help were clearly displayed from the beginning of need entry. The number of workers quitting without entering any needs increased to 61 out of 341 (and were paid for completing training). The presence of this greater number of early departures lowered the median duration, but the remaining workers, on average, submitted more needs and the cumulative total was approximately the same. This increase in needs for the non-zero workers could potentially be attributed to the full availability of early and more help.

The increase in maximum duration shown in Table 3.6 from 116 minutes to 172 minutes corresponds to an increase in maximum allowed time for the study (3 hours instead of 2 hours). While a majority of users exit long before this time, a flexible structure allowing engaged users to continue might benefit total counts.

3.4.8 Limitations and Future Work

A number of limitations to this work should be addressed. Perhaps most important, this data reflects only the total quantity of needs and not the quality or redundancy. Establishing a correlation for finding needs, as has been done for ideation, will require discriminating duplicate and semantically similar needs and then rating unique needs for quality. These methods are described in Parts 2 (Uniqueness) and 3 (Quality).

Second, these results are dependent partly on reputation building and specifics of user interface design. This will have an effect on replicating results within the same crowd, in addition, a strong reputation in one crowd will likely not completely transfer to another when applying this method to long-term application areas such as clinical professionals.

Third, the method relies on what a user says, which may differ from behavior. Additional validation can come in the form of targeted follow-up observations. In addition, a quality assessment process (described in Part 3) can be used to average a large number of ratings to minimize the influence of individual users.

Lastly, the topic areas used in these studies were intentionally general. The similarity in outcomes for the three general topics is not definitive evidence that the method will

be equally effective in a specialized topic, but these studies provide strong rationale and motivation for such work. Specialized topics will require specialized user groups and may require alternate recruiting strategies, but these groups often already have existing crowd structures potentially available for a large-crowd study. An example in the medical application area would be a national association or annual conference of a medical specialty. The incentive structures devised for this study may not be appropriate for future application areas, but relevant incentives exist for specialized crowds, including peer recognition [119] and formal education credits [120]. A medical case study including assessment of need quality will provide valuable data to further assess these limitations.

Chapter 4

Part 2: Assessing Uniqueness of Need Statements

4.1 Part 2 Methods Overview (Uniqueness)

The two STS algorithms described in Sections 2.2.4-2.2.5 were used to test similarity of previously generated need statements. First, the performance of two automated algorithms was evaluated to determine a preferred algorithm. Second, the preferred algorithm was used to test for uniqueness. The objective was to differentiate between sets of duplicate and unique need statements. A summary of the multiple training, test, and analysis data sets used for this series of studies is shown in Figure 4.1. In summary, Figure 4.1 describes how key inputs were determined prior to use in the uniqueness study. These inputs were the final, preferred automated algorithm and the cutoff score to differentiate between unique and duplicate statements. Details are described below.

The methods described in Part 1: Quantity Study 3 were used to generate the analysis set of need statements for conducting the uniqueness study. The analysis set is summarized in Table 5.1. Algorithms were trained using a training set of need statements independently collected during earlier Part 1 Quantity Studies.

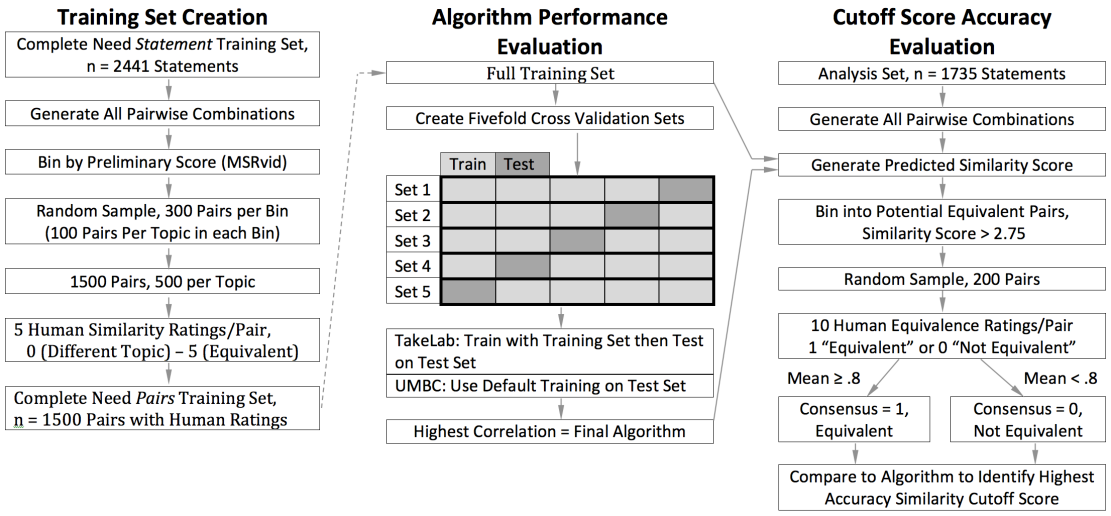

Figure 4.1: Overview of Study Work Flow, Data, and Analyses

4.1.1 Need Statement Preprocessing

The list of submitted needs consisted of a sequential Need ID number, a need statement, an optional story, an AMT user ID, a record of all stimulus information previously viewed, and a topic area key. Need statements were separated based on topic to create three independent data sets.

Preprocessing steps were executed using Python scripts. A first step removed contractions and non-standard characters. A second step generated tab-separated files of full-text sentence pairs as input files for the algorithm similarity rating. These input files provided necessary combinations in order to compare each need to all others. The results would determine which statements might be duplicate or redundant. A matrix with all needs on both the horizontal and vertical axes provided all pairwise combinations. A need statement would be compared to itself; however, if two statements, A and B, were represented with A:B, then the reverse B:A would be redundant and was not required (half of the complete matrix was used).

4.1.2 Need Statement Training Sets

The complete need statement training set (see Figure 4.1) was processed into pairs and initially rated using a provided, trained algorithm. This algorithm was the TakeLab system trained using provided MSRvid training data. The result was a preliminary score to help create a smaller subset with an approximately uniform distribution of similar and unique pairs. The ratings were binned by preliminary score, and randomly sampled for 300 pairs per bin. The complete need pairs training set was comprised of 1500 pairs (500 pairs per topic) across the 0-5 score range. This training set was rated using AMT workers in a fashion similar to MSRvid data [105]. This human ratings similarity study is referred to as Pilot 1a, and the complete details for the AMT and Zoho interfaces are provided as an Appendix in Sections B.1.1 and B.1.2, respectively. A screen capture of the sentence similarity data collection user interface is shown in Figure 4.2. In order to evaluate algorithm performance, the complete need pairs training set was partitioned into five training and five test sets (each with 100 pairs per topic) for a fivefold cross validation.

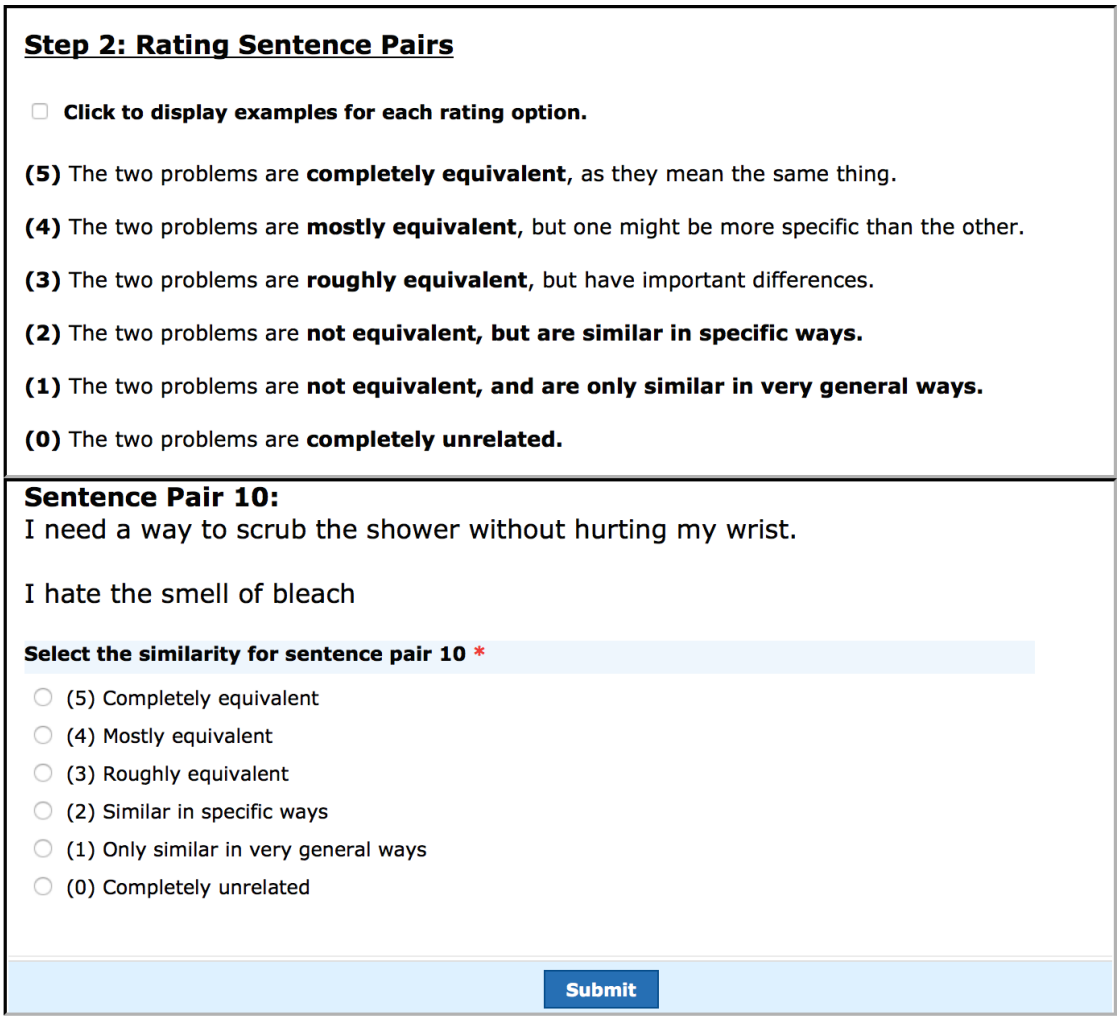

Figure 4.2: Screen Capture of Sentence Pair Similarity Data Collection

4.1.3 Similarity Cutoff Scores

The analysis of potential duplicates used a cutoff score to divide pairs of need statements into potential duplicates or potential unique entries. The analysis assumed the cutoff score would represent a point where two statement were considered equivalent in meaning. In order to evaluate the accuracy of a cutoff score, a sample of need statements was taken from the need pair analysis set with predicted similarity ratings of 2.75-5 (see Figure 4.1). This set of 200 test pairs was in the approximate range of an equivalency

cutoff. These sentence pairs were rated using human raters recruited from AMT. Each rater was randomly assigned sentence pairs, and each pair was rated 10 times. The rating was a binary selection of 1 ("Equivalent") or 0 ("Not Equivalent"). This human ratings equivalency study is referred to as Pilot 1b, and the complete details for the AMT and Zoho interfaces are provided as an Appendix in Sections B.1.1 and B.1.2, respectively. A screen capture of the sentence equivalency data collection user interface is shown in Figure 4.3. The mean equivalence rating of each pair was calculated. A mean value of .8 or higher was used to represent consensus that the pair was equivalent, and a consensus value was set to 1. Mean values of less than .8 represented not equivalent, and a consensus value was set to 0.

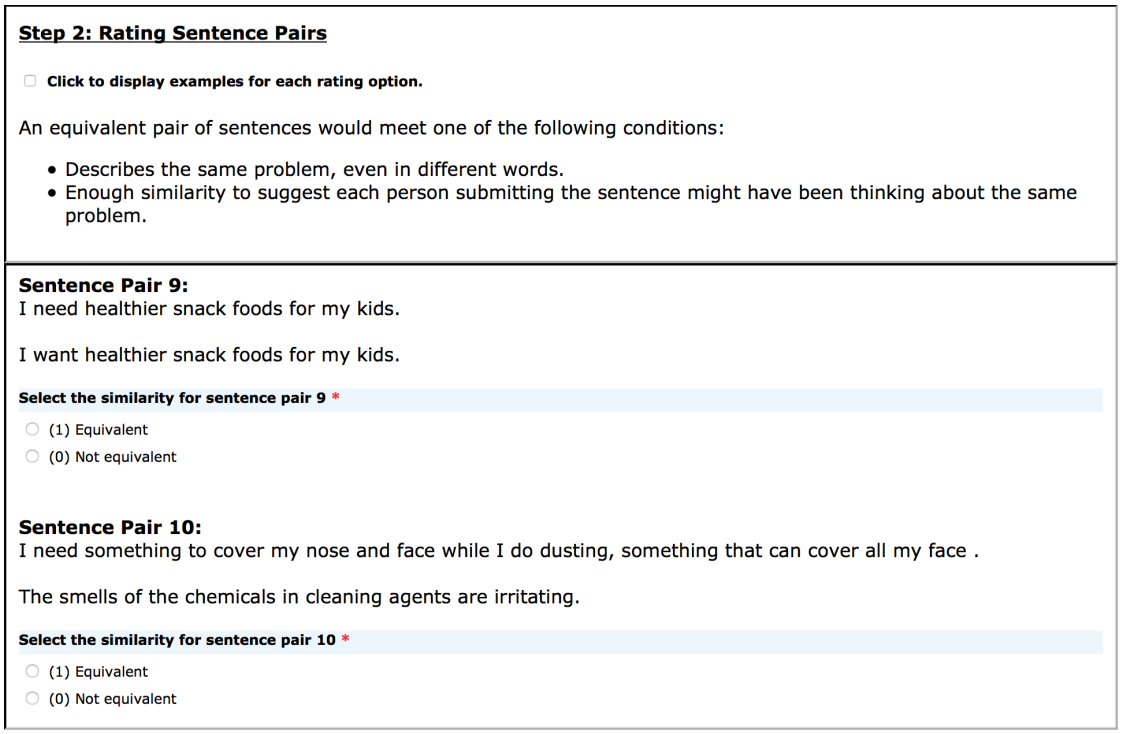

Figure 4.3: Screen Capture of Sentence Pair Equivalency Data Collection

The predicted similarity scores using TakeLab-simple trained with need statement data was compared to the binary consensus values (0 or 1) for the 200 test pairs. The accuracy of predicted ratings for a range of cutoff scores was plotted using the ROCR

package in R [121], and a cutoff representing equivalent meaning was selected based on best-case accuracy.

Additional cutoff scores were evaluated during later analyses to test the effect of filtering data based on criteria other than equivalency (e.g. mostly similar or somewhat similar).

4.1.4 Algorithm 1 Analysis: TakeLab-simple

The TakeLab source files included large (greater than 1GB) word corpus files (New York Times Annotated Corpus and Wikipedia Corpus) that had been previously filtered to include only the words present in the SemEval data sets. The complete corpus files were obtained and filtered for only the words present in the set of all training and analysis need statements. This step creates manageable file sizes and does not impact ratings.

The TakeLab-simple system was trained on each of the five need statement validation training sets. The corresponding five validation test sets were analyzed, producing an output value of the predicted similarity score on a scale of 0 to 5. A Pearson correlation value was used to compare predicted similarity scores to human ratings. Correlation values were calculated for the five test sets using both the original SemEval MSRvid model and each respective new model from need statement training data.

Following the cross validation, the TakeLab-simple system was trained using the complete training set of 1500 pairs. This final need statement training model was used for uniqueness study performed on the entire need statement analysis set collected during Part 1: Quantity Study 3.

4.1.5 Algorithm 2 Analysis: UMBC-PairingWords

The UMBC system was not tested as a local system; therefore, the system was trained using only (default) SemEval data. The Python API accesses the system via a URL consisting of embedded pairs of text passages. The output value is a predicted similarity score in the range of 0 to 1, and this output is scaled to match the SemEval range of 0 to 5. The UMBC-PairingWords system has an additional parameter to select one of three configuration types, denoted as type 0, 1, or 2. This value modifies the number of parameter combinations tested (1, 2, and 4, respectively) to determine maximum similarity.

The combinations include parameters such as “ignore common adverbs”, “use extended stopwords” and “support acronym” (Lushan Han, UMBC, personal communication).

The five test sets from the cross validation study were analyzed accessing the system via embedded URL parameters. All three configuration types were tested. A Pearson correlation value was used to compare predicted similarity scores to human ratings.

4.1.6 Uniqueness Study: Identifying Unique and Redundant Entries

The algorithm performance evaluation and the cutoff score evaluation (see Figure 4.1) determined the final algorithm and cutoff score used as inputs for the uniqueness study. The need statement analysis set was processed as described in Section 4.1.1. Resulting pairs were analyzed using the Takelab-simple system trained using needs data as described in Section 4.1.2. The final analysis used only the system with the highest Pearson correlation values as reported in Section 4.2.

The set of potential duplicate sentence pairs was assumed to be those with a similarity score above the cutoff as described in Section 4.1.3. Although the complete analysis of similarity included pairs where a single sentence was compared to itself, these pairs were omitted from the set of potential duplicates as these pairs did not represent a sentence submitted multiple times.

If a single sentence was included in the set of potential duplicates multiple times, the total count of pairs was recorded for each baseline sentence. Within a pair of sentences, the baseline was considered to be the sentence submitted first (e.g. with the lowest ID number).

The accuracy of predicted ratings above and below the cutoff score was evaluated using human ratings of equivalency as described in Section 4.1.3. A false negative rating would be a high human consensus of equivalence, but a low predicted similarity score. A false positive would be a low human consensus of equivalence, but a high predicted similarity score.

4.1.7 Analysis of Crowd Size Permutations

The analysis described in Section 4.1.6 includes needs submitted from all users in each topic group. A permutation analysis was used to determine how the relative quantity of

unique needs changes with increasing group size. In this analysis, random subsamples were repeated at varying group sizes to simulate sizes from small to large groups. Figure 4.4 shows a schematic representation of analyzing one permutation. Analyzing group size characteristics required input files with each sentence replaced with the sequential need ID. This was appended with the user ID of the participant submitting each need and the similarity score of the sentence pair.

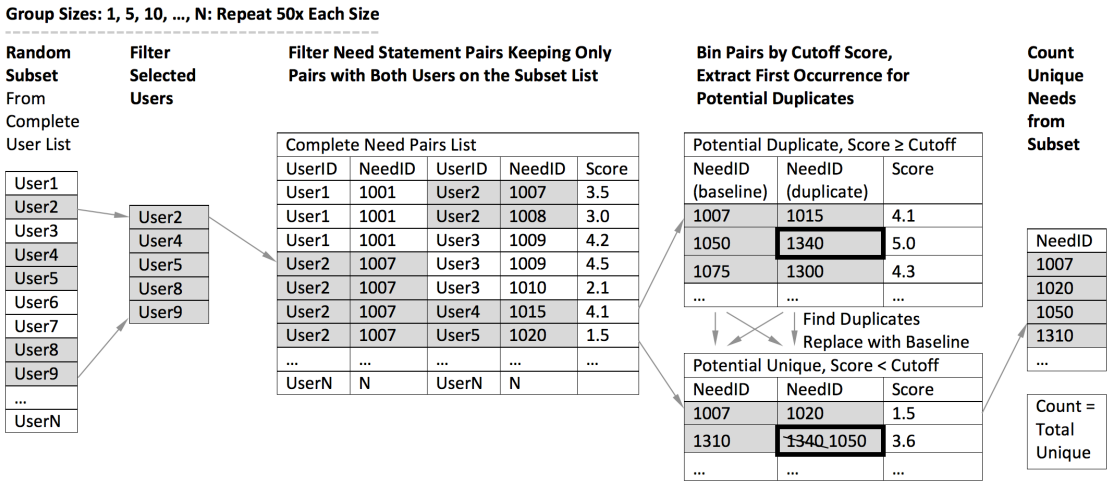

Figure 4.4: Schematic of Group Size Permutation Analysis (Represents a Single Iteration Using a Group Size of Five)

In the Figure 4.4 example, shaded cells represent data from users included in the single permutation, and non-shaded cells represent data from other users. Users 2,4,5,8, and 9 were randomly selected out of all users for a simulated group size of 5. The complete list of sentence pairs was filtered to only include pairs where both sentences were submitted by users in the permutation group. Pairs where a sentence was compared to itself (same ID) were omitted. This complete list was divided into potential duplicate pairs (score $\geq$ cutoff) and potential unique pairs (score $<$ cutoff).

The list of potential unique pairs was then compared to the list of potential duplicates and if a sentence ID was found in the potential unique list that was present in the potential duplicate list, the duplicate sentence ID was replaced with the baseline ID of the duplicate pair. As before, the baseline was the first sentence to be submitted out of any pair of potential duplicates.

After substituting for all potential duplicates, each resulting list of unique sentence pairs was assessed for the count of all unique sentence ID's. This value was calculated for all 50 permutations of a given group size, and the mean and standard error for each group size was calculated and plotted. In order to determine the effects of cutoff scores representing varying degrees of similarity, the permutation analysis was repeated for a range of cutoff scores from 1 to 4.

4.2 Part 2 Results (Uniqueness)

The data collection process generated 1,735 need statements for the analysis set. After dividing into 3 topics and generating pairwise combinations, the total of three sets of sentence pairs was 507,074. Table 4.1 shows a summary of the counts of need statements and pairwise combinations for each topic group.

Table 4.1: Summary of Need Statement Data Collection

Topic	Users	Need statements	Combination pairs
Cooking	104	568	161,596
Cleaning	121	650	211,575
Travel	116	517	133,903
Total	**341**	**1,735**	**507,074**

4.2.1 Performance of Algorithms

Table 4.2 lists Pearson correlation values for all STS systems. The TakeLab-simple system was tested for models trained using MSRvid data and each cross validation training set. The UMBC-PairingWords was tested using three configuration settings. The UMBC-PairingWords system was available only as trained with SemEval data. All comparisons were relative to human ratings of the test sets as described in Section 4.1.2.

The TakeLab-simple system trained using a need statement training set resulted in a mean Pearson correlation of .85 and the UMBC system trained using SemEval data resulted in a value of .83 for the type 2 configuration setting.

Table 4.2: Algorithm Performance Compared to Human Ratings

	Manually Trained		"Off-the-Shelf"		
Test Set	TakeLab (MSRvid)	TakeLab (Needs)	UMBC (type 0)	UMBC (type 1)	UMBC (type 2)
Set 1	.70	.85	.66	.66	.84
Set 2	.72	.89	.63	.63	.85
Set 3	.71	.84	.64	.64	.83
Set 4	.71	.86	.63	.63	.83
Set 5	.67	.82	.83	.83	.82
Mean	.70	**.85**	.68	.68	**.83**

4.2.2 Uniqueness Study: Summary of Potential Duplicates

Table 4.3 provides the number of sentences in each topic with a potential duplicate based on a cutoff score of 4.0. Table 4.4 includes example text of potential duplicate sentences with the highest similarity scores in each topic. Similarity scores relative to baseline sentences are provided for each. Text shown is after preprocessing as described in Section 4.1.1.

Table 4.3: Summary of Potential Duplicates at Cutoff = 4

Topic	Total Pairs, Score ≥ 4	Count of Baseline Needs	Max. Duplicates per Sentence
Cooking	7	7	1
Cleaning	33	21	6
Travel	6	6	1

Table 4.4: Highest Similarity Sentences and Scores

Score	Need Statements: Cooking, Cleaning, Travel (top to bottom)
Baseline	I need a better way to store lids for my pots and pans.
5.0	I need a way to store my pots and pans.
Baseline	I need a way to keep cool while cooking in the kitchen.
5.0	A way to keep cool in the kitchen while cooking
Baseline	A way to make food cook more evenly in the microwave
4.97	I wish there were a way to make food cook evenly in the microwave.
Baseline	I need a way to scrub the kitchen floor without getting on my hands and knees.
5.0	I need a way to scrub the floors without getting on my hands and knees
Baseline	My knees hurt when I am scrubbing the floor.
5.0	My knees hurt when I am scrubbing the floor.
Baseline	You can spray cleaners in your eyes
5.0	You can spray cleaners in your mouth
Baseline	Trying to figure our how long the trip will take.
4.88	Trying to figure out how long the trip will take.
Baseline	I need a way to lessen my anxiety when it comes to flying.
4.42	I need a way to lessen my anxiety on long drives.
Baseline	I need a way to find driving directions easier
4.26	I need an easier way to get driving directions for when we travel.

4.2.3 False Negatives and False Positives

The accuracy of predicted ratings over a range of cutoff scores is shown in Figure 4.5. The average accuracy plateaus at approximately .75 at a cutoff value of 4.0. A cutoff

value of 4.0 was therefore used to divide predicted values into lists of potential duplicates and potential unique statements. One potential source of inaccuracy is the demonstrated variability, or lack of consensus, in human ratings of equivalence. Figure 4.6 shows a wide band of pairs with an equivalence rating between .3 and .7. This represents the range where at least 3 out of 10 raters differed from the majority.

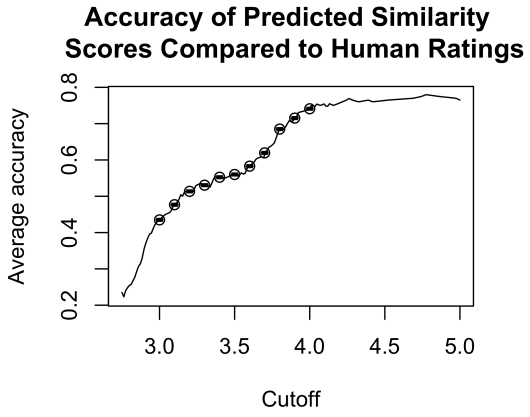

Figure 4.5: Accuracy of Predictions, Points Shown at Values for Cutoff = 3 - 4, by .1

Tables 4.5 and 4.6 show worst-case examples of inaccurate predicted ratings. A single pair from each topic was selected based on the highest discrepancy between predicted similarity scores and human-rated equivalence.

Table 4.5: False Positive Examples: Predicted Similarity is Too High

Similarity Score	Equivalent Rating	Need Statements: Cooking, Cleaning, Travel (top to bottom)
Baseline		I need a way to know if my rice is done.
4.7	0.1	I need a easy way to know if my steak is done.
Baseline		I need a easier way to clean the outside of my windows.
4.5	0.0	I need an easier way to clean my bathtub
Baseline		I need a way to guarantee that my luggage will arrive.
3.6	0.0	I need an easy way to pack my luggage.

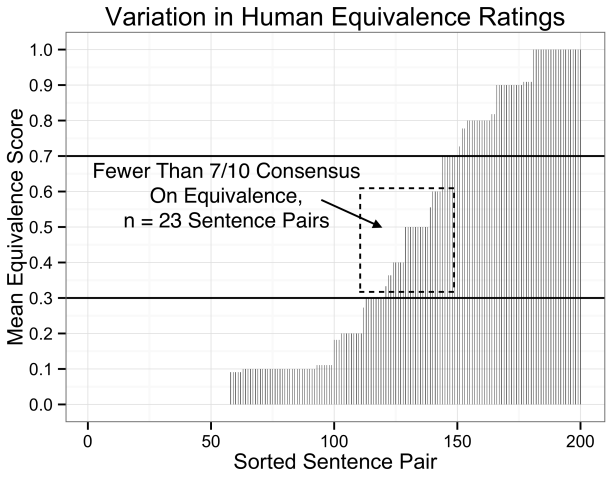

Figure 4.6: Sorted Mean Equivalence Scores

Table 4.6: False Negative Examples: Predicted Similarity is Too Low

Similarity Score	Equivalent Rating	Need Statements: Cooking, Cleaning, Travel (top to bottom)
Baseline		I need a way to lift a partially cut open can lid out of the can.
2.8	1.0	I need a better way to open cans.
Baseline		I need a good way to clean the top of a ceiling fan.
3.1	1.0	Ceiling fans are difficult to clean.
Baseline		I need a convenient way to make a checklist of things to pack
2.8	0.9	I need a better way to pack when traveling.

4.2.4 Uniqueness Study: Unique Statements and Crowd Size

Figure 4.7 shows plots of each topic for the mean quantity of unique need statements at each group size. Points represent the mean count of unique needs for 50 permutations of the group size. Curves for unique needs vs. group size are repeated for a range of cutoff values representing very little similarity (cutoff = 1) to equivalent (cutoff = 4). The results support hypothesis 1. Each plot demonstrates that the count of unique needs increases nearly linearly at high cutoff values, as there are few pairs above this cutoff and few substitutions are made due to duplicates. As the cutoff value decreases, the number of unique needs appears asymptotic for large crowd sizes. Figure 4.8 shows all topics plotted together using a cutoff score of 4, representing equivalent meaning of

sentence pairs. While the slopes of each curve for the 3 topics are similar, at a group size of 100, the count of needs in the travel group is approximately 80 lower than the count in Cleaning and Cooking groups. In Figs. 4.7 and 4.8, error bars are shown but are smaller than displayed data points.

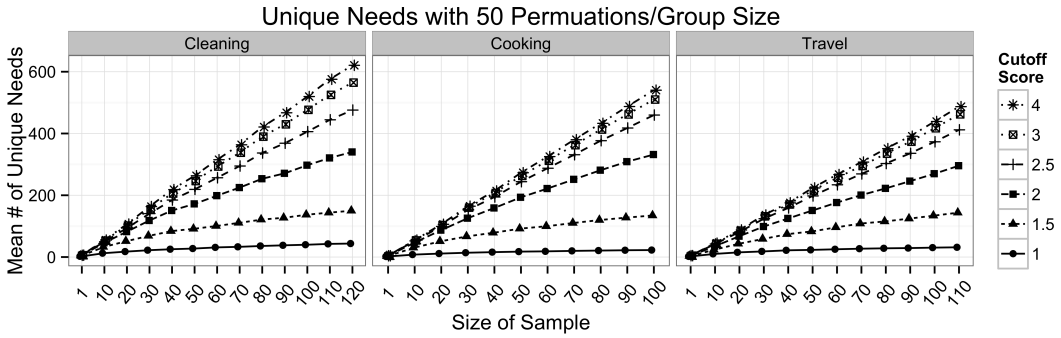

Figure 4.7: Quantities of Unique Statement with Increasing Group Sizes [Standard Error (SE) Bars Smaller Than Points]

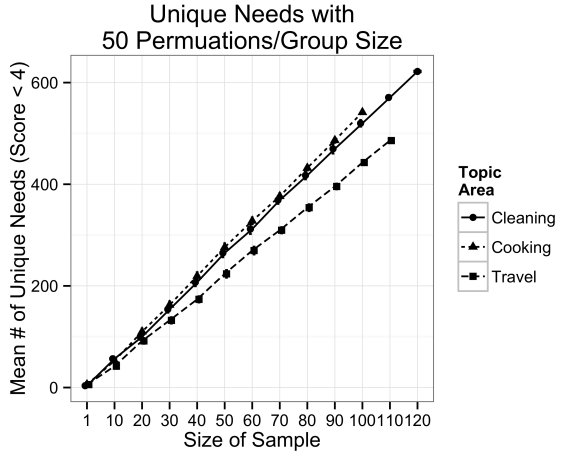

Figure 4.8: Quantities of Unique Statements at Cutoff Score = 4 [Standard Error (SE) Bars Smaller Than Points]

4.3 Part 2 Discussion (Uniqueness)

This study confirms NLP algorithms can potentially overcome the resource-intensive process of assessing open innovation data through automated screening of duplicates. Two state-of-the-art algorithms are presented and offer the ability to generate accurate predictions using both task-specific training data and generalized training data.

4.3.1 STS Can Detect Duplications and Uniqueness in Needs-Based Open Innovation Data

The results show the ability of STS algorithms to detect duplicate need statements among three independent lists of more than 500 sentences each. Table 4.4 shows examples of both exact duplicates as well as statements using similar language to convey equivalent meaning. Correlations were generally high (up to .85 when using an algorithm trained with need-statement data). The results support future work to apply STS methods to needs-based data as well as other common open innovation applications for potential solutions and ideas if submitted via text.

The specific ability to detect a lack of similarity (rather than detecting duplication) may also have wide application in open innovation data management and supports future work in this area. Previously, duplicate statements would be defined by having a STS score above a cutoff (typically a high cutoff, such as four on a scale of zero to five). However, testing for pairs below a very low cutoff would indicate a lack of similarity to each other. In resource-constrained situations, an organization may have the ability to evaluate a set maximum number of options. STS rankings create the potential to determine a cutoff score based on the required final count of statements with the lowest similarity to each other. For example, the same similarity scores could be analyzed to determine the lowest cutoff score resulting in 100 highly unique statements. If the cutoff to achieve this total was 1.0, this means for all 100 statements, none were rated with a score greater than 1.0 relative to any other statement in the entire set. These statements would potentially allow rapid exploration of the entire data set. A further application of STS may then seek the similar variations of any high-quality statement from this exploration set. Lowest similarity scores may be exploited to find tacit or latent needs or novel ideas–those that are rarely articulated.

4.3.2 Reduced Resources for Automated Methods

The approximate computation time for rating the complete set of need pairs for each topic was 1 day. The computation times for larger sets would be limited only by processing speed. While previous examples, such as the 3 month Cisco project described in Section 2.2.1, likely performed assessments beyond only duplication, automated methods would likely compare favorably for this specific step. For example, a common sorting method such as affinity mapping can group like entries and identify duplicates.

While human sorting of large numbers of need statements could potentially be conducted, e.g. recruiting manuals sorters via Amazon Mechanical Turk, automated methods may remain advantageous. Efficient human duplication detection may lack the same graded score or comparison possible when using algorithms ratings. Graded ratings allow for systematic reviews of highly unique statements as well as duplicates, as described in Section 4.3.1. Automated algorithms also offer flexibility to train on and analyze statements including jargon (e.g. clinical or medical terms). This flexibility may permit analysis of need statements from specialized users more easily than human raters from the general population and potentially less costly than recruiting specialized human sorters.

4.3.3 Potential for Future Increases in Accuracy

The accuracy of true positives in predicted scores decreased considerably when a cutoff score reflects similar, but not equivalent sentences. While this increased subtlety may be challenging for current technology, human raters also demonstrated poor consensus in this region, and the continuing attention dedicated to NLP research may soon reduce the gap between prediction and human ratings. In this study alone, comparing 2012 to 2013, the results demonstrated a significant performance increase. Our best results for the 2012 algorithm required specialized tuning to an open-innovation application. The 2013 algorithm achieved almost equivalent performance with no manual tuning to our application area. The accuracy of STS may be further improved with application-specific development. For example, in some instances the predicted scores indicated A=B, B=C, but A$\neq$C, indicating this logic structure is not accounted for.

In addition, while STS algorithms are appropriate for statements relating to common

consumer goods and services, similar NLP methods dedicated to specialized language, such as clinical vocabulary, are also enjoying research interest and the potential for ongoing improvements [122]

4.3.4 Evidence Against Fraud or Malicious Use

Table 4.4 includes some examples of exact duplication; however, exact duplicates are rare, and there is no evidence of widespread malicious submission (e.g. for increased pay) even when compared to sets of statements shown as the shared need stimulus (data not shown in results). One example in particular was traced to the same user submitting the same statement in rapid succession; however, after also submitting other unique needs. One potential explanation is inadvertently clicking the submit button multiple times.

4.3.5 Limitations and Future Work

As described as a limitation in Part 1, these topic were intentionally chosen to be broad enough to be relevant to a large pool of recruited participants. Rates of duplication may be dependent on the specificity of the topic. The three topics used in this study resulted in similar rates of unique entries over a wide range of group sizes. Initial expectations were to see first suggestions that are obvious and hence often similar; however, the quantity of submitted data and the degree of uniqueness exceeded expectations. A more narrow focus with similar group sizes will likely increase duplication; although this may be desirable as it indicates saturation in the data. Recruiting larger groups for broad topics is also possible; however, this may generate quantities of data that require larger computational resources if evaluated post data collection. However, future developments of real-time algorithm implementation would distribute computational analysis over a longer period. In this scenario, each new need would be immediately compared to all previously submitted needs (either all needs or filtered for only unique baseline entries) and processed at the time of submission. NLP algorithms are highly amenable to such use.

STS will only be applicable to submitted data in text form. This provides further

justification for use with text-based need descriptions; however, idea submissions incorporating a visual component, such as a sketch, would not be suitable. Also, assessing a large pool of open innovation data specifically for similar and unique subsets is only one component of managing the data. A similarity rating does not imply whether the submission is of high quality (e.g. important to users, representing a market opportunity) and should be pursued further; however, the rating may facilitate later steps by reducing redundant evaluations and simplifying initial explorations of the space. Evaluating the quality of ideas has been thoroughly studied [56]. Research relating to the quality of need statements is more limited [42] and is further motivated by the results of the Part 2 current work.

Only two algorithms were tested for this study. Candidates were chosen based on existing data suggesting superior performance over many other candidates; however, the specifics of comparing need or idea statements might result in outcomes where other algorithms produce better results. In addition, because the UMBC system source code was not publicly available, there was no way to retrain the system with new task-specific data. Therefore, UMBC does not use the same training data as the final TakeLab system, introducing an additional unknown in this comparison.

Chapter 5

Part 3: Assessing the Quality of Need Statements Submitted by Users

5.1 Part 3 Methods Overview (Quality)

The quality study described in Part 3 represented the culmination of the overall body of validation studies. After the completion of quality assessments, a wide range of hypothesis were tested. The hypotheses listed in Table 1.2 are reviewed and described with additional detail below.

(H2) Increasing the number of participants submitting needs increases the number of high-quality needs as judged by users.

(H3) Increasing the quantity of needs contributed per person increases the number of high-quality needs as judged by users.

These two hypotheses generally related to whether user research would benefit primarily from increasing the size of the user group or from applying more in-depth methods to help individuals articulate more needs or a combination of both.

(H4) Increasing levels of self-rated user expertise will not significantly increase the number of high-quality needs per person.

The fourth hypothesis can inform what characteristics of a group (in addition to group size) can improve the outcome of user research.

(H5) Needs submitted first would be less likely to be high quality than needs submitted after a sustained period of time.

When providing users with improved methods to articulate their own needs, the resulting output will include a list of need statements. It is possible that needs that come to mind first will represent overly general or superficial statements. These might be commonly duplicated and potentially lower quality than statements submitted after an opportunity for more prolonged consideration.

(H6) Semantically similar need statements would be rated as equivalent in quality.

Within a large group of users, several individuals might describe essentially the same underlying need. A valid quality metric should result in equivalent quality for similar wordings of semantically equivalent statements.

(H7) Need statements would be rated as higher quality if a detailed description of the need context was available.

Need statements submitted by users are typically one sentence long and are intended as a synopsis. Detailed contextual information was often provided as well. This detailed information may be of value to users who are rating the quality of need statements, and might change the perceived quality.

5.1.1 Need Statement Data

The need statement quality assessment analyzed data collected during Part 1: Quantity Study 3. The data consisted of sentence-length need statements and paragraph-length stories providing additional context and detail. Results from Part 2 (Uniqueness) were incorporated to exclude potential duplicate need statements from the quality assessment. Table 5.1 includes a breakdown of need statements for each topic area, the proportion submitted with an optional story, and potential duplicates removed from analysis. The quantity of need statements (500+ per topic) significantly exceeds most prior work as described in Section 2.3.

Table 5.1: Summary of Need Statement and Topics

Topic	Users	Need statements	Including Stories
Cooking	104	568	439
Cleaning	121	650	422
Travel	116	517	385
Original	341	1,735	1,246
STS Duplicates	N/A	-38	-30
Phase 1	341	1,697	1,216

5.1.2 Quality Rating Data Collection

All quality ratings were collected using a custom online survey interface built using Zoho Creator and recruiting participants from AMT. Each participant was randomly assigned to one of the same topics originally used for need collection. Participants would see instructions that read in part: "The ratings help prioritize which problems could be solved to help the most people." Each participant would then complete one page of optional demographics questions for age, gender, expertise (self rated), experience (self rated hours per week related to topic), and whether any user description was

applicable. Examples of user descriptions for cooking include: "family member with diet restrictions," "cook for small children." Multiple selections were allowed. Descriptions were created after reviewing problem statement data and were implemented to allow optional analysis of population segments.

Next, participants would read detailed instructions describing reasons to flag a statement ("the statement is already a solution not a need" or "the statement is unclear") and could review examples of statements appropriate for flagging. The participants then read details for the two quality criteria as described in Section 5.1.4.

Each participant was shown a random selection of 10 problem statements related to the assigned topic. If a statement included a full story, this was displayed under the statement. There were options to flag a statement and to rate the statement for Importance and Satisfaction. If the statement was flagged, the Importance and Satisfaction criteria were replaced with a question for the type of flag. Flagged statements were not rated for quality. Participants were paid \$0.50 for rating 10 statements. Repeat participants automatically bypassed demographics questions and proceeded to rate 10 new statements within the original topic.

One statement provided in the random selection was a trick question to check attention, and it read in part "Leave all questions for this statement blank to confirm you have read the full statement." If a participant did not leave these criteria blank, all 10 ratings in that set would be labeled as an attention "fail." These were omitted from analysis. The complete details of the AMT and Zoho interfaces for the quality data collection are provided as an Appendix in Sections D.1.1 and D.1.2, respectively.

5.1.3 Need Statement Quality Rating Phases

The quality ratings for need statements were collected in three sequential rounds of recruiting in order to efficiently use resources and minimize cost of rating low quality statements. The first phase began with the complete set as described in Table 5.1. Subsequent phases began with a modified set after preliminary analysis as shown in Table 5.2. All statements were initially rated by a minimum of 5 participants. These preliminary results were used to remove flagged statements and the lowest quartile of mean quality rankings. In phase 2, the remaining set was rated 10 additional times to reach a minimum of 15 ratings each. For phase 3, flagged statements were again

removed, and the top quartile proceeded for an additional 15 ratings to reach a total of 30 ratings per statement. Flagged statements were again removed after Phase 3, and before final analysis.

Table 5.2: Overview of Exclusion Criteria for Phases in Quality Study

	Ratings/Need	Exclusion[E] Criteria
Phase 1	5+ Ratings	E: N/A (All were included)
Phase 2	15+	E: Flagged 3+ Times after Phase 1
		E: Mean Rating in Bottom Quartile
Phase 3	30	E: Flagged 8+ Times after Phase 2
		E: Mean Rating in Bottom 3 Quartiles
Final	30	E: Flagged 16+ Times after Phase 3

5.1.4 Quality Metric

The two criteria in this study were: how important the problem was to the need statement rater, and how satisfied the rater was with existing solutions. Importance was rated from 1 ("Unimportant") to 5 ("Very Important"), and Satisfaction was rated from 1 ("No Solution or Very Unsatisfied") to 5 ("Very Satisfied"). Similar work by Ulwick does not indicate verbatim labels (anchors) for the scale.

The final quality rating was a linear combination of the two criteria scores as defined by Equation 5.1. The value of Satisfaction is inverted by subtraction from 6, essentially to mean that a high quality is a combination of a need with high Importance and high "Unsatisfaction". However, rating for Satisfaction was considered more common and less likely to create confusion.

$$Quality = Importance + (6 - Satisfaction) \tag{5.1}$$

5.1.5 Analysis Methods for Effects of User Characteristics

The effects of user group size on overall need quality (hypothesis 2) was evaluated using a permutation analysis for each topic. In this analysis, random subsamples were repeated at varying group sizes to simulate sizes from small to large groups. Figure 5.1 shows a schematic representation of analyzing one permutation. Each user and each need statement was replaced with its sequential ID number. A new matrix combined

the need ID with the user ID of the participant submitting each need and the quality score calculated from mean Importance and Satisfaction ratings.

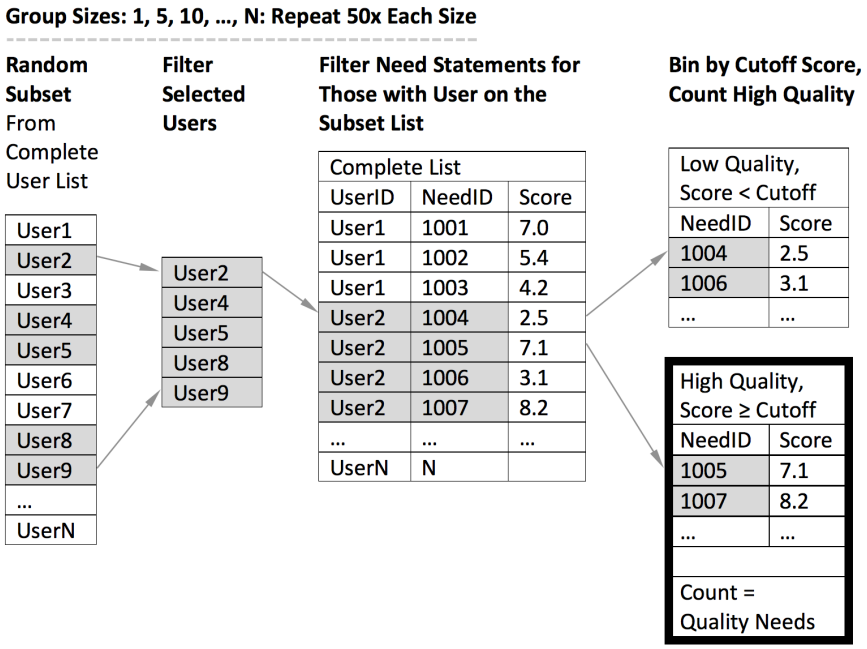

Figure 5.1: Process for Analysis of One Group Size Permutation

In Figure 5.1, shaded cells represent data included in the single permutation, and non-shaded cells represent those that were excluded. The total list of users for each topic was randomly sampled with sizes of 1, 5, 10,... n, where n equals the total users for each topic. Each group sample size was repeated for 50 different permutations. For each group permutation, the complete list of need statements was filtered to only include statements submitted by users in the permutation group. This complete list was divided into high and low quality bins based on range of a quality score cutoff values (e.g. scores approximating the top 1% or 5%). A count of quality statements was created for each group and cutoff, and mean count values and standard errors for 50 permutations were plotted.

In the Figure 5.1 example, users 2, 4, 5, 8, and 9 were randomly selected out of all users for a simulated group size of 5. Only the needs from these users were included and were binned based on the quality score cut-off (which varies for different analyses).

The high-quality bin was used for the count of quality needs for this permutation. After 50 repetitions at group size 5, the mean count (and standard error) for this group size was calculated.

For hypothesis 3, the high-quality needs per person were analyzed using a metric of the count of top quartile needs per person. This metric was used in favor of mean quality scores per person because a high count of quality needs would emphasize the objective of a needfinding process (e.g. if a participant submits 5 high-quality needs and 20 low quality, the mean might be equivalent to a different participant with 1 high quality and 4 low quality; however, the former case would be a more valuable outcome).

The same metric of top quartile needs was used to evaluate hypothesis 4, the effects of user demographics (submitter or rater). Count data was not a normal distribution and was therefore analyzed using a likelihood-ratio test to determine the best fit model comparing Poisson and negative binomial models. A regression analysis was used for the best fit model to test differences of groups (by default, models tested differences of log(means)). In addition, a multiple comparison test (multcomp R package using "Tukey" parameter [117]) was used on the generalized linear model to test pairwise combinations of user demographic groups. For each demographic included in the analysis, if the response was blank, the quality data was excluded.

Descriptive statistics were employed to visualize trends in the data, such as quality distributions.

5.1.6 Analysis Methods for Effects of Need Statement Characteristics

The effects of need statement sequence on overall need quality (hypothesis 5) was analyzed with descriptive statistics. Two metrics were used to represent quality for groups of need statements. The first metric was a median of quality ratings per group (represented by box plots) where the progression of groups would be all needs submitted first by users, all needs submitted second by users, etc. This has a benefit of capturing high sample sizes; however, the disadvantage is an undesired influence of low quality need statements. When assessing the value of an aggregated list of needs, the value of high quality needs would not be diminished regardless of the quantity of low quality entries. A second metric was used to emphasize this perspective and counted only those needs rated in the top quartile for quality. Ratios of counts of top quartile needs were plotted

for each group. Finally, descriptive trend lines were plotted using scatter plots.

Hypothesis 5 was also tested using two data sets, further described in Section 5.2. One data set represented the set of need statements from all users. This data set provides an overall trend. The advantage of the full user set is a progression throughout the entire range of needs (e.g. a need submitted first up to 25th by an individual). However, this set includes wide variation in group sizes, as few users submitted more than 10 needs. In addition, groups for the first and second need statements might include those needs from users only submitting one or two total entries. Comparing needs submitted first with those submitted second or third may still include very different groups of people. A second data set limited the analysis to the same individuals - those who submitted seven need statements. This was chosen as the highest number with at least 20 individuals. This data set provides a more narrow range of the sequence and ensures uniform sample sizes for each point in the sequence with the same individuals in the group. Each metric described above was applied to these two data sets.

For hypothesis 6, only need statement pairs previously identified as potential duplicates based on STS algorithms were analyzed. The difference in rated quality for each pair was calculated. The distribution of these differences was plotted. The similarity score of these pairs fell within the range of four to five (out of a total range of zero to five). A cutoff score of four was chosen based on previous analysis as described in Section 4.2.3. Pairs were created using the first submitted need, the "baseline", and each STS duplicate. A two-sided t-test was performed using quality scores for the paired data. The descriptive trend line was plotted for the STS similarity scores and corresponding differences in rated quality.

Hypothesis 7 was tested using a random sample of need statements from the total set. A sample of 45 need statements (all including a detailed story for context) was generated using 15 statements per topic. Each need statement was duplicated exactly; however, the detailed story was omitted. The story/no story pairs were included in the random sequence of all need statements to be rated for quality. A two-sided t-test was performed using quality scores for the paired data. The distribution of differences in quality scores for statement pairs was plotted.

Descriptive statistics were employed to visualize additional, secondary analyses, such as effects of the type of stimulus viewed prior to submitting a need statement and effects

of need statement originality based on STS similarity scores. As shown in Figure 3.4, the user interface included three buttons for three types of stimulus information. If a user selected "Help A" to view a collection of images, and then submitted a need statement, this need statement was tagged as following an image stimulus. The ratio of top quartile needs to total needs for each type was plotted. A statistical analysis was not performed for the effect of stimulus type because stimulus types were not randomly assigned for this data.

A metric of originality can be calculated based on STS scores. This assumes if a need statement has few other needs scored as similar to itself, it might be considered original. If a need statement has a high number of needs scored as similar to itself, it might be considered non-original. The STS algorithm was used to score the similarity of all pairwise combinations of all needs. A cutoff score of 2.5 was used to indicated similar meaning. For each baseline need (submitted first), the count of pairs including this statement was calculated and plotted with corresponding quality ratings.

5.2 Part 3 Results (Quality)

The data collection process generated a total of 25,837 ratings across the three phases for the total set of 1697 need statements. Table 5.3 includes the initial counts of need statements used for each phase and the counts of those need statements excluded prior to the start of the following phase. The final phase (phase 3) included 289 need statements and included a minimum of 30 ratings per statement before exclusions. Table 5.4 shows a summary of the counts of ratings collected for all phases and the number of individual ratings excluded because of flags or the participant failed the attention question as described in Section 5.1.2

Table 5.3: Summary of Need Statement Data Sets

Criteria	Phase 1	Phase 2	Phase 3
Rated Statements	1,697	1,168	289
E: Flagged*	-66	-5	0
E: Bottom Quartile(s)*	-463	-874	N/A
After Exclusions	1,168	289	289

* (E) represents Exclusions

Table 5.4: Summary of Need Statement Quality Ratings

	Phase 1	Phase 2	Phase 3	Total
Ratings Submitted	9,739	11,854	4,244	25,837
E: Failed Attention Question* †	-658	-859	-340	-1,857
E: Marked as Flag*	-940	-925	-273	-2,138
Ratings Analyzed	8,141	10,070	3,631	21,842

†All 10 survey ratings were omitted for a failed attention question
* (E) represents Exclusions

Flagged ratings were excluded from analysis even if the number of flags for a particular need statement was not high enough to exclude the need statement. For example, zero need statements were excluded due to 15+ flags after phase 3; however, 273 flags were submitted in this phase distributed among the included need statements. After exclusions, 21,842 ratings were analyzed. The combined data collection duration of all three survey phases was approximately 6 days. The target sample size for the final phase was 30 ratings per need. After removing flags and attention fails, the actual median count of ratings was 26 per statement.

Table 5.5: Summary of Need Statement Data Sets for Hypotheses 5-7

Analysis Description	Need Statements	Quality Ratings
H5, Raw Data	*1,697*	*25,837*
H5, All Users Analysis (Fig. 5.9)	1,626	21,717
H5, All Users Top Quartile Analysis (Fig. 5.10)	406	8,492
H5, Seven Needs per User Analysis (Fig. 5.11)	144	1,867
H5, Seven Needs per User Top Quartile Analysis (Fig. 5.12)	37	780
H6, STS Pairs, Raw Data	*66*	*1,146*
H6, STS Pairs Analysis (Figs. 5.13 - 5.14)	64	992
H7, Story/No Story Pairs, Raw Data	*90*	*2,759*
H7, Story/No Story Pairs Analysis (Fig. 5.15)	84	2,181
Stimulus Type Analysis (Fig. 5.16)	*1,626*	*21,717*
Statement Uniqueness Analysis (Fig. 5.17)	*191*	*2,674*

Hypotheses 5-7 included only a portion of the total data collected. Table 5.5 includes the initial counts of need statements and quality ratings used for each analysis. Differences between the complete analysis set and the raw data set are due to exclusions because of flags or the participant failed the attention question. All analysis sets in Table 5.5 are after exclusions.

5.2.1 Need Quality Distribution

Figure 5.2 shows the stacked distribution of mean quality ratings for all need statements (aggregated topics) included in each phase. The quality equation is described in Section 5.1.4. Descriptively, the distribution appears approximately normal and subsets of need statements used in different phases maintain general groupings for bottom, mid, and top quartiles. The Phase designation represents the final phase, for example, Phase 2 needs include those selected from Phase 1 to continue but were then excluded from Phase 3.

5.2.2 Need Quality for Varying Group Sizes

The results for the group size permutation analysis support hypothesis 2 for each topic and are shown in Figure 5.3. The entire population (e.g. all segments) is included. Curves for high-quality needs vs. group size are repeated for a range of cutoff values representing varying degrees of quality (7, 7.25, 7.5, and 8). Each point represents the

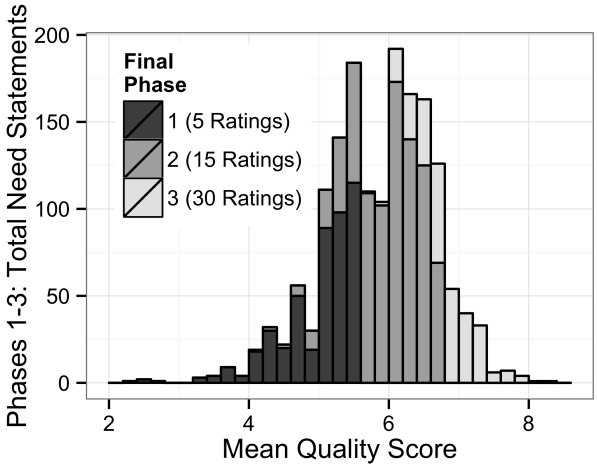

Figure 5.2: Stacked Distribution of Quality Scores (All Phases)

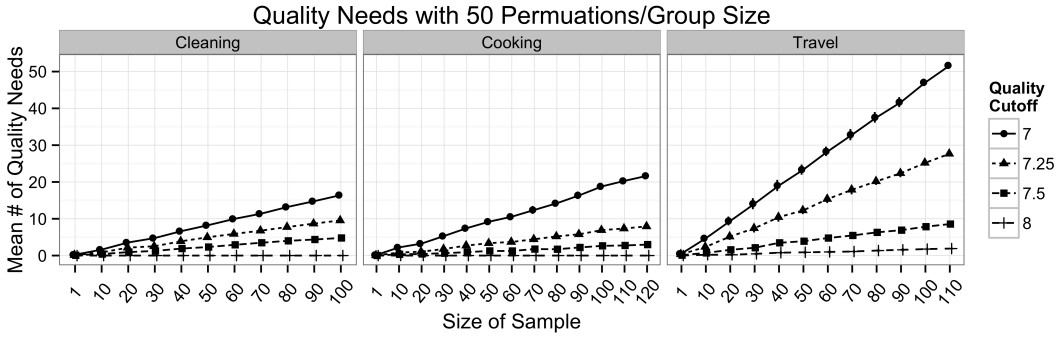

Figure 5.3: High-Quality Needs for Increasing Group Sizes [Error Bars Indicate Standard Errors]

mean of 50 random subsamples as described in Section 5.1.5. Error bars are shown, but are occasionally smaller than the data point. The plots using cutoff values less than 8 demonstrate a nearly linear relationship, where the number of high-quality needs increases with group size. Only the travel topic included mean ratings greater than 8.

Figure 5.4 shows all topics plotted together using a cutoff score of 7.5, representing a cutoff where the maximum for each topic is ten or fewer. Plots display a similar linear nature for each topic; however, slopes vary and the group size required to attain

a certain count of high-quality needs may vary approximately by a factor of 3 depending on topic.

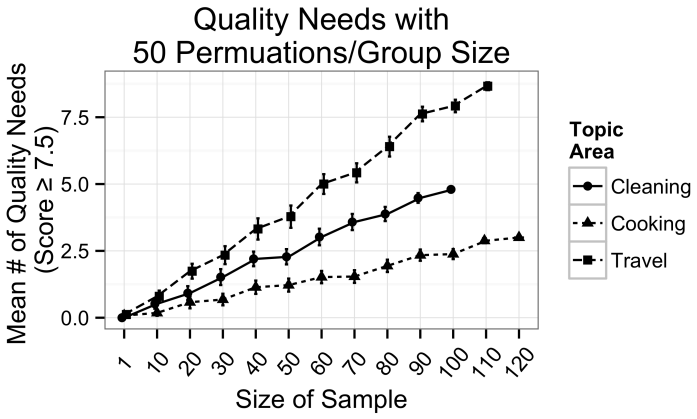

Figure 5.4: High-Quality Needs (Cutoff Score = 7.5) for All Topics and Group Sizes [Error Bars Indicate Standard Errors]

5.2.3 High Quality and High Quantity

For each participant submitting needs, the total number of needs submitted was compared to the count of top quartile needs (hypothesis 3). The trend of greater top quartile needs with increasing total counts is shown in Figure 5.5. The data represents integer values; however, overlapping points are offset for clarity.

5.2.4 High Quality and User Expertise

Figures 5.6-5.7 descriptively represent the effects of user demographics on need quality (hypothesis 4). Figure 5.6 summarizes the number of top quartile need statements submitted by users in each self-rated expertise group. Figure 5.7 summarizes the number of top quartile need statements submitted by users in each experience group (self-rated hours per week spent on a given topic). The data excluded due to blank demographics questions was less than 3% for both expertise and experience.

A Poisson regression analysis was used based on likelihood-ratio test results and goodness of fit tests. The model was preferred because the additional parameter of

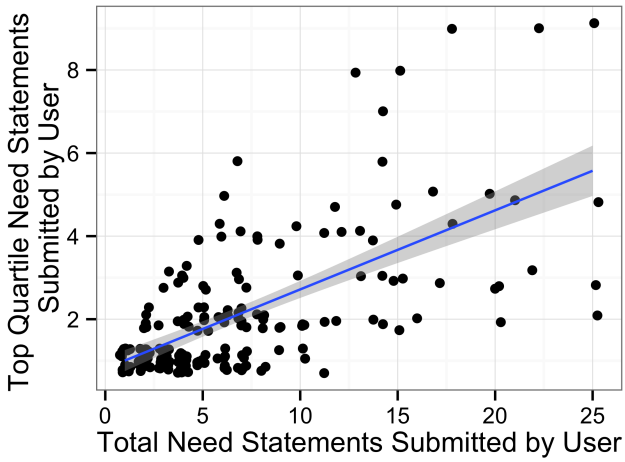

Figure 5.5: Top Quartile Needs for Users with Increasing Total Need Counts [Shading Indicates 95% CI]

the negative binomial model did not improve fit. The analysis tested differences of log(means) for expertise and experience. The relative contributions of level of expertise (self-rated), experience (self-rated hours per week), and topic area were tested with likelihood-ratio tests for Poisson regression models. The topic was a significant factor (p-value = 0.012). The level of expertise was not a significant factor (significant at p <.05). The level of experience (hours per week) was a significant factor (p-value = 0.032). While experience and topic were included in the final regression model, there were no individual pairwise comparisons for experience with a statistically significant difference (lowest p-value was 0.056 for No Hours:Up to 5 Hours).

5.2.5 Need Rater and Need Submitter Experience

The demographics of participants was recorded for both the need statement submitter and raters. Need statements were randomly assigned to raters, so random variation resulted in needs submitted by novice users rated by experts and vice versa. As a variation for hypothesis 4, the difference in user experience (hours per week) was calculated subtracting the experience group number of the need rater from the group number of the need submitter, e.g. a -3 would represent a need submitted by a lowest-experience user

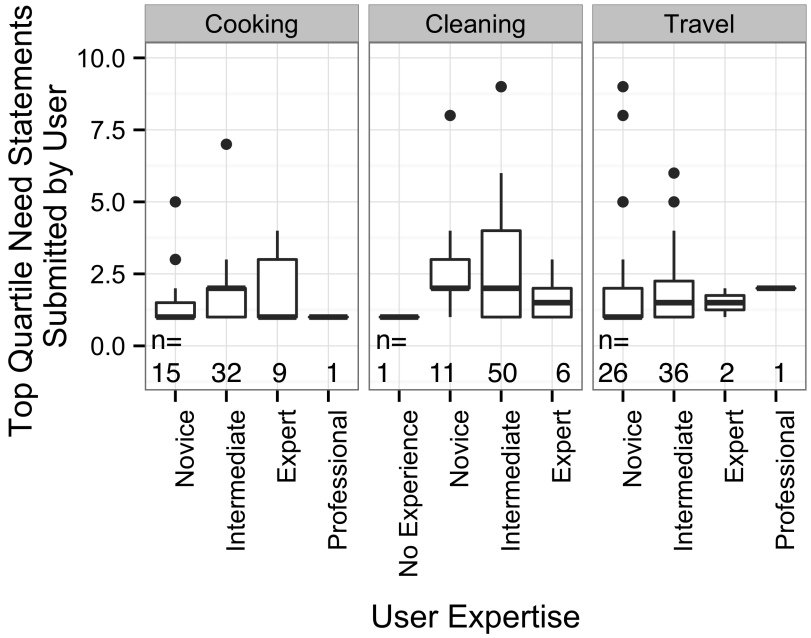

Figure 5.6: Top Quartile Needs for All Topics and Expertise Groups (Group Size, n, Shown)

(group 1) and rated by a most-experienced user (group 4). Figure 5.8 shows the mean quality score for each level of submitter-rater difference for experience groups. There is no trend indicating the degree of similarity of submitter and rater demographics (e.g. experience) affects the quality rating.

5.2.6 Highest-Rated Need Statements

Top-rated need statements, both overall and for a selection of segments are listed in Table 5.6 for a single representative topic of cleaning. The top-rated overall need includes ratings from all users. Ratings for population segments include only those raters identifying with the user description shown as described in Section 5.1.2 (e.g. a user in the cleaning group who is a "pet owner"). These top rated need statements paired with initial quality screening data would represent the output of the method in practice. Rating counts per need statement among segments varied widely; therefore, only top statements with at least 15 ratings for a segment are shown.

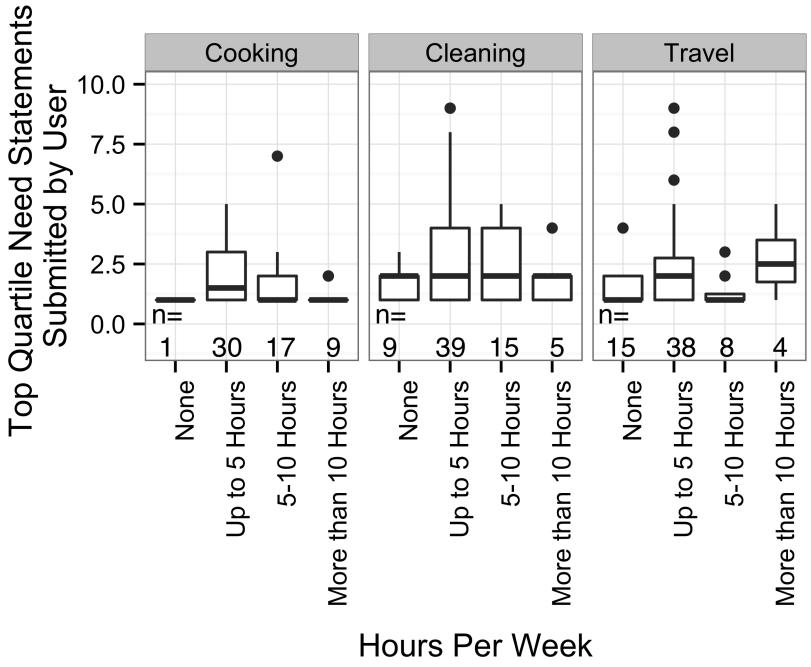

Figure 5.7: Top Quartile Needs for All Topics and Experience Groups (Group Size, n, Shown)

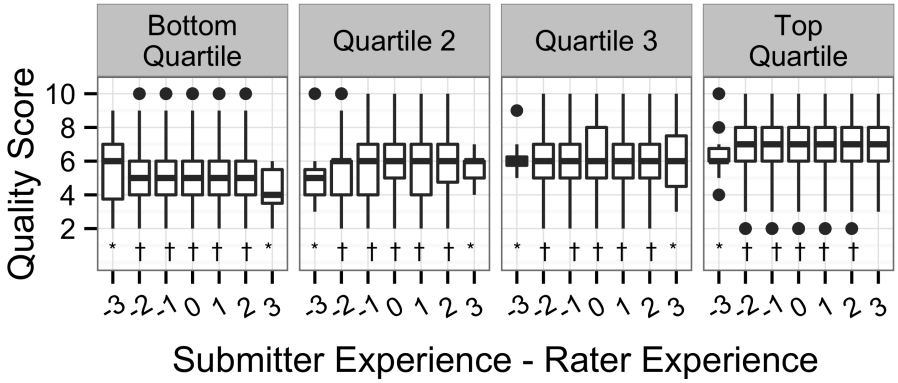

Figure 5.8: Mean Ratings for Differences in Submitter and Rater Experience [Negative Difference: Need from Low-Experience User Rated by High-Experience User], Group Sizes: * for n <20; †for n >100

Table 5.6: Examples of Highest Rated Need Statements from Overall Population and Selected Populations Segments

Topic	Need Statement	Importance	Satisfaction	Quality
Cleaning: Overall	Dirt and grime build up on my computer keyboard. Story: I have tried several different options to clean my keyboard but I cannot get down in there. It is easy to clean the tops of the keys but there is a lot that gets down in there that cannot be reached. I'm looking at it right now.	4.00	2.23	**7.77**
Cleaning: Pet Owner	The vacuum isn't strong enough to get pet hair completely out of the carpet	3.94	2.41	**7.53**
Cleaning: Wood Floors	I never feel sure that I got ALL the shards of broken glass. Story: If I drop a clear piece of glassware it's going to shatter and scatter, and of course the pieces are going to be nearly impossible to see. I always clean from a very wide area just because I can't trust that the little splinters will be visible, or that they will get picked up.	3.88	2.06	**7.81**

Table 5.7 provides a comparison of need statements rated as highest, potentially including any topic area. Examples show the highest overall quality score, highest Importance (only), and lowest Satisfaction (indicating a high value of 6-Satisfaction).

Table 5.8 provides a comparison of need statements rated as low, potentially including any topic area. Examples show the lowest overall quality score, lowest Importance (only), and highest Satisfaction (indicating a low value of 6-Satisfaction). Because lowest quality ratings were excluded from phases 2 and 3 as described in Section 5.1.3, most lowest scores have 5 or fewer ratings. Table 5.8 shows only statements with at least 10 ratings each in order to show examples with low scores and a greater sample size.

5.2.7 Need Quality and Sequence

The analysis of need statement quality from the first need a user submits to the last need a user submits addresses hypothesis 5. Hypothesis 5 was not confirmed. The best

Table 5.7: Examples of Highest Rated Need Statements Overall and Individual Metrics

Topic	Need Statement	Importance	Satisfaction	Quality
Highest: Overall	What if you are late for one of your flights / trains? Story: I ran late at a meeting in DC once, and the cab didn't get me to the airport in time. I was supposed to meet someone in New Orleans, but had to take a later plane, and had no way to let them know. Had another issue where the plane had mechanical problems but had our luggage, and we got where we were going, but the luggage didn't.	4.0	1.77	**8.23**
Lowest: Satisfaction	It would be nice to be able to bring drinks larger than 3oz on flights that were purchased outside the airport.	2.74	**1.63**	7.11
Highest: Importance	I need a way to reserve a place to stay at my destination. Story: If I intend to stay overnight at my destination, I'll need a place to sleep. It would be nice to have a way to reserve my room ahead of time.	**4.43**	4.14	6.28

Table 5.8: Examples of Low Rated Need Statements with 10 or More Ratings

Topic	Need Statement	Importance	Satisfaction	Quality
Lowest: Overall	Have to bend over to use a dustpan	2.36	4.18	**4.18**
Highest: Satisfaction	It would be nice to be able to find things to do in the places I travel to.	3.71	**4.21**	5.5
Lowest: Importance	I need to find a hotel that is pet friendly so I can take them with me.	**1.42**	1.67	5.75

fit lines for quality score over the sequence range are approximately horizontal for both the complete user group and the group of users with 7 needs. While the best fit lines for top quartile needs trend down in Fig. 5.10 and trend up in Fig. 5.12, the confidence intervals in both cases are larger than the deviation from horizontal.

Each analysis used differing sets of quality ratings as shown in Table 5.5. Figures 5.9 - 5.10 include results for all users combined. The need statement number represents

the sequence of the need statement per individual. For a need statement number of 5, the data point includes only need statements submitted fifth by the included group of users. Figures 5.11 - 5.12 show the same analysis including only the approximately 20 users who submitted 7 needs. Figures 5.9(a) and 5.11(a) show median values of quality scores over the respective sequence ranges. Figures 5.9(b) and 5.11(b) show scatter plots with linear best fit lines and 95% confidence intervals. Figures 5.10 and 5.12 represent the number of top quartile need statements for the same sequence ranges. Values are normalized to account for different total quantities in each group. For example, as shown in Figure 5.10, there were 324 needs submitted first and of these 83 were in the top quartile for rated quality. The quantity of top quartile needs per 100 would be 83/(324/100) or 83/3.24, giving approximately 25 top quartile needs per 100.

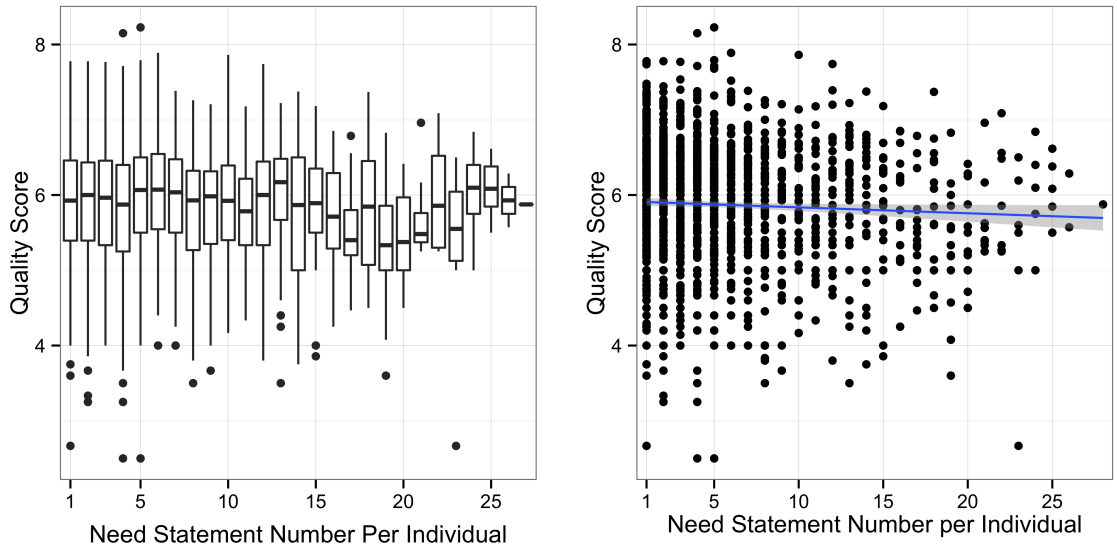

Figure 5.9: (a) Left, (b) Right: Quality of Need Statements for the Sequence of Needs per User [Shading Indicates 95% CI, All Users]

5.2.8 Quality of Duplicate Statements

Results support Hypothesis 6. The paired t-test results showed no significant difference between quality scores of duplicate pairs (p-value = 0.37). The difference in quality

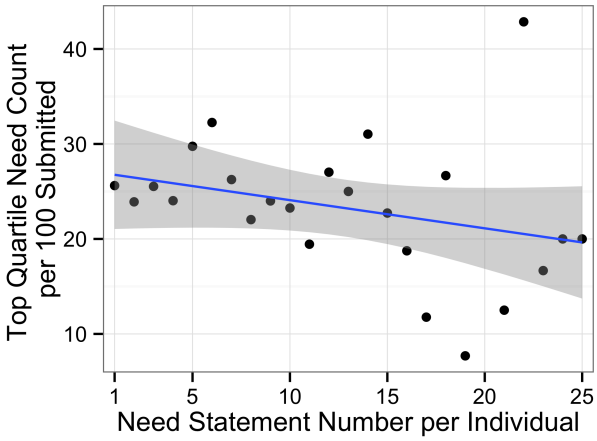

Figure 5.10: Count of Top Quartile Need Statements for the Sequence of Needs per User [Shading Indicates 95% CI]

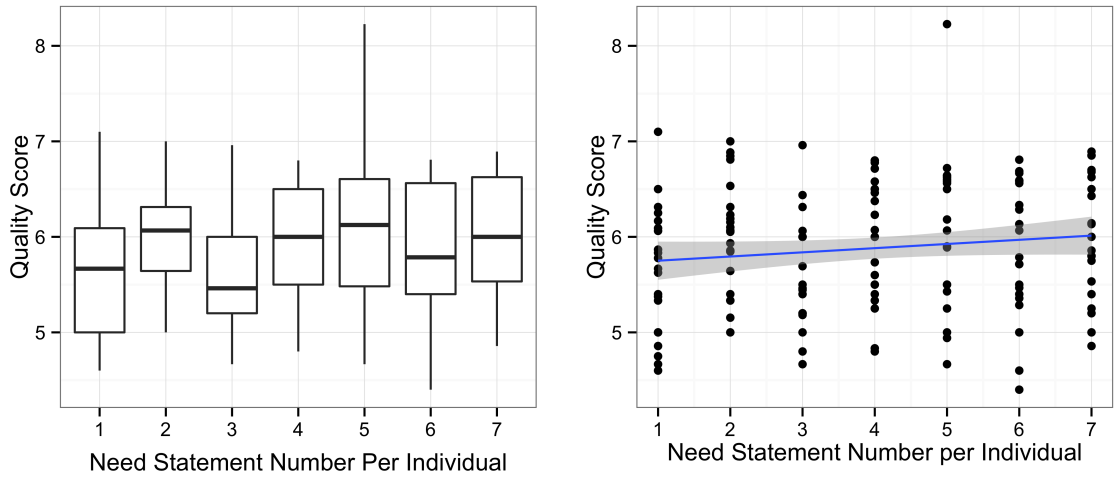

Figure 5.11: (a) Left, (b) Right: Quality of Need Statements for the Sequence of Needs per User [Shading Indicates 95% CI, Only Users Submitting 7 Needs]

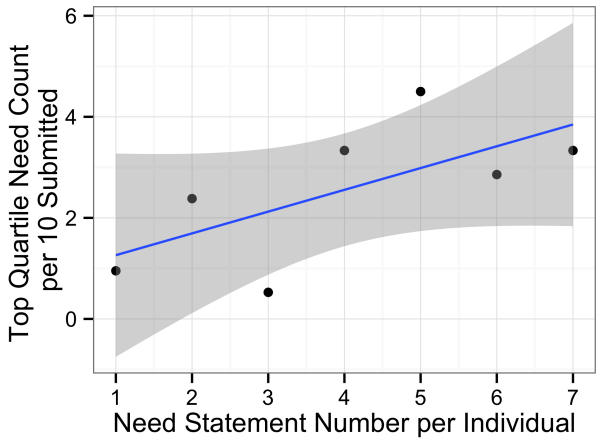

Figure 5.12: Count of Top Quartile Need Statements for the Sequence of Needs per User [Shading Indicates 95% CI, Only Users Submitting 7 Needs]

score was calculated for each pair. Figure 5.13 shows the distribution of differences in quality scores.

As shown in Table 4.3, there were 46 pairs of statements scored by STS algorithms as potentially duplicate. These 46 pairs included 66 total unique need statements given that in some cases multiple pairs included duplicates of the same baseline statements. After exclusions for flags and attention question failures, 37 pairs with 64 unique statements were analyzed. Paired data included the quality score of the baseline statement and also the quality score of the STS duplicate statement.

The range of algorithm scores represents a range in the degree of similarity (e.g. 4 might represent different statements with equivalent meanings, and 5 might represent nearly exact duplicate statements). The difference in quality scores might be dependent on the degree of similarity. Figure 5.14 shows each need statement pair with the STS score and the corresponding absolute value of the difference in quality scores. The best fit line trends slightly downward; however, the confidence intervals are larger than the deviation from horizontal.

Table 5.9 provides examples of representative points on the axis extremes in Fig. 5.14. One example combines the lowest STS score (4.0) with greatest difference in quality score (2.4) The second example combines the highest STS score (5.0) with the

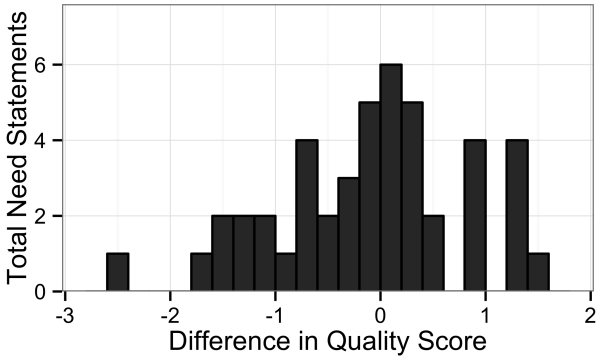

Figure 5.13: Distribution of Variation in Quality for STS Duplicates

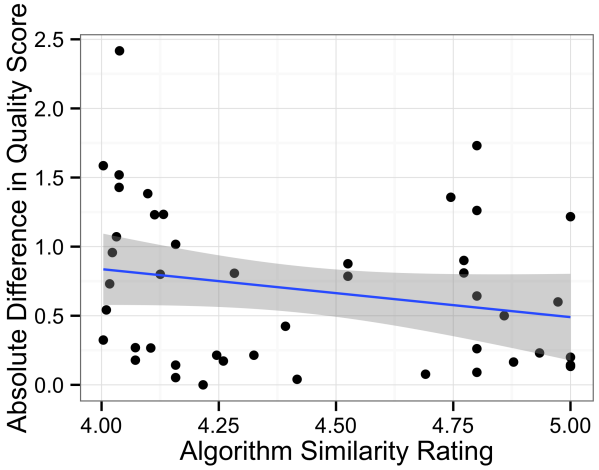

Figure 5.14: Different in Quality for Duplicate Needs for Varying Similarity [Shading Indicates 95% CI]

lowest difference in quality score (0.1). Full text need statements are shown for each example.

Table 5.9: Examples of Quality Ratings for STS Duplicate Statements

Baseline	Duplicate Statement	Similarity Score	Difference in Quality
I need an easier way to clean up after my dog.	I need an easier way to clean up pet stains.	4.0	2.4
I need a way to scrub the kitchen floor without getting on my hands and knees.	I need a way to scrub the floors without getting on my hands and knees.	5.0	0.1

5.2.9 Need Statements With and Without Detailed Stories

The sample of 45 need statement pairs (one with a detailed story and one without) was reduced to 42 pairs or 84 need statements after exclusions for flags and failing the attention question. Paired data included the quality score of each statement with and without the original detailed story. Results do not support hypothesis 7. The paired t-test results showed no significant difference between quality scores of duplicate pairs (p-value = 0.33). The difference in quality score was calculated for each pair. Figure 5.15 shows the distribution of differences in quality scores.

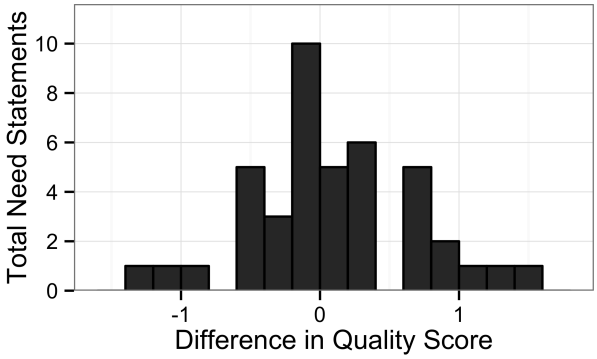

Figure 5.15: Distribution of Variation in Quality for Omitted-Story Duplicates

5.2.10 Quality of Need Statements after Viewing a Stimulus

One secondary research question (see Table 1.2) related to whether viewing a specific type of stimulus would affect the quality of later need submissions. Figure 5.16 shows an analysis evaluating the top quartile needs for each stimulus type. Plotted bars represent a ratio. For example, in the cooking topic there were 225 total needs submitted prior to any stimulus ("None") and of these 29 were in the top quartile for a ratio of 0.13. Quartiles were calculated cumulatively for all topics combined. Patterns for effects of stimulus type are not consistent across the three topics, (e.g. Cleaning shows little change for different types, and Travel shows the Images ratio as less than 50% of others.

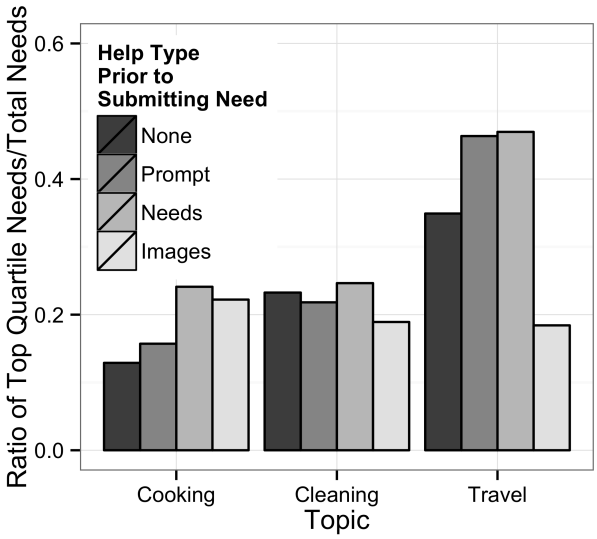

Figure 5.16: Quality of Need Statement for Users Viewing Different Stimulus Types

5.2.11 Need Statement Quality and Statement Uniqueness

An additional research question (see Table 1.2) related to whether the degree of uniqueness or lack of similarity to other needs would affect rated quality. Figure 5.17 shows the quality score of each baseline need and the corresponding counts of similar statements. For example, if the number of similar need statements was 20, this baseline need was included in 20 pairs of similar statements. In other words, this need statement was not

likely original because there were 20 other statements rated as similar. This analysis used a cutoff score different from a previous cutoff of four. The cutoff of four represents equivalent meaning, and the sample size was low. A cutoff of 2.5 was used in this analysis to represent similar meaning in order to increase the number of pairs; however, the trend for a cutoff of four was similar (results not shown). Descriptively, the best-fit linear trend line does not indicate the number of similar need statements is correlated to a change in quality.

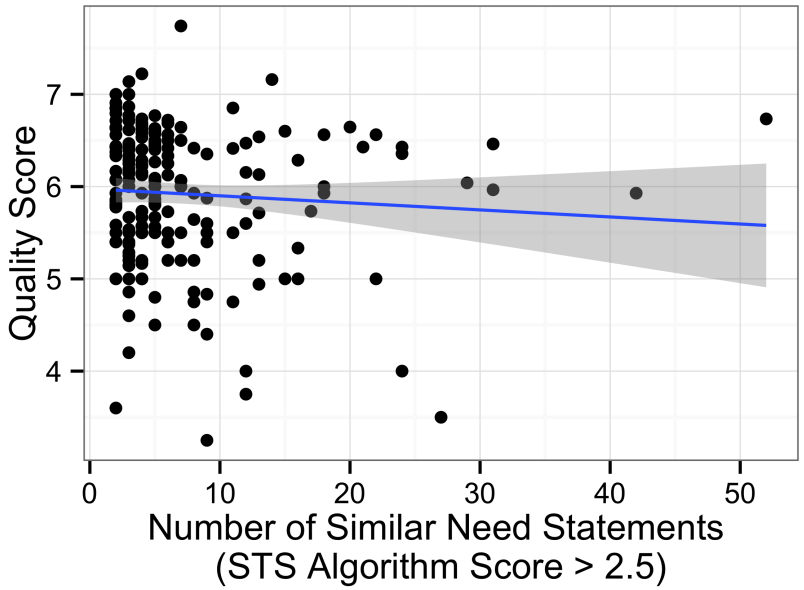

Figure 5.17: Quality of Need Statements and Algorithmically-Rated Uniqueness [Shading Indicates 95% CI]

5.3 Part 3 Discussion (Quality)

The overall goal of this study is to demonstrate a rapid quality rating method for needs and evaluate effective user group characteristics. The results showed the quality rating method can serve as an initial prioritization mechanism for lists of over 500 need statements per topic. The analysis of effects of user and group characteristics provided several important observations to inform large scale needfinding.

5.3.1 Higher Need Statement Quantity Leads to Higher Quality

The results demonstrate a correlation between high need quantity and high need quality for both groups (hypothesis 2) and individuals (hypothesis 3). Figure 5.3 shows that the number of high-quality needs increases with the size of the group contributing need statements. The higher total counts of need statements from larger groups has previously been shown to include increasing counts of unique needs (See Part 2). The increase in unique statements also results in a higher quantity of top-rated needs. Figure 5.5 presents an increasing trend of individual need counts and high quality. The results together indicate a benefit both for increasing individual quantity through relevant stimuli during need collection (see Part 1: Quantity Study 2) and also through recruiting large, diverse groups. While current results were limited to diversity of expertise and experience, the analysis of gender and age demographics (not shown) also demonstrated no significant association with need quality. The consistency of these results suggest that diversity of other demographics (e.g. ethnic or socioeconomic) may also be valuable to capture greater portions of a complete need space.

Results are consistent with previous studies finding an increase in need quantity as group size increases. While previous studies have shown an asymptotic curve with diminishing returns for groups larger than 30 [28], the specificity of the topic might influence this outcome. To our knowledge, no previous results confirm a correlation for quantity and quality of needs. The same correlation has been shown in the analogous process of concept ideation, both for cumulative group quantity [58, 59, 60] and also individuals within a group, where the relationship was similarly linear [55].

While there was an effect for topic area on the number of high-quality needs submitted, results from Part 1 do not reflect this effect for total quantity. This suggests that

the group size required to collect a given total number of needs may be consistent across differing topics, but the final number of users needed to capture a specified number of high-quality needs could vary by topic.

5.3.2 Expertise Does Not Predict User-Rated Quality

While current practice often emphasizes input from experts, in particular, during development of specialized products for health care users [47], these results support a contrary hypothesis (No. 4) that increasing levels of self-rated user expertise will not significantly increase the number of high-quality needs per person. This is consistent with results in Part 2 (Uniqueness) where experts do not submit a higher quantity of total needs. This is not confirmation that experts generally will be equivalent to non-experts, as the data does not reflect a binary classification. Rather there does not appear to be a trend of increasing quality with increasing self-rated expertise. While this study did not characterize quality for a specialized topic, the consistent results across the three topics used supports the further study of need statements for specialized topics (e.g. health care) collected from specialized users. In addition the results suggest inclusion of all levels of experience regardless of topic.

The results do not suggest that similarity of submitter and rater experience will affect perceived quality. In other words, experienced and inexperienced users are equally likely to submit a need that is rated as high quality by raters from the same experience group or a different group.

5.3.3 High-Volume Quality Rating is Feasible

The results support the use of simple quality metrics to provide an initial prioritization for large groups of need statements. Additional user feedback and analysis may provide additional insight when considering a subset of needs relevant to a specific project. When resources permit, a single phase with a high count of ratings for all statements will simplify analysis; however, staggered phases are feasible for constrained resources. While not used here, preliminary manual screening by the development team, as previously described [17], may be beneficial. This allows a focus on the least obvious statements and decreases required ratings when recruiting high numbers of raters is less feasible.

The quality rating methods and analysis based on overall ratings or segment ratings appear to identify relevant need statements. The participants of this study were recruited from AMT, a community of workers spending significant time performing online tasks at a computer. It is noteworthy that the highest-rated cleaning need overall (see Table 5.6) related to computer keyboards, but this was not the case when isolating single segments. However, this study primarily sought to generate ratings for overall populations and intentionally did not assign individuals in particular segments to need statements relevant to this segment. For example, Table 5.8 includes a need statement with a low mean Importance score relating to pet friendly hotels. This need may be low importance to the overall population, but higher importance to pet owners. However, zero of the 15 raters of this need were pet owners; therefore, comparing overall results segment results in this case is not possible.

The examples of actual high-rated need statements submitted by users and initial quality rating data represent information to inform later need assessment activities (e.g. based on market size or intellectual property). After additional review, one item from such a list might be selected as an area to address during a concept generation phase.

The total duration for collecting quality data on over 1500 need statements was 6 days. This does not reflect continuous analysis time, only the duration of the data collection phases. This duration might increase if motivated raters were less available; however, this study demonstrates feasibility. Comparing analysis durations to existing methods is challenging due to insufficient data for existing methods.

5.3.4 The First Needs to Come to Mind Are Not Lower Quality than Later Needs

The results suggest that on average, a user is equally likely to list a top quality need if it is the first one to come to mind as if they spend prolonged periods of time reviewing images and examples and listing a need arising fifth or tenth or twentieth. This is consistent with prior results showing that users who submit a higher quantity of needs are more likely to have a higher number of high quality statements (e.g. top quartile); however, this was not consistent with an assumption that more tacit or latent needs might be articulated later and that such needs would be rated as higher quality. Our hypothesis that quality would be lower for the earliest submissions was not supported.

The hypothesis was based on expectations that the quality of needs might follow the same trend as the quality of ideas during ideation. While there is evidence to suggest that the earliest idea entries during ideation are often superficial or not novel, the results suggest this trend does not apply to quality of need statements.

This finding potentially impacts future work as it indicates that there may not be a penalty in quality for using a higher quantity of individuals rather than longer engagement with each person. There is a benefit to prolonged engagement in a higher quantity of needs per person overall; however, if the same number of needs was generated by a larger group with fewer per person, the end result may not be significantly different. It is possible that the additional diversity of the larger group provides new perspectives leading to new needs in a similar fashion as prolonged engagement can encourage new perspectives for a given individual. When performing user research of this kind, the relative costs of retaining each person for long periods of time should be balanced against the costs of recruiting additional individuals for shorter times.

The results do not conclusively demonstrate whether the rate of tacit or latent needs changes over time and do not address whether a tacit need would be rated as high quality using this metric. The results warrant additional research to address these issues. One potential explanation is that for some topic areas, readily articulated needs have not necessarily been addressed to the complete satisfaction of users. When aggregating large lists of needs from many people, some needs might be difficult to articulate for 99% of the group. However, one percent might have an experience or background allowing the need to be more readily articulated, and this need may be recognized as high quality by a high proportion of the group. Based on previous results, this difference in background is not necessarily a greater level of expertise, as user expertise or experience was not a significant factor for need quality.

5.3.5 Algorithmically-Rated Unique Need Statements Are Not Higher Quality than Those with Many Similar Variants

If a need statement is submitted and found to have many similar variations, the quality of this need is not significantly different from a need statement with few similar statements. This is a key finding as this differs from analogs in ideation where novel and less-common ideas are often considered more valuable (e.g. the objective is to generate creative ideas

and novelty is a metric for creativity). One perspective could be that highly-unique need statements are equally likely to be low quality. This differs from an assumption that tacit or hard to articulate needs would be scored as highly original and also as high quality. Additional research is required to understand if the STS algorithm is effective in identifying this type of need.

The finding of equivalent quality for STS-scored duplicates provides further support both for the use of automated algorithms in assessing large data sets, and for the quality metrics used for quantitative prioritization. While Fig. 5.13 shows a small number of STS duplicates resulting in a difference in rated quality of over 1.5 points, this can potentially be attributed to known rates of false positives for algorithm scores as well as large variation in human gold-standard ratings. Overall the result confirms that need statements scored as equivalent will typically have equivalent quality ratings.

5.3.6 Omitting Detailed Context When Rating Need Statement Quality May Not Effect Ratings

Performing a quantitative screening to prioritize lists of hundreds of need statements requires a large number of ratings. The quality rating metrics were devised to be simple and allow rapid throughput; however, these studies assumed that any need statement that included a detailed story should have the story available during the quality rating. In this scenario, if a user was rating the need statement and was unsure of the context or meaning, the story could provide this background. This process takes significantly more time, and based on this result, it is not necessary. The results suggest that the average final quality ratings will be equivalent even if the background story is omitted during rating.

This does not suggest that the background story should not be collected in the need collection phase. The background story may have value for other purposes. A user may have sufficient information to rate the need using simple metrics based on a summary sentence, but later phases where additional validation information is collected and potential solutions are proposed might benefit from this contextual information.

5.3.7 Each Type of Stimulus May Result in Quality Need Statements

In a real-world scenario for large-scale needfinding described here, users are able to select any type of stimulus information that is of interest. The results here suggest there is no penalty for user-directed selections. There was no type of stimulus that was dramatically better or worse than others since there was no evident trend for a higher proportion of high-quality needs. Additional research using randomized methods would be valuable to confirm this finding.

5.3.8 Limitations and Future Work

Our results apply to our methods, specifically using a content-rich web application to collect user needs, and other methods, such as focus groups or interviews, might have different results. While the limited effects of expertise are consistently demonstrated for these studies, additional research is warranted to confirm this result for additional methods.

The results primarily represent an analysis of overall population priorities. While the same types of analysis can be performed using population segments, results may vary more widely when considering a large range of diverse segments, in part because sample sizes per need statement per segment were much less uniform.

The need statements reflect verbatim content from users and do not include modifications (e.g. to increase consistency or restated to consider a related root cause of a problem). Data was collected without strict requirements on format or grammatical structures in order to avoid a cognitive demand that might decrease need count. The structure of need statements can impact later phases of development, and verbatim statements prioritized with this method can be further refined and iterated as more information is collected. Additional study is warranted to evaluate new methods to potentially maintain high quantity while collecting more structured statements from users or to apply previous methods to systematically rephrase existing statements [123]. These results also support further study to identify effects of additional need statement characteristics, such as the availability of a detailed story or whether the need was submitted early or later on that user's list.

Sets of problem statements do not represent inclusive lists of all needs. The topic

areas were intentionally selected as broadly applicable to a large population; however, the rate of unique statements exceeded expectations. Because each additional group member for a topic often added unique needs, there is little evidence of saturation of the qualitative data in this study. A more specific topic may have higher rates of duplicates, demonstrating saturation with a smaller group, and might suggest fewer unarticulated needs are remaining.

The data analysis frequently relied on quantitative metrics for quality ratings. These quality ratings included the overall group, even if some individuals within the group were not in a relevant segment of the population. The ratings represented an overall prioritization. For example, some needs for the topic of cleaning might be specific to those who own pets. A non-pet owner might rate this need as unimportant even if most pet owners rate it highly.

Chapter 6

Part 4: Medical Simulation Manikin Case Study

6.1 Medical Simulation Manikin Background

The needfinding methods described in Parts 1-3 (Chapters 3 - 5) were intentionally validated on general topics with users from the general population. This approach allowed numerous advantages such as decreased costs to recruit users, a large available pool of potential users, and high throughput for recruiting and data collection. However, as described in Section 2.0.3, technical topics such as medical technology may be a desirable application area.

A pilot case study was performed for the specialized topic of medical simulation manikins (synonymous with "mannequins"). These products are human-scale, physical patient simulators generally used for training health care providers. One example is shown in Figure 6.1. Other models are commercially available and include varying feature sets and sizes (e.g. infant models are available). Training activities range from team communication activities to simulated drug delivery.

A wide range of features are currently available to mimic human physiology for training purposes. The manikins might have articulating air chambers to act as lungs, magnetic actuators to simulate a pulse, and tubing and pumps to circulate body fluids. In addition, modern manikins are often wirelessly connected and can be programmed with different training scenarios or procedures. A scenario might dictate how and when the physiology of the manikin changes to be consistent with human symptoms. While the technology available in modern simulators is quite advanced and complex, these devices continue to suffer from a variety of limitations. For example, it is easy to see from Figure 6.1 that this is not really a human, and a caregiver in training would be unable to communicate with it in the same manner as a real patient. As the technology progresses, collecting a thorough set of existing unmet needs and problems could inform priorities for future development.

A number of research and training advocacy groups exist to disseminate effective teaching methods, including simulation-based teaching methods. One example is the American College of Surgeons (ACS) Accredited Education Institutes (AEI) Consortium. The Consortium includes 89 U.S. and international sites. The accreditation program is dedicated to developing a community of institutions committed to furthering surgical education. Given the common use of simulators in surgical education, the

representatives at these sites would have first-hand knowledge of the needs of surgical educators and students, how simulators have been used effectively in local programs, and how use could be improved.

Figure 6.1: iStan: Commercial Medical Simulation Manikin

6.2 Part 4: Case Study Methods

A pilot case study was conducted using an existing contact list maintained by AEI. The study's online needs assessment tool was reviewed by AEI Research and Development committee members and was approved to be sent to member sites. An email with a link to the assessment was sent directly by an AEI administrator email account. The email went to a representative from each of 89 sites. The instructions included a request to forward to other local program participants; therefore the exact number who received the email is not known.

The user interface and data collection methodology validated in Chapter 3 was employed in this study but with content specific to the medical simulation area. The online tool contained the same functionality as Part 1: Study 3. Respondents would review an introductory page with descriptions of the objectives. The stated objective was to collect a wide range of unmet needs relevant to next generation surgical simulators. Respondents then answered demographics questions customized for this technical field. Questions asked for general use environment (military or civilian), general role (Provider/Educator or Simulation Support), and years of experience.

The instructions for submitting the desired format of a need statement were the same as previous studies. Need statements should be a complete sentence and should not include an invention. Instructions were followed by a series of examples related to a generic topic of reading books. Respondents could then choose to take an optional quiz to test understanding of the instructions. The choices were to take the quiz or skip directly to entering need statements.

Previous studies have required completing and passing the quiz to continue. However, this step is likely to reduce completion rates when respondents are not incentivized in the same manner as Amazon Mechanical Turk workers. Previous studies have not collected data from those who failed the quiz, so there is currently no data to determine the difference in quality that might result when recruiting respondents who skip or fail the quiz.

Respondents viewed the same need statement entry screen with the same three "Help" choices as previously used in Part 1: Study 3. The selection of narrative prompts was reworded to relate to simulation manikins, but used the same overall traits and

matrix as described in Section 3.2.5. A randomly selected prompt was displayed when the prompt button was chosen.

The "shared images" help provided a random selection of 10 images, as previously done. A total pool of 140 images was curated from new photo sessions of manikin simulation training and also from the University of Minnesota Center for Research in Education and Simulation Technologies (CREST) and SimPORTAL digital archives. Images were processed to mask identifying details (e.g. faces, name badges). The existing collection of internal images was diverse and the benefit of additional pilot studies to request images from users was considered limited in value. The shared images content included a new selection criteria not used in previous studies. Some images available in the digital archives were clearly more appropriate for specific stakeholders. Displaying these images to different stakeholders might increase confusion. An example would be an image of a circuit board or mechanical pump inside the body of the simulator. A repair technician (in the role of Simulation Support) might immediately recognize these components and recall experiences using or repairing them. However, a training instructor who teaches students using only fully assembled and operational manikins might be confused seeing these unfamiliar inner workings. There is a risk this confusion could divert a respondent's attention away from his or her own experiences. Therefore, the set of images was first filtered to be consistent with the respondent's demographic answers (in particular, the general role), and then a random group of 10 filtered images was presented.

The "shared needs" help was generated using transcripts and audio recordings from previous need assessment focus group sessions. This process differs from those in Part 1 (see Section 3.2.6). Previous studies omitted the shared needs help from the first study, and these were added only after first collecting needs during the first study. Due to the availability of need statement data in the form of audio from manikin users, this content was used in place of verbatim content from online studies. The audio files included approximately 6 hours of discussion across two different sites, each site including 5-10 people at different times. Spoken statements were transcribed in a manner consistent with how a user might have written them. These editorial changes were made to allow the shared needs to model expected need statement formats as well as to stimulate new perspectives during the session. A total of 80 need statements were generated. In some

cases the audio recordings included an expanded description or discussion as well, and this became the basis for full length stories in selected cases. When the shared needs button was selected, a random list of 10 need statements was displayed.

Respondents could enter as few or as many need statements as desired and could view an unlimited number of help selections.

Data analysis differed from previous studies. The quantity of data was expected to be smaller than previous studies, and a complete assessment using established statistical tests and quality metrics was not necessarily warranted. However, the data did allow a comparison of need statements collected from the online application to those derived from focus group audio recordings. This comparison was performed using a combined list of new and previous need statements. This list was used to create a matrix of all pairwise comparisons. The list of statement pairs was analyzed using the UMBC semantic similarity algorithm as described in Section 4.1.5. This algorithm was chosen for use in the case study based on previous evidence it would be more robust when rating statements with different content relative to the data used to train the algorithm.

6.3 Part 4: Case Study Results

The online application for the needs assessment was accessed 21 times. Responses were anonymous, and the system did not track potential repeat users. The total number of unique users was likely approximately 21. Not all respondents who began the assessment continued to the end and submitted data. A total of seven respondents submitted 20 need statements. The maximum per person was seven statements and the minimum was one.

The results of the semantic analysis were similar to previous study results in a number of ways. There was no evidence of malicious duplications, and also little evidence of redundant entries in general. In cases where new need statements were similar to previous focus group data, the similar needs were not included in sets of sample needs shown to the particular respondent.

Table 6.1 provides samples from the set of 20 new need statements and similar previous statements derived from focus group audio recordings. The samples reflect the three highest similarity scores where one need from the pair was in the focus group set and the other was a new statement submitted by an AEI respondent. The highest scores from this data were not has high as in previous consumer product studies (see Part 2 in Section 4.2). One potential reason is the presence of medical and anatomical language which may not have been present in algorithm training data. The limited scope of this analysis did not warrant validating a specialized algorithm; however, testing similarity of clinical text is an active area of research [122] and specialized algorithms may be relevant for future work.

Table 6.1: Data from Web-based Needfinding Methods Compared to Focus Groups Using Automated Algorithm

Web-based Needfinding	Focus Group	Score
The manikin's response needs to be life-like.	I want an immediate response or reaction to an input to the manikin.	3.7
I wish radial pulse spot was more anatomically correct.	Manikin only has a pulse on the right radial wrist.	3.1
The manikin needs to elicit a human connection with the trainee.	Current manikins lack a human connection.	3.0

In several cases, web-based needfinding (crowd-sourced) provided unique statements

markedly different from the pool identified by the focus groups. Samples are provided in Table 6.2. Selections were chosen as representative of the specificity and technical detail included in some need statements.

Table 6.2: Unique Needs from Web-based Needfinding Methods (Not Similar to Those Identified in Focus Group Data)

Trauma man: the window for chest tube insertion is too low in the axilla. We teach the students only to place the tubes at nipple line or higher, and only a very small proportion of the window is above the nipple line.
Trauma man: Intercostal vessels are not anatomically correct- they get cut and fluid spills out when students insert chest tubes.
We would like to be able to place organs in Sim man's abdominal cavity so we do laparoscopy sim within an inter-professional education simulation e.g. with anesthesia and nursing staff in an OR.

6.4 Part 4: Case Study Discussion

6.4.1 Users Articulate Similar as Well as Unique Needs Compared to Focus Groups

The example need statements shown in Table 6.1 demonstrate that many of the unmet needs identified in a focus group setting can also be directly articulated by users in an asynchronous online application. The small sample of new need statements suggest there will be overlap in the data collected from these differing methods. Many needs might be similar; however, each method might also identify a number of unique needs not produced by the alternate method. Currently, there is insufficient data to determine if unique needs derived from a focus group and not identified during large-scale needfinding might be higher or lower quality than unique needs from the current methods. Given the objective for a divergent search of unmet needs during many development projects, input from multiple sources may be a benefit.

6.4.2 Lengthy Instructions Contributed to Low Completion Rate

The pilot case study was performed expecting the response rate would likely be lower than studies described in Part 1. The number of respondents beginning the assessment was comparable to the response rate of previous surveys conducted by AEI. A survey conducted approximately 2 years prior resulted in a response rate of 41% with a similar number of target sites. The current pilot case study achieved a rate of 25% beginning the assessment, but the rate of submitting need statements was much lower, at under 10% of the total list. As described in Section 3.3 for Part 1, a large number of AMT participants quit before reaching the need submission phase; however, overall rates of quitting were slightly lower in the manikin study. Approximately 50% of AMT participants submitted usable needs in Part 1 compared with under 35% of those who started the manikin study. In case studies where the total email distribution list is greater than 1,000, the final quantity of data might approach previous AMT studies; however, increasing this rate presents an opportunity for future research.

There are likely multiple causes for the lower completion rate; however, one comment in the feedback section is important to consider. The individual responded, “Instructions are complicated and are a deterrent to completing the survey.” Previous studies

described in Part 1 relied on participants who were self-selected with a willingness to complete online surveys and tasks with limited compensation. These participants had likely seen a wide variety in the level of detail of instructions and training for online tasks. The manikin case study respondents may have had less exposure to non-traditional survey tools, and were also not participating for the same types of incentives as AMT workers.

Future work will likely examine appropriate incentive methods, possibly incorporating incentives previously described for online technical communities and training, such as peer recognition and continuing education credit [120, 119]. However, additional review of the need statement instructions and training is also warranted. To date, no study using these methods has compared the quality of need statements submitted by those who review detailed instructions and pass a quiz to need statements submitted by those who did not. Early studies relied on a conservative approach to maintain integrity of data knowing that the pool of study recruits is effectively limitless and cost of recruiting is low. Future need assessments with technical professionals will be limited to fewer potential respondents and maximizing completion rates will be an important element of a successful outcome. If data quality remains high with a more limited depth to the instructions and training, the perceived time commitment will be lower, potentially increasing completion rates. Submitting the needs assessment to a larger email contact list might produce a larger, comprehensive data set and would justify the resources to prioritize the list of need statements as done in Part 3.

Chapter 7

Conclusions

While in-depth user research methods (e.g. traditional ethnography) can effectively capture a wide range of user perspectives, short-duration approaches in large-scale needfinding are also effective for a divergent process of capturing needs. The specific methods used during needfinding can positively effect a user's ability to articulate high-quality needs, and these methods include web applications with visual and textual stimuli. This suggests further improvements in articulating tacit and latent needs can be achieved given additional research. The results confirm there is significant value in recruiting a large group with a wide range of user demographics if the objective is to capture a large portion of the total unmet needs space for a given topic, even if participants submit a small number each. The high count of cumulative unique needs from large groups suggests the portion of the total needs space that has been captured can be comparable to in-depth methods with smaller groups of users. Equally important, the methods described here can minimize the resources required to manage data from these large groups. Lastly, the evidence that the first needs to be articulated are not lower quality than later needs supports a balanced approach to recruiting new users and retaining existing users for immersive and in-depth input and suggests that a large enough diverse group contributing only a few needs per person can ultimately produce a valuable list of high-quality needs. Table 7.1 provides a summary of all hypotheses and research questions and hypothesis results.

The combined studies in Part 1 (Quantity) provide strong evidence that users will

Table 7.1: Overview of Hypotheses, Research Questions, and Results (Bold Indicates Key Contribution)

Hypothesis/Research Question	**Section**	**Result**
Does any specific type of stimulus help a user articulate a higher quantity of needs?	3.3.2	
Do levels of expertise or experience affect the quantity of needs a user can articulate?	3.3.4	
Can automated machine learning algorithms detect duplication among textual need statements?	4.2.2	
H1: Increasing the number of participants contributing found needs increases the quantity of unique needs.	4.2.4	Confirmed
H2: Increasing the number of participants submitting needs increases the number of high-quality needs as judged by users.	5.2.2	Confirmed
H3: Increasing the quantity of needs contributed per person increases the number of high-quality needs as judged by users.	5.2.3	Confirmed
H4: Increasing levels of self-rated user expertise will not significantly increase the number of high-quality needs per person.	5.2.4	Confirmed
H5: Needs submitted first would be less likely to be high quality than needs submitted after a sustained period of time.	5.2.7	Not Confirmed
H6: Semantically similar need statements would be rated as equivalent in quality.	5.2.8	Confirmed
H7: Need statements would be rated as higher quality if a detailed description of the need context was available.	5.2.9	Not Confirmed
Does the type of stimulus seen before entering a need affect need quality?	5.2.10	
Does the uniqueness of a need statement affect the need quality?	5.2.11	
Will online needfinding result in similar and/or unique needs compared to focus groups?	6.3	

have the ability to articulate needs directly when the interaction is mediated by sufficient background and instructions, incentives, and stimuli. The results of these studies indicate that specific stimulus types, such as shared needs and contextual images, can significantly increase the quantity of needs collected. However, users can generate nearly half of all needs prior to specific stimuli, suggesting that a specific stimulus might be beneficial but is not required. The incentive structure and user interface can be used to influence user behavior, such as to increase the amount of background information provided with each need. User expertise did not result in a significant difference of needs generated, and although the level of experience was significant, the degree of this change may not warrant exclusive use in practice of high-experience users for need quantity generation.

The Part 2 (Uniqueness) results demonstrate that as the group size increases, the group generates a higher quantity of unique cumulative needs. The results of automated analysis support the use of Semantic Textual Similarity (STS) algorithms for rapidly assessing needs-based open innovation textual data to facilitate comprehensive evaluations of quality. Automated analysis via STS algorithms can scale to the large volumes typical in open-innovation data and overcome a significant impediment to its widespread use: the bottleneck of prohibitive human resource costs to process such data. The accuracy of this method is promising, particularly given the low consensus levels of manual human ratings. The correlations of predicted similarity values to human ratings were generally high (up to .85) and indicated that specialized tuning of the algorithm to need statement applications is beneficial but not required. The accuracy of a similarity rating for statements with equivalent meaning approached .75, and the decrease in accuracy for ratings of lower similarity is partially attributed to a lack of consensus in human ratings for sentence equivalency. The improvement in results for non-tuned algorithms from 2012 to 2013 provides further support that NLP research will lead to ongoing improvements relevant to this application.

The results in Part 3 (Quality) support the use of simplified metrics of Importance and Satisfaction to initially screen and prioritize large numbers of need statements and provide further support for the feasibility of methods to perform large-scale needfinding using large groups of diverse users. The results confirm that the number of high-quality need statements directly articulated from users will increase when asking a larger group

and also when using known methods to help users articulate more needs per person. User demographics (e.g. self-rated expertise and hours per week) were not significantly associated with increasing quantities of high-quality needs for users with greater than zero hours per week. A need statement submitted by one experience group (e.g. Up to 5 hours/wk) could be rated for quality by the same experience group or any different experience group (e.g. 5-10 hours/wk) without significantly effecting quality scores. The results show the needs that first come to mind are not lower quality than needs that come to mind later. This enables a user researcher to combine available methods based on constraints on accessing user environments, recruiting respondents, and retaining participation for long periods of time. The results also suggest prioritizing need statements using quality ratings of a summary sentence would be equivalent to a more resource-intensive process of reviewing detailed contextual information regarding the needs (e.g. background stories). Quality ratings of semantically similar statements were equivalent, providing support for the use of automated algorithms to identify duplicate statements.

The case study results demonstrate the feasibility of collecting need statements using similar methods applied to a target application area of professionals in a medical setting, although further refinement of the data collection interface and instructions may be beneficial to increase response rates. The set of needs collected through a web-based application partially overlapped a set of needs identified through focus groups but also generated additional unique needs to complement these other sources (e.g. focus groups).

References

[1] Cory R Schaffhausen and Timothy M Kowalewski. Large-Scale Needfinding: Methods of Increasing User-Generated Needs from Large Populations. *Journal of Mechanical Design*, 137(7):071403, 2015.

[2] C. Schaffhausen and T.K. Kowalewski. Large-Scale Needs-Based Open Innovation Via Automated Semantic Textual Similarity Analysis. In *Proceedings of the ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Boston, MA, 2015. In Press, Full Citation Available at Time of Publication.

[3] Cory R Schaffhausen and Timothy M Kowalewski. Assessing Quality of User-Submitted Need Statements from Large-Scale Needfinding: Effects of Expertise and Group Size. *Journal of Mechanical Design*, 2015, In Review.

[4] Cory R Schaffhausen and Timothy M Kowalewski. Assessing Quality of Unmet User Needs: Effects of Need Statement Characteristics. *Design Studies*, 2015, In Review.

[5] E. Von Hippel. Lead Users: A Source of Novel Product Concepts. *Management Science*, 32(7):791–805, 1986.

[6] S.R. Rosenthal and M. Capper. Ethnographies in the Front End: Designing for Enhanced Customer Experiences. *Journal of Product Innovation Management*, 23(3):215–237, 2006.

[7] G.R. Schirr. Flawed Tools: The Efficacy of Group Research Methods to Generate Customer Ideas. *Journal of Product Innovation Management*, 29(3):473–488, 2012.

[8] S. Hyysalo. Some Problems in the Traditional Approaches to Predicting the Use of a Technology-Driven Invention. *Innovation: The European Journal of Social Science Research*, 16(2):117–137, 2003.

[9] C. Pope. Conducting Ethnography in Medical Settings. *Medical Education*, 39(12):1180–1187, 2005.

[10] Dev Patnaik and Robert Becker. Needfinding: The Why and How of Uncovering People's Needs. *Design Management Journal (Former Series)*, 10(2):37–43, 1999.

[11] D Patnaik. *Needfinding: design research and planning*. CreateSpace Independent Publishing Platform, Lexington, KY, 3rd edition, 2014.

[12] Stefanos A Zenios, Josh Makower, Paul G Yock, Todd J Brinton, Uday N Kumar, Lyn Denend, and Thomas M Krummel. *Biodesign: the process of innovating medical technologies*. Cambridge University Press, New York, NY, 2010.

[13] Jacob W Getzels. Problem Finding: a Theoretical Note. *Cognitive science*, 3(2):167–172, 1979.

[14] Sara L Beckman and Michael Barry. Innovation as a Learning Process: Embedding Design Thinking. *California Management Review*, 50(1):24–56, 2007.

[15] J.L. Martin, E. Murphy, J.A. Crowe, and B.J. Norris. Capturing User Requirements in Medical Device Development: The Role of Ergonomics. *Physiological Measurement*, 27(8):R49–R62, 2006.

[16] R. Cooper and S. Edgett. Ideation for Product Innovation: What are the Best Methods? *PDMA Visions Magazine*, 1(1):12–17, 2008.

[17] Karl T Ulrich and Steven D Eppinger. *Product design and development*. McGraw-Hill/Irwin, New York, NY, 3rd edition, 2004.

[18] T. Kelley and J. Littman. *The Art of Innovation*. Number ISBN-13: 978-0385499842. Crown Business, New York, NY, 2001.

[19] E. Gundlin. *The 3M Way to Innovation*. Number ISBN-13: 978-4770024763. Kodansha Intl., Tokyo, Japan, 1st edition, 2000.

[20] Mahir Alkaya, Froukje Sleeswijk Visser, and Christine De Lille. Supporting NPD Teams in Innovation: Structuring User Data on the Foundations of Empathy. In *Leading Innovation Through Design: Proceedings of the Design Management Institute 2012 International Research Conference*, pages 1–8. Design Management Institute, 2012.

[21] Merlijn Kouprie and Froukje Sleeswijk Visser. A Framework for Empathy in Design: Stepping Into and Out of the User's Life. *Journal of Engineering Design*, 20(5):437–448, 2009.

[22] Daniel G Johnson, Nicole Genco, Matthew N Saunders, Paul Williams, Carolyn Conner Seepersad, and Katja Hölttä-Otto. An Experimental Investigation of the Effectiveness of Empathic Experience Design for Innovative Concept Generation. *Journal of Mechanical Design*, 136(5):051009–1–9, 2014.

[23] Dev Patnaik. *Wired to care: How companies prosper when they create widespread empathy*. Rearson Education, Inc., Upper Saddle River, New Jersey, USA, 2009.

[24] R. Elliott and N. Jankel-Elliott. Using Ethnography in Strategic Consumer Research. *Qualitative Market Research: An International Journal*, 6(4):215–223, 2003.

[25] W.C. Savenye and R.S. Robinson. Qualitative research issues and methods: An introduction for educational technologists. *Handbook of Research for Educational Communications and Technology*, pages 1171–1195, 1996.

[26] R. Weiss. *Learning From Strangers: The Art and Method of Qualitative Interview Studies*. Number ISBN-13: 978-0684823126. Free Press, New York, NY, 1st edition, 1995.

[27] Herbert J Rubin and Irene S Rubin. *Qualitative interviewing: The art of hearing data*. SAGE Publications, Incorporated, Thousand Oaks, CA), 2011.

[28] Abbie Griffin and John R Hauser. The Voice of the Customer. *Marketing science*, 12(1):1–27, 1993.

[29] Barry L. Bayus. Understanding customer needs. In Scott Shane, editor, *The handbook of technology and innovation management*, pages 115–141. John Wiley & Sons, West Sussex, England, 2008.

[30] Jacob Goldenberg, Donald R Lehmann, and David Mazursky. The Idea Itself and the Circumstances of its Emergence as Predictors of New Product Success. *Management Science*, 47(1):69–84, 2001.

[31] Ravi Balachandra and John H Friar. Factors for Success in R&D Projects and New Product Innovation: a Contextual Framework. *Engineering Management, IEEE Transactions on*, 44(3):276–287, 1997.

[32] Morris B. Holbrook and Elizabeth C. Hirschman. The Experiential Aspects of Consumption: Consumer Fantasies, Feelings, and Fun. *Journal of Consumer Research*, 9(2):pp. 132–140, 1982.

[33] Barry J. Babin, William R. Darden, and Mitch Griffin. Work and/or Fun: Measuring Hedonic and Utilitarian Shopping Value. *Journal of Consumer Research*, 20(4):pp. 644–656, 1994.

[34] Pieter MA Desmet and Paul Hekkert. Framework of Product Experience. *International Journal of Design*, 1(1):57–66, 2007.

[35] Josip Mikulic and Darko Prebezac. A Critical Review of Techniques for Classifying Quality Attributes in the Kano Model. *Managing Service Quality: An International Journal*, 21(1):46–66, 2011, http://dx.doi.org/10.1108/09604521111100243.

[36] Timothy W Simpson, Aaron Bobuk, Laura A Slingerland, Sean Brennan, Drew Logan, and Karl Reichard. From User Requirements to Commonality Specifications: An Integrated Approach to Product Family Design. *Research in Engineering Design*, 23(2):141–153, 2012.

[37] Phillip Cormier, Andrew Olewnik, and Kemper Lewis. Toward a Formalization of Affordance Modeling for Engineering Design. *Research in Engineering Design*, 25(3):259–277, 2014.

[38] Benjamin T Ciavola, Chunlong Wu, and John K Gershenson. Integrating Function-and Affordance-Based Design Representations. *Journal of Mechanical Design*, 137(5):051101, 2015.

[39] Matthew G Green, Palanisamy Kuppuraj Palani Rajan, and Kristin L Wood. Product Usage Context: Improving Customer Needs Gathering and Design Target Setting. In *ASME 2004 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages 393–403. 2004.

[40] Dominique Scaravetti, Jean-Pierre Nadeau, Jérôme Pailhès, and Patrick Sebastian. Structuring of Embodiment Design Problem Based on the Product Lifecycle. *International Journal of Product Development*, 2(1-2):47–70, 2005.

[41] Anthony W Ulwick. *What customers want: using outcome-driven innovation to create breakthrough products and services*, volume 71408673. McGraw-Hill, New York, NY, 2005.

[42] Anthony W Ulwick. Turn Customer Input into Innovation. *Harvard Business Review*, 80(1):91–7, 2002.

[43] Kurt Matzler and Hans H Hinterhuber. How to Make Product Development Projects More Successful by Integrating Kano's Model of Customer Satisfaction into Quality Function Deployment. *Technovation*, 18(1):25–38, 1998.

[44] Xiaojuan Ma, Li Yu, Jodi L Forlizzi, and Steven P Dow. Exiting the Design Studio: Leveraging Online Participants for Early-Stage Design Feedback. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 676–685. 2015.

[45] J. Martin and J. Barnett. Integrating the Results of User Research into Medical Device Development: Insights from a Case Study. *BMC Medical Informatics and Decision Making*, 12(74):1–10, 2012.

[46] Daniel B Kramer, Shuai Xu, and Aaron S Kesselheim. How Does Medical Device Regulation Perform in the United States and the European Union? A Systematic Review. *PLoS Medicine*, 9(7):e1001276, 2012.

[47] A. Money, J. Barnett, J. Kuljis, M. Craven, J. Martin, and T. Young. The Role of the User Within the Medical Device Design and Development Process: Medical Device Manufacturers' Perspectives. *BMC Medical Informatics and Decision Making*, 11(15):1–12, 2011.

[48] A. Shah and S. Alshawi. The Role of User Requirements Research in Medical Device Development. In *Proceedings of the European and Mediterranean Conference on Information Systems*, pages 1–25. EMCIS2010, Abu Dhabi, UAE, 2010.

[49] B.C. Poulton. User Involvement in Identifying Health Needs and Shaping and Evaluating Services: Is It Being Realised? *Journal of Advanced Nursing*, 30(6):1289–1296, 2001.

[50] M.B. Ram, N. Campling, P. Grocott, and H. Weir. A Methodology for a Structured Survey of the Healthcare Literature Related to Medical Device Users. *Evaluation*, 14(1):49–73, 2008.

[51] K. Vredenburg, J.Y. Mao, P.W. Smith, and T. Carey. A Survey of User-Centered Design Practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves*, pages 471–478. 2002.

[52] Lauren M Aquino Shluzas, Martin Steinert, and Larry J Leifer. Designing to Maximize Value for Multiple Stakeholders: A Challenge to Med-tech Innovation. In *DS 68-10: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 10: Design Methods and Tools pt. 2, Lyngby/Copenhagen, Denmark, 15.-19.08. 2011*, pages 159–166. 2011.

[53] S.G.S. Shah and I. Robinson. Benefits of and Barriers to Involving Users in Medical Device Technology Development and Evaluation. *International Journal of Technology Assessment in Health Care*, 23(1):131–137, 2007.

[54] M. Kauppinen, S. Kujala, T. Aaltio, and L. Lehtola. Introducing Requirements Engineering: How to Make a Cultural Change Happen in Practice. In *Requirements Engineering, 2002. Proceedings. IEEE Joint International Conference on*, pages 43–51. 2002.

[55] Barry Matthew Kudrowitz and David Wallace. Assessing the Quality of Ideas from Prolific, Early-Stage Product Ideation. *Journal of Engineering Design*, 24(2):120–139, 2013.

[56] Douglas L Dean, Jillian M Hender, Thomas L Rodgers, and Eric L Santanen. Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *Journal of the Association for Information Systems*, 7(10):646–698, 2006.

[57] Alex F. Osborn. *Applied imagination.* Scribner's, New York, NY, 1953.

[58] Michael Diehl and Wolfgang Stroebe. Productivity Loss in Brainstorming Groups: Toward the Solution of a Riddle. *Journal of Personality and Social Psychology*, 53(3):497–509, 1987.

[59] Karen Leggett Dugosh and Paul B Paulus. Cognitive and Social Comparison Processes in Brainstorming. *Journal of Experimental Social Psychology*, 41(3):313–320, 2005.

[60] Paul B Paulus, Nicholas W Kohn, and Lauren E Arditti. Effects of Quantity and Quality Instructions on Brainstorming. *The Journal of Creative Behavior*, 45(1):38–46, 2011.

[61] Paul B Paulus, Mary T Dzindolet, George Poletes, and L Mabel Camacho. Perception of Performance in Group Brainstorming: the Illusion of Group Productivity. *Personality and Social Psychology Bulletin*, 19(1):78–89, 1993.

[62] Brian Mullen, Craig Johnson, and Eduardo Salas. Productivity Loss in Brainstorming Groups: A Meta-Analytic Integration. *Basic and Applied Social Psychology*, 12(1):3–23, 1991.

[63] Michael Diehl and Wolfgang Stroebe. Productivity Loss in Idea-Generating Groups: Tracking Down the Blocking Effect. *Journal of personality and social psychology*, 61(3):392, 1991.

[64] L Mabel Camacho and Paul B Paulus. The Role of Social Anxiousness in Group Brainstorming. *Journal of personality and social psychology*, 68(6):1071, 1995.

[65] Paul B Paulus and Mary T Dzindolet. Social Influence Processes in Group Brainstorming. *Journal of Personality and Social Psychology*, 64(4):575, 1993.

[66] Steven J Karau and Kipling D Williams. Social Loafing: a Meta-analytic Review and Theoretical Integration. *Journal of Personality and Social Psychology*, 65(4):681, 1993.

[67] Jr. Nunamaker, Jay F., Robert O. Briggs, Daniel D. Mittleman, Douglas R. Vogel, and Pierre A. Balthazard. Lessons from a Dozen Years of Group Support Systems Research: A Discussion of Lab and Field Findings. *Journal of Management Information Systems*, 13(3):pp. 163–207, 1996.

[68] Luis L Martins, Lucy L Gilson, and M Travis Maynard. Virtual Teams: What Do We Know and Where Do We Go from Here? *Journal of management*, 30(6):805–835, 2004.

[69] Arthur B VanGundy. Brain Writing for New Product Ideas: An Alternative to Brainstorming. *Journal of Consumer Marketing*, 1(2):67–74, 1984.

[70] R Brent Gallupe, Alan R Dennis, William H Cooper, Joseph S Valacich, Lana M Bastianutti, and Jay F Nunamaker. Electronic Brainstorming and Group Size. *Academy of Management Journal*, 35(2):350–369, 1992.

[71] Alain Pinsonneault, Henri Barki, R Brent Gallupe, and Norberto Hoppen. Electronic Brainstorming: The Illusion of Productivity. *Information Systems Research*, 10(2):110–133, 1999.

[72] Darleen M DeRosa, Carter L Smith, and Donald A Hantula. The Medium Matters: Mining the Long-Promised Merit of Group Interaction in Creative Idea Generation Tasks in a Meta-Analysis of the Electronic Group Brainstorming Literature. *Computers in Human Behavior*, 23(3):1549–1581, 2007.

[73] Joseph S Valacich, Alan R Dennis, and Terry Connolly. Idea Generation in Computer-Based Groups: A New Ending to an Old Story. *Organizational Behavior and Human Decision Processes*, 1994.

[74] Pao Siangliulue, Kenneth C Arnold, Krzysztof Z Gajos, and Steven P Dow. Toward Collaborative Ideation at Scale: leveraging Ideas from Others to Generate More Creative and Diverse Ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 937–945. 2015.

[75] John Morgan and Richard Wang. Tournaments for Ideas. *California Management Review*, 52(2):77–97, 2010.

[76] Terri R Kurtzberg and Teresa M Amabile. From Guilford to Creative Synergy: opening the Black Box of Team-level Creativity. *Creativity Research Journal*, 13(3-4):285–294, 2001.

[77] David G Jansson and Steven M Smith. Design Fixation. *Design Studies*, 12(1):3–11, 1991.

[78] JS Linsey, I Tseng, K Fu, J Cagan, KL Wood, and C Schunn. A Study of Design Fixation, its Mitigation and Perception in Engineering Design Faculty. *Journal of Mechanical Design*, 132(4):e041003, 2010.

[79] Vimal K Viswanathan and Julie S Linsey. Design Fixation and its Mitigation: A Study on the Role of Expertise. *Journal of Mechanical Design*, 135(5):051008, 2013.

[80] K Anders Ericsson and Herbert A Simon. Verbal Reports as Data. *Psychological review*, 87(3):215, 1980.

[81] Ted Boren and Judith Ramey. Thinking Aloud: reconciling Theory and Practice. *Professional Communication, IEEE Transactions on*, 43(3):261–278, 2000.

[82] IDEO. Method cards: 51 ways to inspire design. 2003.

[83] Joseph Lin and Carolyn Conner Seepersad. Empathic Lead Users: The Effects of Extraordinary User Experiences on Customer Needs Analysis and Product Redesign. In *ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages 289–296. 2007.

[84] Henry W Chesbrough. The Era of Open Innovation. *Managing innovation and change*, 127(3):34–41, 2006.

[85] Steven Dow, Elizabeth Gerber, and Audris Wong. A Pilot Study of Using Crowds in the Classroom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 227–236. ACM, New York, NY, USA, 2013.

[86] Thomas Stone and Seung-Kyum Choi. Extracting Consumer Preference from User-Generated Content Sources Using Classification. In *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages V03AT03A031–V03AT03A031. 2013.

[87] Haakon Faste. Opening "Open" Innovation. In *Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces*, DPPI '11, pages 54:1–54:8. ACM, New York, NY, USA, 2011.

[88] Gabriela Goldschmidt and Anat Litan Sever. Inspiring Design Ideas with Texts. *Design Studies*, 32(2):139–155, 2011.

[89] Yoram Reich, Suresh L Konda, Ira A Monarch, Sean N Levy, and Eswaran Subrahmanian. Varieties and Issues of Participation and Design. *Design Studies*, 17(2):165–180, 1996.

[90] Bernard A Nijstad, Wolfgang Stroebe, and Hein FM Lodewijkx. Cognitive Stimulation and Interference in Groups: Exposure Effects in an Idea Generation Task. *Journal of experimental social psychology*, 38(6):535–544, 2002.

[91] Karen Leggett Dugosh, Paul B Paulus, Evelyn J Roland, and Huei-Chuan Yang. Cognitive Stimulation in Brainstorming. *Journal of Personality and Social Psychology*, 79(5):722 – 735, 2000.

[92] Nicholas W Kohn, Paul B Paulus, and YunHee Choi. Building on the Ideas of Others: An Examination of the Idea Combination Process. *Journal of Experimental Social Psychology*, 47(3):554–561, 2011.

[93] Alan R Dennis, Randall K Minas, and Akshay P Bhagwatwar. Sparking Creativity: Improving Electronic Brainstorming with Individual Cognitive Priming. *Journal of Management Information Systems*, 29(4):195–216, 2013.

[94] Christine A Toh and Scarlett R Miller. The Impact of Example Modality and Physical Interactions on Design Creativity. *Journal of Mechanical Design*, 136(9):091004, 2014.

[95] Daren C Brabham. Crowdsourcing as a Model for Problem Solving an Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies*, 14(1):75–90, 2008.

[96] Marion K. Poetz and Martin Schreier. The Value of Crowdsourcing: Can Users Really Compete with Professionals in Generating New Product Ideas? *Journal of Product Innovation Management*, 29(2):245–256, 2012.

[97] Kevin J Boudreau and Karim R Lakhani. Using the Crowd as an Innovation Partner. *Harvard Business Review*, 91(4):60–69, 2013.

[98] E. Enkel, O. Gassmann, and H. Chesbrough. Open r&d and open innovation: Exploring the phenomenon. *R&D Management*, 39(4):311–316, 2009.

[99] Adam Westerski. *Semantic Technologies in Idea Management Systems: A Model for Interoperability, Linking and Filtering.* PhD thesis, Universidad Politécnica de Madrid (UPM), Madrid, 2013.

[100] Guido Jouret. Inside Cisco's Search for the Next Big Idea. *Harvard Business Review*, 87(9):43 – 45, 2009.

[101] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in Action: Covers Apache Lucene 3.0.* Manning Publications Co., Greenwich, CT, USA, 2010.

[102] A. Westerski, C.A. Iglesias, and J.E. Garcia. Idea Relationship Analysis in Open Innovation Crowdsourcing Systems. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on*, pages 289–296. Pittsburgh, PA, USA, Oct 2012.

[103] Gobinda G. Chowdhury. Natural Language Processing. *Annual Review of Information Science and Technology*, 37(1):51–89, 2003.

[104] E. Cambria and B. White. Jumping NLP Curves: A Review of Natural Language Processing Research. *Computational Intelligence Magazine, IEEE*, 9(2):48–57, May 2014.

[105] Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Montréal, Canada, 2012.

[106] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. SEM 2013 Shared Task: Semantic Textual Similarity, Including a Pilot on Typed-Similarity. In **SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics*. Atlanta, Georgia, USA, 2013.

[107] Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. Takelab: Systems for Measuring Semantic Text Similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 441–448. Association for Computational Linguistics, Montréal, Canada, 7-8 June 2012.

[108] Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 44–52. Atlanta, Georgia, USA, 2013.

[109] Noriaki Kano, Nobuhiko Seraku, Fumio Takahashi, and Shinichi Tsuji. Attractive Quality and Must-Be Quality. *Journal of the Japanese Society for Quality Control*, 14(2):147–156, 1984.

[110] Chun-Chih Chen and Ming-Chuen Chuang. Integrating the Kano Model into a Robust Design Approach to Enhance Customer Satisfaction with Product Design. *International Journal of Production Economics*, 114(2):667–681, 2008.

[111] A. M. M. Sharif Ullah and Jun'ichi Tamaki. Analysis of Kano-Model-Based Customer Needs for Product Development. *Systems Engineering*, 14(2):154–172, 2011.

[112] Fatemeh Zahedi. The Analytic Hierarchy Process: A Survey of the Method and Its Applications. *Interfaces*, 16(4):pp. 96–108, 1986.

[113] Jože Duhovnik, Janez Kušar, Marko Starbek, et al. Development Process with Regard to Customer Requirements. *Concurrent Engineering*, 14(1):67–82, 2006.

[114] V Agouridas, H Winand, A McKay, and A de Pennington. Early Alignment of Design Requirements with Stakeholder Needs. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 220(9):1483–1507, 2006, http://pib.sagepub.com/content/220/9/1483.full.pdf+html.

[115] Gabriele Paolacci, Jesse Chandler, and Panagiotis Ipeirotis. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

[116] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46(4):1023–1031, 2014.

[117] Frank Bretz, Torsten Hothorn, and Peter Westfall. *Multiple comparisons using R*. CRC Press, Boca Raton, FL, 2010.

[118] Winter Mason and Duncan J Watts. Financial Incentives and the Performance of Crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.

[119] Oded Nov, David Anderson, and Ofer Arazy. Volunteer Computing: A Model of the Factors Determining Contribution to Community-Based Scientific Research. In *Proceedings of the 19th International Conference on World Wide Web*, pages 741–750. 2010.

[120] Vernon R Curran and Lisa Fleet. A Review of Evaluation Outcomes of Web-Based Continuing Medical Education. *Medical education*, 39(6):561–567, 2005.

[121] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. ROCR: Visualizing Classifier Performance in R. *Bioinformatics*, 21(20):3940–3941, 2005.

[122] Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. SemEval-2014 Task 7: Analysis of Clinical Text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, volume 199, pages 54–62. Dublin, Ireland, 2014.

[123] David Vernon and Ian Hocking. Thinking Hats and Good Men: Structured Techniques in a Problem Construction Task. *Thinking Skills and Creativity*, 14:41–46, 2014.

Appendix A

Part 1: Quantity

A.1 Part 1: Quantity Study 1 Additional Material

The exact content of the first study is included in this section. Actual content is reflected in normal font, and editorial information is *provided in italics.* Not all text formatting used in the final user interface is reflected in this section.

A.1.1 Amazon Mechanical Turk Interface for Study 1

- *Title for the HIT*

  Describe problems with common products and services

- *Summary description shown to AMT workers*

  This is a research study where you can submit examples of problems with common products and services. You will see the specific topic after you begin. Write more to earn more. Earn frequent bonuses of about $.10 per paragraph.

- *Keywords entered into AMT*

  needs problems study research survey product

- *Main body of the HIT displayed above the box to paste in the completion code*

  Tell us how products and services you use haven't worked like they should.

  The steps include:

  - Review instructions
  - Review training activity ($0.65 payment)
  - Describe problems related to your topic to earn bonuses

  Responses are open-ended and after every 5th response, you are eligible for a $0.20 bonus payment. Earn additional bonus payments of $0.10 if you give detailed reasons for your responses. Minimum Participation times will be approximately 5-15 minutes and you are encouraged to continue beyond this and earn additional payments after each additional 5th response as long as you would like. Incomplete or duplicate responses may not be approved.

  Survey link: Open survey in new tab

A.1.2 Zoho Creator App for Study 1

- *Page 1*

  Thank you for taking the next step on our HIT.

  Your survey is about: [topic area].

  [topic area description]

If you do not have significant experience [topic area], that is OK, as long as you are interested in explaining what problems you have and what you think could be improved.

The text above was calculated and displayed after the participant was assigned to a specific topic area. The text inserted by these calculations is shown below.

Table A.1: *Calculated Text to Describe each Topic Area*

Topic Area	**Topic Area Description**
Preparing food and cooking	This includes any step you take to start with food on the shelf or in the refrigerator and end with a meal ready to eat.
Doing housecleaning and household chores	This includes cleaning up messes, whether dirt or clutter, doing laundry, sorting mail, and other jobs around the house.
Planning a trip	This includes any travel beyond daily routines. Trips might be work-related, vacations, local, abroad, by yourself, or with others.

- *Page 2*

  Consent Information:

  The information you provide going forward will be stored and used for the research study. Review the information below and indicate you agree to participate if you would like to continue. You are invited to be in a research study of how common products and services might fail to meet your needs. You were selected as a possible participant because you selected this HIT on Amazon Mechanical Turk. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

  This study is being conducted by: Timothy Kowalewski, PhD, and Cory Schaffhausen, University of Minnesota

  Procedures:

  If you agree to be in this study, we would ask you to do the following things: The steps include:

  – Review instructions
  – Review training activity ($.65 standard payment)
  – Answer questions

  Responses are open-ended and after every 5th response, you are eligible for a $.20 bonus payment. Earn additional bonus payments of $.10 if you give detailed

reasons for your responses. Participation times will be approximately 5-15 minutes and you are encouraged to continue beyond this and earn additional payments after each additional 5th response. Incomplete or duplicate responses may not be approved.

Confidentiality:

The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify a subject. Research records will be stored securely and only researchers will have access to the records. Voluntary Nature of the Study: Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota. If you decide to participate, you are free to not answer any question or withdraw at any time without affecting those relationships.

Contacts and Questions:

The researcher(s) conducting this study is (are): Timothy Kowalewski, Assistant Professor. You may call or email with any questions you have now. If you have questions later, you are encouraged to contact him at University of Minnesota, ME 207, 111 Church St, Minneapolis, MN 55455, 625-626-0054, timk@umn.edu. If you have any questions or concerns regarding this study and would like to talk to someone other than the researcher(s), you are encouraged to contact the Research Subjects Advocate Line, D528 Mayo, 420 Delaware St. Southeast, Minneapolis, Minnesota 55455; (612) 625-1650. You are encouraged to retain a copy of this information to keep for your records.

In order to continue, you must agree to the following:

- I am age 18 or older.
- I have read and understood the information above.
- I want to participate in this study and continue with the HIT.

I agree Yes No

- *Page 3*

  Step 1: Review Instructions

  Please follow the instructions listed below. We want to know what would make [topic area] a better experience. Examples can be very broad, for example: [examples of better].

  You will type in descriptions of problems or unmet needs you face [topic area]. You want to describe these so someone could make improvements or offer solutions in the future. Try to think of as many as you can.

  The instructions for this HIT include:

1. Clearly and completely describe the problem, ideally in a complete sentence.
2. You should NOT include inventions or ideas to solve the problem. You should only list the problem itself.

The code you need for Amazon Mechanical Turk payment will be presented when you are ready to stop.

The text above was calculated and displayed after the participant was assigned to a specific topic area. The text inserted by these calculations is shown below.

Table A.2: *Calculated Text to Describe What Better Would Mean for Each Topic Area*

Topic Area	**Examples of Better**
Preparing food and cooking	more convenient, less effort, safer, easier to understand, cheaper, more consistent, or faster
Doing housecleaning and household chores	more convenient, less effort, safer, easier to understand, cheaper, or faster
Planning a trip	more convenient, more enjoyable, safer, easier to understand, a better value, or faster to use

- *Page 4*

  Step 2: Training

  Read this page of instructions carefully so you can pass the quiz on the next page. You will earn a payment of $.65 if you complete the training, but you are only able to continue and earn bonuses if you pass.

  Your training and quiz will cover a different topic: Reading Books

  This includes reading, holding and carrying a paper book.

  Review these examples and whether they would be consistent with the instructions you were given. It doesn't matter whether you agree that these are needs, for now just focus on the instructions.

  Example 1:

  Hard to hold.

  No, this is not consistent with the instructions. This is not a complete description of a problem. You will need to be more specific. A better example:

  I need an easier way to hold books in the winter time when I am wearing gloves.

  Example 2:

  I really need an LED light that can shine on the page and double as a bookmark.

No, this is not consistent with the instructions. This is an invention using an LED light to solve the person's problem. Try to think only about the problem – in this case the problem might be storing the reading light.

It is hard to find a convenient place to store a light with my book.

Example 3:

I wish books had pages that give fewer paper cuts.

Yes, this is acceptable. It is OK to describe a specific feature. The key is to focus only on features that result in a problem rather than on features that might be a solution or invention.

Click here to see an optional training video [1:40 min]

See the next page for the quiz.

- *Page 5*

  Step 3: Training Quiz

  You must identify which of the following are consistent with the HIT instructions in order to continue.

  Check here to Show the HIT instructions

  I need a way to open pages of a book so the words are easier to see near the binding. * Yes No

  I wish the book would stay open to the right page without constantly holding it. * Yes No

  I wish I could hold the book open with one hand without my fingers getting tired. * Yes No

  Easier to read. * Yes No

  A book with a photodetector light sensor to say when there was enough light to avoid eye strain. * Yes No

- *Page 6*

  Congratulations!

  Your answers are correct.

  You have completed the training HIT and now you can earn bonus payments. Your code will be shown when you are finished.

- *Page 7*

  Please tell us a little about yourself. This information is used to analyze responses from like participants. It is not used to identify you. Responses are optional.

  What is your age?

  - 18-25
  - 26-35
  - 36-45
  - 46-55
  - 56-65
  - 66+

  What is your gender? Male Female

  Even though experience with planning a trip is not required, we want to know more about how much experience you have. How much time do you spend during an average week?

  - None
  - Up to 5 hours
  - 5-10 hours
  - More than 10 hours

  How would you rate your expertise?

  - No experience
  - Novice
  - Intermediate
  - Expert
  - Professional

- *Page 8*

  How would you rate yourself on these general characteristics? Where would you place yourself on this scale of introversion (I keep to myself) to extroversion (I am very outgoing)?

  - Strongly Introverted
  - Moderately Introverted
  - In the Middle
  - Moderately Extroverted
  - Strongly Extroverted

  I am a very detail oriented person.

  - Strongly Agree
  - Agree
  - Neither Agree nor Disagree
  - Disagree

  - Strongly Disagree

  I am good at seeing the big picture.

  - Strongly Agree
  - Agree
  - Neither Agree nor Disagree
  - Disagree
  - Strongly Disagree

  I consider myself to be a gadget person.

  - Strongly Agree
  - Agree
  - Neither Agree nor Disagree
  - Disagree
  - Strongly Disagree

  I consider myself to be a people person.

  - Strongly Agree
  - Agree
  - Neither Agree nor Disagree
  - Disagree
  - Strongly Disagree

- *Page 9*

  Step 4: Problems and Needs

  Now you are ready to begin answering questions about [topic area].

  Dont worry about whether the benefit is worth the cost. We simply want lots of suggestions.

  There are opportunities to get help. You can watch a short video to explain the page. If you arent sure what to say click "I'm Stuck".

- *Page 10*

  Step 4: Problems and Needs

  Click here to view a 1 minute video about using this form.

  Need 1: Enter a problem you think should be solved or a need that should be met in order to make planning a trip a better experience.

  Enter only one problem or need, then click Enter Another before entering more. You earn a $.20 bonus each time you reach a total of 5 needs.

Enter One Need

Text box here

[optional] Tell a story that explains the setting or the background information that made you think of the problem above. You can earn an additional $.10 bonus for this story, but it must be a complete paragraph. Try to include as much context as possible, such as who, what, when, where, and why, but avoid identifying other individuals with real names.

Enter One Story

Text box here

Buttons: "Enter Another", "I'm Stuck"

- *Page 11*

  You might be running out of ideas at this point, but if you are willing to enter any more, you will earn a guaranteed $0.20 bonus for your next entry regardless of whether you reach 5.

  We hope you will read the short paragraph below and see if thinking in this way helps uncover more problems and needs.

  [Narrative prompt text]

  All possible options for the narrative prompt are described in Section A.1.3. For participants assigned to the control group, only the first line was displayed.

A.1.3 Narrative Prompt Options

Each prompt within the matrix was assigned a unique prompt ID and was also labeled with the matrix row, matrix column and matrix cell variation. Table A.3 shows each prompt ID and the location relative to each matrix row, column, and cell variation (these are cases with multiple prompts within a single cell).

For each prompt ID shown in Table A.3, the user interface included 3 unique wordings of the prompt to correspond to each topic area. Table A.4 gives verbatim text used for each prompt specific to each topic area.

[1] A Prompt ID of 1 was assigned to participants in the control group. This group was shown no prompt.

Table A.3: Prompt ID Numbers Assigned in Each Matrix Cell

			Matrix Columns		
			1st Person	Product	3rd Person
			1	2	3
Row Label	**Matrix Row**	**Cell Variation**	**Prompt ID**		
Emotion	1	1	2 [1]	3	4
		2		5	
Habits	2	1	6	7	8
		2	9		10
Communication	3	1	11	12	13
Uncertainty	4	1	14	15	16
		2	17		
Expertise	5	1	18	19	20
		2	21	22	23
		3	24		25
		4	26		
Technology	6	1	27	28	29
		2		30	

Table A.4: Prompt Text Specific to Each Topic Area

	Topic Area		
	Preparing food and cooking	Doing housecleaning and chores	Planning a trip
ID	**Complete Prompt Text**		
1	*Control - No prompt given	*Control - No prompt given	*Control - No prompt given
2	Think of a time while you were preparing food and cooking when you found yourself saying"why do I have to do this" or "I never want to do this again." Maybe you felt like you were wasting time. Or maybe you felt disappointed, frustrated, or confused. Maybe you had a reservation or fear or you simply wanted to skip the worst part. What were you doing at the time? Do these unpleasant experiences help identify specific problems?	Think of a time while you were doing housecleaning and household chores when you found yourself saying "why do I have to do this?" or "I never want to do this again." Maybe you felt like you were wasting time. Or maybe you felt disappointed, frustrated, or confused. Maybe you had a reservation or fear or you simply wanted to skip the worst part. What were you doing at the time? Do these unpleasant experiences help identify specific problems?	Think of a time while you were planning a trip when you found yourself saying "why do I have to do this?" or "I never want to do this again." Maybe you felt like you were wasting time. Or maybe you felt disappointed, frustrated, or confused. Maybe you had a reservation or fear or you simply wanted to skip the worst part. What were you doing at the time? Do these unpleasant experiences help identify specific problems?
3	Think of something you have used or wanted to use to prepare food and cook, but it was intimidating to try. Maybe it was an unfamiliar tool or process and you didn't think you would use it correctly. Maybe you expected to be criticized or expected to feel embarrassed. What was the tool or process that was intimidating? What problems are created if you feel this way?	Think of something you have used or wanted to use to do housecleaning and household chores, but it was intimidating to try. Maybe it was an unfamiliar tool or process and you didn't think you would use it correctly. Maybe you expected to be criticized or expected to feel embarrassed. What was the tool or process that was intimidating? What problems are created if you feel this way?	Think of something you have used or wanted to use to plan a trip, but it was intimidating to try. Maybe it was an unfamiliar tool or process and you didn't think you would use it correctly. Maybe you expected to be criticized or expected to feel embarrassed. What was the tool or process that was intimidating? What problems are created if you feel this way?

Continued on next page

Table A.4 – continued from previous page

ID	Cooking	Housecleaning	Planning a trip
4	Think of a time when have you seen or heard about other people trying to prepare food and cook the same way you do, and you know they don't like to do it. Maybe you have observed displeasure while watching people or you have heard complaints afterwards. What were problems that other people encountered?	Think of a time when have you seen or heard about other people trying to do housecleaning and household chores the same way you do, and you know they don't like to do it. Maybe you have observed displeasure while watching people or you have heard complaints afterwards. What were problems that other people encountered?	Think of a time when have you seen or heard about other people trying to plan a trip the same way you do, and you know they don't like to do it. Maybe you have observed displeasure while watching people or you have heard complaints afterwards. What were problems that other people encountered?
5	Think of something that you have used to prepare food and cook, but it made you really frustrated, irritated or even upset. You wanted to throw it in the garbage. What tools or steps of the task do you associate with this kind of frustration? What problems could be solved to make these less frustrating?	Think of something that you have used to do housecleaning and household chores, but it made you really frustrated, irritated or even upset. You wanted to throw it in the garbage. What tools or steps of the task do you associate with this kind of frustration? What problems could be solved to make these less frustrating?	Think of something that you have used to plan a trip, but it made you really frustrated, irritated or even upset. You wanted to throw it in the garbage. What tools or steps of the task do you associate with this kind of frustration? What problems could be solved to make these less frustrating?
6	Think of a time when you were preparing food and cooking and you needed to try something new. You tried it once (or a few times) but then stopped or decided to go back to what you usually do. Maybe you thought you would enjoy it and didnt, or it was something you thought you should do, but you decided it wasnt worth the effort. Or the old way is just better. What were you doing, and what was it that made you stop? What problem could be addressed to help you stick with the best possible approach?	Think of a time when you were doing housecleaning and household chores and you needed to try something new. You tried it once (or a few times) but then stopped or decided to go back to what you usually do. Maybe you thought you would enjoy it and didnt, or it was something you thought you should do, but you decided it wasnt worth the effort. Or the old way is just better. What were you doing, and what was it that made you stop? What problem could be addressed to help you stick with the best possible approach?	Think of a time when you were planning a trip and you needed to try something new. You tried it once (or a few times) but then stopped or decided to go back to what you usually do. Maybe you thought you would enjoy it and didnt, or it was something you thought you should do, but you decided it wasnt worth the effort. Or the old way is just better. What were you doing, and what was it that made you stop? What problem could be addressed to help you stick with the best possible approach?

Continued on next page

Table A.4 – continued from previous page

ID	Cooking	Housecleaning	Planning a trip
7	Think of something that you could potentially use to prepare food and cook, but you try to avoid it because it isn't the best way to get the job done. Maybe you end up taking shortcuts that aren't the best approach or the tools you are using make it more likely you will end up with sloppy results. What is it that you try to avoid and what problems can be solved to have the ability to improve your process?	Think of something that you could potentially use to do housecleaning and household chores, but you try to avoid it because it isn't the best way to get the job done. Maybe you end up taking shortcuts that aren't the best approach or the tools you are using make it more likely you will end up with sloppy results. What is it that you try to avoid and what problems can be solved to have the ability to improve your process?	Think of something that you could potentially use to plan a trip, but you try to avoid it because it isn't the best way to get the job done. Maybe you end up taking shortcuts that aren't the best approach or the tools you are using make it more likely you will end up with poor results. What is it that you try to avoid and what problems can be solved to have the ability to improve your process?
8	Think of a time when have you seen or heard about other people struggling to prepare food or cook because they aren't able to change an old habit. Maybe you have changed to use new or better tools, but others haven't. Maybe they don't realize it is not the best approach or maybe the habit is too ingrained and they haven't found a motivation to change. What is the habit, and how does this lead to problems?	Think of a time when have you seen or heard about other people struggling to do housecleaning and household chores because they aren't able to change an old habit. Maybe you have changed to use new or better tools, but others haven't. Maybe they don't realize it is not the best approach or maybe the habit is too ingrained and they haven't found a motivation to change. What is the habit, and how does this lead to problems?	Think of a time when have you seen or heard about other people struggling to plan a trip because they aren't able to change an old habit. Maybe you have changed to use new or better tools, but others haven't. Maybe they don't realize it is not the best approach or maybe the habit is too ingrained and they haven't found a motivation to change. What is the habit, and how does this lead to problems?
9	Think of a time when you were forced to change how you normally prepare food and cook because you were in a new situation or environment. Maybe you didn't have tools you usually use or available tools were unfamiliar. Maybe you were in a new place or with new people and had to improvise. Can you think of something that was difficult to change, and what problems could be solved to ease this transition?	Think of a time when you were forced to change how you normally do housecleaning and household chores because you were in a new situation or environment. Maybe you didn't have tools you usually use or available tools were unfamiliar. Maybe you were in a new place or with new people and had to improvise. Can you think of something that was difficult to change, and what problems could be solved to ease this transition?	Think of a time when you were forced to change how normally plan a trip because you were in a new situation or environment. Maybe you didn't have things you usually use or available tools were unfamiliar. Maybe you were in a new place or with new people and had to improvise. Can you think of something that was difficult to change, and what problems could be solved to ease this transition?

Continued on next page

Table A.4 – continued from previous page

ID	Cooking	Housecleaning	Planning a trip
10	Think of a time when have you seen or heard about other people who got a new gadget to help prepare food and cook but ended up not using it. Maybe the gadget wasn't what they expected or it didn't make the task easier. Maybe they thought familiar was more important than new or better. What problem was the gadget supposed to help the person overcome, and does this problem still need a new solution?	Think of a time when have you seen or heard about other people who got a new gadget to help do housecleaning and household chores, but ended up not using it. Maybe the gadget wasn't what they expected or it didn't make the task easier. Maybe they thought familiar was more important than new or better. What problem was the gadget supposed to help the person overcome, and does this problem still need a new solution?	Think of a time when have you seen or heard about other people who got a new gadget or app to help plan a trip, but ended up not using it. Maybe the gadget wasn't what they expected or it didn't make the task easier. Maybe they thought familiar was more important than new or better. What problem was the gadget supposed to help the person overcome, and does this problem still need a new solution?
11	Think of a time you when you had to explain to someone else something about preparing food and cooking, and you thought it was really difficult to explain clearly. Maybe you couldn't think of a good way to explain it or maybe you thought you were saying the right things but it wasn't getting across. What were you helping with, and what problem could be solved to help make this easier to explain or to make the process less complex?	Think of a time you when you had to explain to someone else something about doing housecleaning and household chores, and you thought it was really difficult to explain clearly. Maybe you couldn't think of a good way to explain something or maybe you thought you were saying the right things but it wasn't getting across. What were you helping with, and what problem could be solved to help make this easier to explain or to make the process less complex?	Think of a time you when you had to explain to someone else something about planning a trip, and you thought it was really difficult to explain clearly. Maybe you couldn't think of a good way to explain something or maybe you thought you were saying the right things but it wasn't getting across. What were you helping with, and what problem could be solved to help make this easier to explain or to make the process less complex?

Continued on next page

Table A.4 – continued from previous page

ID	Cooking	Housecleaning	Planning a trip
12	Think of something you wanted to use while preparing food and cooking, but you weren't sure how it worked and you couldn't get clear help from the instructions or other people. This could be anything that didn't seem obvious when you first tried it. Maybe you spent time trying to figure it out, but then you felt like you were stuck. What were you trying to figure out and what problem could be solved to make it easier?	Think of something that you wanted to use while doing housecleaning and household chores, but you weren't sure how it worked and you couldn't get clear help from the instructions or other people. This could be anything that didn't seem obvious when you first tried it. Maybe you spent time trying to figure it out, but then you felt like you were stuck. What were you trying to figure out and what problem could be solved to make it easier?	Think of something that you wanted to use while planning a trip, but you weren't sure how it worked and you couldn't get clear help from the instructions or other people. This could be anything that didn't seem obvious when you first tried it. Maybe you spent time trying to figure it out, but then you felt like you were stuck. What were you trying to figure out and what problem could be solved to make it easier?
13	Think of a time when you have seen other people preparing food and cooking, but they had stopped or waited because they were confused. Maybe they struggled to understand what they should do next or what approach would be best. Maybe they weren't getting enough help. What are some of the reasons why the next steps might have been unclear, and what problems could be solved to overcome this?	Think of a time when you have seen other people doing housecleaning and household chores, but they had stopped or waited because they were confused. Maybe they struggled to understand what they should do next or what approach would be best. Maybe they weren't getting enough help. What are some of the reasons why the next steps might have been unclear, and what problems could be solved to overcome this?	Think of a time when you have seen other people planning a trip, but they had stopped or waited because they were confused. Maybe they struggled to understand what they should do next. Maybe they weren't getting enough help. What are some of the reasons why the next steps might have been unclear, and what problems could be solved to overcome this?
14	Think of a time when you tried preparing food and cooking, and the result did not end up how you had hoped or wanted. You were expecting to get a certain result, but that isn't what happened. Can you identify any reasons why you didnt get the outcome you expected? What problem could be addressed to help get the outcome you wanted?	Think of a time when you tried doing housecleaning and household chores, and the result did not end up how you had hoped or wanted. You were expecting to get a certain result, but that isn't what happened. Can you identify any reasons why you didnt get the outcome you expected? What problem could be addressed to help get the outcome you wanted?	Think of a time when you tried to plan a trip, and the result did not end up how you had hoped or wanted. You were expecting to get a certain experience, but that isn't what happened. Can you identify any reasons why you didnt get the outcome you expected? What problem could be addressed to help get the outcome you wanted?

Continued on next page

Table A.4 – continued from previous page

ID	Cooking	Housecleaning	Planning a trip
15	Think of something you needed to use to prepare food and cook, but you were getting inconsistent results. Sometimes it would work and sometimes it wouldn't, even if you thought you were using it the same way each time. What was the thing that gave unexpected results? What problems could be solved to get the result you were wanting?	Think of something you needed to use to houseclean and do household chores, but you were getting inconsistent results. Sometimes it would work and sometimes it wouldn't, even if you thought you were using it the same way each time. What was the thing that gave unexpected results? What problems could be solved to get the result you were wanting?	Think of something you needed to use to plan a trip, but you were getting inconsistent results. Sometimes it would work and sometimes it wouldn't, even if you thought you were using it the same way each time. What was giving unexpected results? What problems could be solved to get the result you were wanting?
16	Think of a time when other people tried to help you while preparing food and cooking, but they weren't sure what you wanted or what they should do to help. This could be based on communication barriers, different points of view, or a general misunderstanding. What was difficult for other people to help with? What problem could be solved to help people know what you wanted?	Think of a time when other people tried to help you while doing housecleaning and household chores, but they weren't sure what you wanted or what they should do to help. This could be based on communication barriers, different points of view, or a general misunderstanding. What was difficult for other people to help with? What problem could be solved to help people know what you wanted?	Think of a time when other people tried to help you with planning a trip, but they weren't sure what you wanted or what they should do to help. This could be based on communication barriers, different points of view, or a general misunderstanding. What was difficult for other people to help with? What problem could be solved to help people know what you wanted?
17	Think of a time when you were preparing food and cooking, and you were unsure how to get the result you wanted. Maybe this was new for you or the things you were trying weren't working the way you expected. You didn't know the best way to move ahead or didn't know where to find the information you needed. What problem could be addressed to help you understand what to do?	Think of a time when you were doing housecleaning and household chores, and you were unsure how to get the result you wanted. Maybe this was new for you or the things you were trying weren't working the way you expected. You didn't know the best way to move ahead or didn't know where to find the information you needed. What problem could be addressed to help you understand what to do?	Think of a time when you tried to plan a trip, and you were unsure whether you would get the result you wanted. Maybe this was new for you or parts of the plan weren't coming together the way you expected. Maybe you didn't know where to find the information you needed. What problem could be addressed to help you understand what to do?

Continued on next page

Table A.4 – continued from previous page

ID	Cooking	Housecleaning	Planning a trip
18	Think of a time when you were preparing food and cooking, and you needed to stop and ask for help (or you wished you could). This could have been asking a friend or coworker, looking online, or making a call. What was the exact problem you needed help with?	Think of a time when you were doing housecleaning and household chores, and you needed to stop and ask for help (or you wished you could). This could have been asking a friend or coworker, looking online, or making a call. What was the exact problem you needed help with?	Think of a time when you were planning a trip, and you needed to stop and ask for help (or you wished you could). This could have been asking a friend or coworker, looking online, or making a call to customer service. What was the exact problem you needed help with?
19	Think of something you have used while preparing food and cooking where you repurposed it to meet your needs. You used something in a way that wasn't originally intended. It was a "work around" that got you what you wanted. You modified something or used it in a new way. What was it that you modified? What problems could be solved to make your new approach more available?	Think of something you have used while doing housecleaning and household chores where you repurposed it to meet your needs. You used something in a way that wasn't originally intended. It was a "work around" that got you what you wanted. You modified something or used it in a new way. What was it that you modified? What problems could be solved to make your new approach more available?	Think of something you have used while planning a trip where you repurposed it to meet your needs. You used something in a way that wasn't originally intended. It was a "work around" that got you what you wanted. What was it that you modified? What problems could be solved to make your new approach more available?
20	Think of a time when you have seen or heard about other people who prepare food and cook similar to you, but they seem to be doing it wrong? Maybe they have difficulty understanding how to do it correctly, or it is hard to measure if one way is better than another. Maybe information is lacking to help decide which way is best. What specific task can you think of, and what problems could be solved to provide better information?	Think of a time when you have seen or heard about other people who do housecleaning and household chores similar to you, but they seem to do it wrong? Maybe they have difficulty understanding how to do it correctly, or it is hard to measure if one way is better than another. Maybe information is lacking to help decide which way is best. What specific task can you think of, and what problems could be solved to provide better information?	Think of a time when you have seen or heard about other people who plan trips similar to you, but they seem to do something wrong? Maybe information is lacking to help decide which way is best. Maybe it is hard to measure if one way is better than another. What specific task can you think of, and what problems could be solved to provide better information?

Continued on next page

Table A.4 – continued from previous page

ID	Cooking	Housecleaning	Planning a trip
21	Think of a time when you thought you had a good idea of how long it would take to finish preparing food and cooking, but you ended up spending much longer than you had planned. Maybe you were trying something for the first time or you thought past experience would be a guide, but the new task turned out to be too different. Can you identify what step resulted in the delay? What problem could be addressed to reduce these kinds of delays?	Think of a time when you thought you had a good idea of how long it would take to finish doing housecleaning and household chores, but you ended up spending much longer than you had planned. Maybe you were trying something for the first time or you thought past experience would be a guide, but the new task turned out to be too different. Can you identify what step resulted in the delay? What problem could be addressed to reduce these kinds of delays?	Think of a time when you thought you had a good idea of how long it would take to plan a trip, but you ended up spending much longer than you had planned. Maybe you were trying something for the first time or you thought past experience would be a guide, but the new trip turned out to be too different. Can you identify what area resulted in the delay? What problem could be addressed to reduce these kinds of delays?
22	Think of something you have used when preparing food and cooking that was so intuitive that you didn't struggle to learn how to use it. You didn't need to read a manual and maybe you felt like it was designed in a way that focused on how users would use it. Have you ever wished something else could be this easy? What would you like to be as intuitive as your example of good design, and what problems do you deal with when a good design is missing?	Think of something you have used when doing housecleaning and household chores that was so intuitive that you didn't struggle to learn how to use it. You didn't need to read a manual and maybe you felt like it was designed in a way that focused on how users would use it. Have you ever wished something else could be this easy? What would you like to be as intuitive as your example of good design, and what problems do you deal with when a good design is missing?	Think of something you have used when planning a trip that was so intuitive that you didn't struggle to learn how to use it. You didn't need to read instructions and maybe you felt like it was designed in a way that focused on how users would use it. Have you ever wished something else could be this easy? What would you like to be as intuitive as your example of good design, and what problems do you deal with when a good design is missing?

Continued on next page

Table A.4 – continued from previous page

ID	Cooking	Housecleaning	Planning a trip
23	Think of a time when have you seen or heard about other people who prepare food and cook similar to you, but you can tell that they are doing it better than you are. This may simply be a case where other people have more experience or more training or maybe it seems like they are just a natural fit for the task. What makes you think other people do something better, and what are problems that could be solved to help get you to the same level?	Think of a time when have you seen or heard about other people who houseclean and do household chores similar to you, but you can tell that they are doing it better than you are. This may simply be a case where other people have more experience or more training or maybe it seems like they are just a natural fit for the task. What makes you think other people do something better, and what are problems that could be solved to help get you to the same level?	Think of a time when have you seen or heard about other people who plan trips similar to you, but you can tell that they are doing it better than you are. This may simply be a case where other people have more experience or or the right tools or maybe it seems like they have information you don't. What makes you think other people do something better, and what are problems that could be solved to help get you to the same level?
24	Think of a time when you were preparing food and cooking and you had to do the same thing more than once. Maybe the first time wasn't right or maybe the steps were mixed up and you had to back up and try again. Maybe after finishing you couldn't eat what you made and needed to make it again. What steps were you working on? What problems existed that resulted in needing to start again?	Think of a time when you were doing housecleaning and household chores and you had to do the same thing more than once. Maybe the first time wasn't right or maybe the steps were mixed up and you had to back up and try again. Maybe after finishing you undid what you had already finished and had to start again. What steps were you working on? What problems existed that resulted in needing to start again?	Think of a time when you were planning a trip and you had to do the same thing more than once. Maybe the first time wasn't right or maybe the steps mixed up and you had to back up and try again. Maybe after finishing you undid what you had already finished and had to start again. What steps were you working on? What problems existed that resulted in needing to start again?
25	Think of a time when you have seen or heard about other people who prepare food and cook similar to you, but these people are having difficulty and need help. Maybe they are less experienced or dont have the same background or have a disability. What kind of problems have you seen others experience or could you imagine others might experience?	Think of a time when you have seen or heard about other people who houseclean and do household chores similar to you, but these people are having difficulty and need help. Maybe they are less experienced or dont have the same background or have a disability. What kind of problems have you seen others experience or could you imagine others might experience?	Think of a time when you have seen or heard about other people who plan trips similar to you, but these people are having difficulty and need help. Maybe they are less experienced or dont have the same background or have a disability. What kind of problems have you seen others experience or could you imagine others might experience?

Continued on next page

Table A.4 – continued from previous page

ID	Cooking	Housecleaning	Planning a trip
26	Think of a time when you were preparing food and cooking and you thought, Oops, I shouldnt have done that. What made you realize later that you could have done something different? What problem could be solved to help you see this possibility from the beginning?	Think of a time when you were doing housecleaning and household chores and you thought, Oops, I shouldnt have done that. What made you realize later that you could have done something different? What problem could be solved to help you see this possibility from the beginning?	Think of a time when you were planning a trip and you thought, Oops, I shouldnt have done that. What made you realize later that you could have done something different? What problem could be solved to help you see this possibility from the beginning?
27	Think of a time when you first saw some gadget to help with preparing food and cooking and immediately wondered what the point of it was. It didn't seem useful to you at all. Maybe you eventually decided it had value or maybe not. What problem was this thing supposed to address? Do you think there are parts of the problem that need a better solution?	Think of a time when you first saw some gadget to help with doing housecleaning and household chores and immediately wondered what the point of it was. It didn't seem useful to you at all. Maybe you eventually decided it had value or maybe not. What problem was this thing supposed to address? Do you think there are parts of the problem that need a better solution?	Think of a time when you first saw some gadget or app to help with planning a trip and immediately wondered what the point of it was. It didn't seem useful to you at all. Maybe you eventually decided it had value or maybe not. What problem was this thing supposed to address? Do you think there are parts of the problem that need a better solution?
28	Think of something you needed to use while preparing food and cooking, and the product had such a bad design that you couldnt help but wonder if the designer had even bothered to use it. You knew there had to be a better way and couldnt imagine how the designer couldnt figure it out. Think about the product with the bad design and what it was supposed to do. What unmet need might still exist because these products were poorly designed?	Think of something you needed to use while doing housecleaning and household chores, and the product had such a bad design that you couldnt help but wonder if the designer had even bothered to use it. You knew there had to be a better way and couldnt imagine how the designer couldnt figure it out. Think about the product with the bad design and what it was supposed to do. What unmet need might still exist because these products were poorly designed?	Think of something you needed to use while planning a trip, and the product or service had such a bad design that you couldnt help but wonder if the designer had even bothered to use it. You knew there had to be a better way and couldnt imagine how the designer couldnt figure it out. Think about the product with the bad design and what it was supposed to do. What unmet need might still exist because these products were poorly designed?

Continued on next page

Table A.4 – continued from previous page

ID	Cooking	Housecleaning	Planning a trip
29	Think of a time when you have seen or heard about other people who tend to rely on the latest gadget or fad to help with preparing food or cooking, but you generally think they aren't doing any better even with all of these new things. Think about what the appeal might be of these gadgets. What problem do you think people hope to solve by buying them?	Think of a time when you have seen or heard about other people who tend to rely on the latest gadget or fad to help with doing housecleaning and household chores, but you generally think they aren't doing any better even with all of these new things. Think about what the appeal might be of these gadgets. What problem do you think people hope to solve by buying them?	Think of a time when you have seen or heard about other people who tend to rely on the latest gadget or app to help with planning a trip, but you generally think they aren't doing any better even with all of these new things. Think about what the appeal might be of these gadgets. What problem do you think people hope to solve by buying them?
30	Think of something you needed to use while preparing food and cooking, and it just stopped working or wore out. Maybe it needed repaired or replaced. Maybe it was very new or just past the warranty. It could be disposable or a top brand, but it eventually failed or broke when you still needed it. What were you trying to use, and what problems might have been created when it stopped working?	Think of something you needed to use while doing housecleaning and household chores, and it just stopped working or wore out. Maybe it needed repaired or replaced or restarted. Maybe it was very new or just past the warranty. It could be disposable or a top brand, but it eventually failed or broke when you still needed it. What were you trying to use, and what problems might have been created when it stopped working?	Think of something you needed to use while planning a trip, and it just stopped working. Maybe it needed repaired or replaced or restarted. Maybe it was very new or just past the warranty, but it eventually failed when you still needed it. What were you trying to use, and what problems might have been created when it stopped working?

A.2 Part 1: Quantity Study 2 Additional Material

The exact content of study 2 is included in this section. Actual content is reflected in normal font, and editorial information is *provided in italics.* Not all text formatting used in the final user interface is reflected in this section. Much of the content is identical to Study 1, and is noted as such.

A.2.1 Amazon Mechanical Turk Interface for Study 2

- *Title for the HIT*

  Describe problems with common products and services

- *Summary description shown to AMT workers*

  This is a research study where you can submit examples of problems with common products and services. You will see the specific topic after you begin. Write more to earn more. Earn frequent bonuses of about $.05 per sentence.

- *Keywords entered into AMT*

  needs problems study research survey product bonus creative brainstorm

- *Main body of the HIT displayed above the box to paste in the completion code*

  Tell us how products and services you use haven't worked like they should.

  The steps include:

  - Review instructions
  - Review training activity ($ 0.65 payment)
  - If you pass the training quiz: Describe problems related to your topic to earn bonuses

  Responses are open-ended and after every response, you are eligible for $ 0.05 bonus payment. Earn additional bonus payments of $ 0.15 if you give detailed reasons for your responses. Minimum participation times will be approximately 5-15 minutes and you are encouraged to continue beyond this and earn additional payments as long as you would like. Incomplete or duplicate responses may not be approved.

  Accept this HIT if you are looking for a creative challenge. We have heard very positive feedback before: This was a very interesting survey/ exercise which made me exercise my creative muscles! Interesting task. Actually made me consider the real problems I face and think more about what solutions I could come up with on my own." Interesting brainstorm activity.

  We pay bonuses honestly and promptly. Check out turker forums.

Warning: Do not use browser reload or back buttons as these may cause errors and prevent a code from displaying.

Survey link: Open survey in new tab

A.2.2 Zoho Creator App for Study 2

- *Page 1*

  See Study 1, Section A.1.

- *Page 2*

  See Study 1, Section A.1, changes are included below.

  If you agree to be in this study, we would ask you to do the following things: The steps include:

  - Review instructions
  - Review training activity ($.65 standard payment)
  - If you pass the training quiz: Describe problems related to your topic to earn bonuses

  Responses are open-ended and after every response, you are eligible for a $ .05 bonus payment. Earn additional bonus payments of $ .15 if you give detailed reasons for your responses. Minimum participation times will be approximately 5-15 minutes and you are encouraged to continue beyond this and earn additional payments as long as you would like. Incomplete or duplicate responses may not be approved.

- *Page 3*

  See Study 1, Section A.1, changes are included below.

  The instructions for this HIT include:

  1. Clearly and completely describe the problem, ideally in a complete sentence.
  2. You should NOT include inventions or ideas to solve the problem. You should only list the problem itself.

  In short: more than a few words (good) and a problem (good) instead of a solution (not good).

- *Page 4*

  See Study 1, Section A.1.

- *Page 5*

  See Study 1, Section A.1, changes are included below.

  A book with an eye strain meter built into the cover using a photodetector light sensor. * Yes No

  The wording of quiz question 5 was modified to remove the mention of a need (detecting eye strain) and use wording with more emphasis on the invention.

- *Page 6*

  See Study 1, Section A.1.

- *Page 7*

  See Study 1, Section A.1.

  Two additional questions were included on Page 7.

  I consider myself to be a people person.

  - Strongly Agree
  - Agree
  - Neither Agree nor Disagree
  - Disagree
  - Strongly Disagree

  I am good at seeing the big picture.

  - Strongly Agree
  - Agree
  - Neither Agree nor Disagree
  - Disagree
  - Strongly Disagree

- *Page 8*

  Page 8 was skipped.

- *Page 9*

  See Study 1, Section A.1.

- *Page 10a*

  Page 10 was displayed with 2 frames on the browser. Pages 10a and 11e were on the left side and shown at the beginning and after clicking "I'm Stuck" a second time, respectively. Pages 10b and 10c were displayed on the right side at the beginning, and after clicking "I'm Stuck", respectively.

Step 4: Problems and Needs

On the right side, enter a problem you think should be solved or a need that should be met in order to make null a better experience.

[optional] Tell a story that explains the setting or the background information that made you think of the problem above. To earn a bonus, it must be a complete paragraph. Try to include as much context as possible, such as who, what, when, where, and why, but avoid identifying other individuals with real names.

You earn* a $ .05 bonus for each need entry and a $ .15 bonus for each story.

- *Page 10b*

  Enter Need 1 Below

  Click here to view a 1 minute video about using this form.

  Enter only one problem or need, then click Enter Another.

  Enter One Need *Text box here*

  Enter A Story About This Need *Text box here*

  Buttons: "Enter Another", "I'm Stuck"

- *Page 10c*

  Enter Need 1 Below

  Click here to view a 1 minute video about using this form.

  Enter only one problem or need, then click Enter Another.

  Enter One Need

  Text box here

  Enter A Story About This Need

  Text box here

  Help A: View images

  Help B: View needs and stories

  Help C: View a prompt

  Buttons: "Enter Another", "Help A", "Help B", "Help C"

- *Page 11a*

  Pages 11a-11d were used to display the 4 treatments on the left side frame. These were the Control, Prompt, Shared Needs, and Shared Images, respectively. Page 11e was a summary page explaining how to view these help screens.

  Step 4: Problems and Needs

  You might be running out of ideas at this point, but if you are willing to enter any more, you will earn a double bonus for your next need entry.

- *Page 11b*

  Step 4: Problems and Needs

  You might be running out of ideas at this point, but if you are willing to enter any more, you will earn a double bonus for your next need entry. *Paragraph was hidden after the first assigned help.*

  We hope you will read the short paragraph below and see if thinking in this way helps uncover more problems and needs.

  [Narrative prompt text]

  All possible options for the narrative prompt are described in Section A.1.3.

- *Page 11c*

  Step 4: Problems and Needs

  You might be running out of ideas at this point, but if you are willing to enter any more, you will earn a double bonus for your next need entry. *Paragraph was hidden after the first assigned help.*

  We hope you will review the list below of stories submitted by other participants. Think of this step like brainstorming using this list as inspiration. Try to think of new needs related to these stories or anything that comes to mind.

  [List of 10 need statements alternating with 10 stories]

- *Page 11d*

  Step 4: Problems and Needs

  You might be running out of ideas at this point, but if you are willing to enter any more, you will earn a double bonus for your next need entry. *Paragraph was hidden after the first assigned help.*

  We hope you will review the list below of images submitted by other participants. Think of this step like brainstorming using this list as inspiration. Try to think of new needs related to these images or anything that comes to mind.

  [List of 10 images]

- *Page 11e*

  Step 4: Problems and Needs

  The first time you clicked "I'm Stuck" you were shown information selected at random for our study.

  Now you are able to choose exactly what type of information you think might help you earn bonuses the fastest.

  Your choices:

Help A: 10 images showing products, services, and problems previously submitted by other users.

Help B: 10 needs and stories (like what you are entering on the right) previously submitted by other users.

Help C: A Narrative prompt with a paragraph description to think about a specific type of need related to *replaced by topic area phrase*.

Please select one to the right. You will be able to view as many as you would like, each help is different.

A.3 Part 1: Quantity Study 3 Additional Material

The exact content of study 3 is included in this section. Actual content is reflected in normal font, and editorial information is *provided in italics.* Not all text formatting used in the final user interface is reflected in this section. Much of the content is identical to previous studies, and is noted as such.

A.3.1 Amazon Mechanical Turk Interface for Study 3

See Study 2, Section A.2.

A.3.2 Zoho Creator App for Study 3

- *Page 1*

  See Study 1, Section A.1.

- *Page 2*

  See Study 2, Section A.2.

- *Page 3*

  See Study 2, Section A.2.

- *Page 4*

  See Study 1, Section A.1.

- *Page 5*

  See Study 2, Section A.2, changes are included below.

  My fingers get tired when I am holding a heavy book and reading in bed or a reclining chair. * Yes No

  The wording of quiz question 3 was modified to state the core problem (fingers getting tired at the beginning.

- *Page 6*

  See Study 1, Section A.1.

- *Page 7*

  See Study 1, Section A.1.

- *Page 8*

  Page 8 was skipped.

- *Page 9*

  See Study 1, Section A.1.

- *Page 10a*

  Page 10 was displayed with 2 frames on the browser. Page 10a was on the left side and shown at the beginning. Page 10c was displayed on the right side.

  Step 4: Problems and Needs

  See Study 2, Section A.2, changes are included below.

  When you are ready for help, you have 3 choices on the right. Each relate to [topic].

  Please select one to the right. You will be able to view as many as you would like, each help is different. This is not required, but we suggest trying each type once and then return to what seems most helpful.

- *Page 10b*

  Page 10b was skipped.

- *Page 10c*

  Enter Need 1 Below

  Click here to view a 1 minute video about using this form.

  Enter only one problem or need, then click Enter Another.

  Enter One Need

  Text box here

  Enter A Story About This Need

  Text box here

  Help A: View images of products & services uploaded by other users

  Help B: View needs and stories previously submitted by other users

  Help C: View a prompt to think about a specific type of need

  Buttons: "Enter Another", "Help A", "Help B", "Help C"

- *Page 11a*

  Page 11a was skipped.

- *Page 11b*

  Pages 11b-11d were used to display the 3 treatments on the left side frame. These were the Prompt, Shared Needs, and Shared Images, respectively.

  Step 4: Problems and Needs

We hope you will read the short paragraph below and see if thinking in this way helps uncover more problems and needs.

[Narrative prompt text]

All possible options for the narrative prompt are described in Section A.1.3.

- *Page 11c*

  Step 4: Problems and Needs

  We hope you will review the list below of stories submitted by other participants. Think of this step like brainstorming using this list as inspiration. Try to think of new needs related to these stories or anything that comes to mind.

  [List of 10 need statements alternating with 10 stories]

- *Page 11d*

  Step 4: Problems and Needs

  We hope you will review the list below of images submitted by other participants. Think of this step like brainstorming using this list as inspiration. Try to think of new needs related to these images or anything that comes to mind.

  [List of 10 images]

- *Page 11e*

  Page 11e was skipped.

Appendix B

Part 2: Uniqueness

B.1 Pilot 1 Additional Material

The exact content of the second pilot is included in this section. Actual content is reflected in normal font, and editorial information is *provided in italics.* Not all text formatting used in the final user interface is reflected in this section. Much of the content is identical to Study 1, and is noted as such.

The pilot study was repeated with two similar rating systems applicable to different training purposes for the semantic similarity algorithm. Pilot 1a evaluated a list of sentence pairs for similarity on a scale of 0-5, and Pilot 1b evaluated a list of sentence pairs using a binary choice of "equivalent" or "not equivalent".

B.1.1 Amazon Mechanical Turk Interface

- *Title for the HIT*

  Rate the similarity of two sentences.

- *Summary description shown to AMT workers*

  Pilot 1a This is a research study where you rate the similarity of two sentences on a scale of 0 to 5. Rate 10 sentence pairs for $.30.

  Pilot 1b This is a research study where you rate the meaning of two sentences as "equivalent" or "not equivalent". Rate 10 sentence pairs for $.25.

- *Keywords entered into AMT*

  needs, problems, study, research, survey, product, similarity, sentences, semantic

- *Main body of the HIT displayed above the box to paste in the completion code*

  Pilot 1a and 1b

  Rate the similarity of two sentences.

  The steps include:

  - Review instructions
  - Rate the similarity of two sentences (total of 10 pairs)

  Each sentence describes a problem that someone experiences that relates to cooking, cleaning, or planning a trip.

  Pilot 1a

  Some pairs will be identical, some will be similar, and some will be very different. You will need to carefully review the options for ratings and make judgments about the meaning of the sentences.

  Pilot 1b

You will need to make a decision whether or not the sentences have equivalent meaning, even if the same words are not used.

Warning: Do not use browser reload or back buttons as these may cause errors and prevent a code from displaying.

Survey link: Open survey in new tab

B.1.2 Zoho Creator App for Pilot 1a

- *Page 1*

  See Study 1, Section A.1 content for Page 1, changes are included below.

  Procedures:

  If you agree to be in this study, we would ask you to do the following things:

  - Review instructions
  - Rate the similarity of two sentences

  You are eligible for a payment of $ 0.30 for rating 10 pairs of sentences.

  You are allowed to complete multiple HITs if available.

- *Page 2*

  You will be rating the similarity of pairs of sentences. Some will be identical, some will be similar, and some will be very different.

  Each sentence describes a problem that someone experiences that relates to cooking, cleaning, or planning a trip.

  Please use the rating scale provided below.

  (5) The two problems are completely equivalent, as they mean the same thing.

  A way to stop from tearing up when one is cutting onions. I wish onions did not make my eyes water and burn so bad when cutting them.

  (4) The two problems are mostly equivalent, but one might be more specific than the other.

  I need to know what attractions in a city meet my interests. I need a way to find odd attractions along the way to our main destination.

  Note: both ask for attractions, but only one is looking for "odd" attractions.

  (3) The two problems are roughly equivalent, but have important differences.

  I wish I had a vacuum cleaner that is more portable on a multi level house. I wish my vacuum was lighter.

Note: portability could be a problem because of reasons other than weight (such as size).

(2) The two problems are not equivalent, but are similar in specific ways.

My hands get dry from dipping them in the bucket of water I use to clean my floors. My knees hurt when I kneel down to wash the floor.

Note: both relate to washing, specifically washing the floor.

(1) The two problems are not equivalent, and are only similar in very general ways.

Older people who are not comfortable with the internet often do not understand how to use it to compare airfare prices. It is often difficult to know where the best prices for gas are in an unknown location.

Note: both relate to prices but are otherwise very different.

(0) The two problems are completely unrelated.

Milk overflows so fast when being boiled. I have a lot of useful storage containers, but I need a good way of storing them.

- *Page 3*

  Sentence Pair 1:

  [Randomly assigned sentence 1]

  [Randomly assigned sentence 2]

  Select the similarity for sentence pair 1

  - (5) Completely equivalent
  - (4) Mostly equivalent
  - (3) Roughtly equivalent
  - (2) Similar in specific ways
  - (1) Only similar in very general ways
  - (0) Completely unrelated

  Repeat for Sentence Pairs 2-10

- *Page 3 instructions sidebar*

  Step 2: Rating Sentence Pairs

  [checkbox] Click to display examples for each rating option. *Displays full instructions from Page 2 if checked*

  - (5) The two problems are **completely equivalent**, as they mean the same thing.

- (4) The two problems are **mostly equivalent**, but one might be more specific than the other.
- (3) The two problems are **roughly equivalent**, but have important differences.
- (2) The two problems are **not equivalent, but are similar in specific ways.**
- (1) The two problems are **not equivalent, and are only similar in very general ways.**
- (0) The two problems are **completely unrelated.**

B.1.3 Zoho Creator App for Pilot 1b

- *Page 1*

  See Study 1, Section A.1 content for Page 1, changes are included below.

  Procedures:

  If you agree to be in this study, we would ask you to do the following things:

  - Review instructions
  - Rate the similarity of two sentences

  You are eligible for a payment of $ 0.25 for rating 10 pairs of sentences.

  You are allowed to complete multiple HITs if available.

- *Page 2*

  Step 1: Review Instructions

  You will be rating the similarity of pairs of sentences. You will need to make a decision whether or not the sentences have equivalent meaning, even if the same words are not used.

  Each sentence describes a problem that someone experiences that relates to cooking, cleaning, or planning a trip.

  Sentences must be rated as either "equivalent" or "not equivalent" , even if they might seem to be in between. Different people will answer differently, that is OK.

  An equivalent pair of sentences would meet one of the following conditions:

  - Describes the same problem, even in different words.
  - Enough similarity to suggest each person submitting the sentence might have been thinking about the same problem.

  Example of Equivalent sentences:

  - 1: I need a better way to keep the vacuum cleaner cord from getting tangled up.
  - 2: I wish I could use my vacuum cleaner without stopping so often to fix the cord.

- *Page 3*

  Sentence Pair 1:

  [Randomly assigned sentence 1]

  [Randomly assigned sentence 2]

  Select the similarity for sentence pair 1

  - (1) Equivalent
  - (0) Not equivalent

  Repeat for Sentence Pairs 2-10

- *Page 3 instructions sidebar*

  Step 2: Rating Sentence Pairs

  [checkbox] Click to display examples for each rating option. *Displays full instructions from Page 2 if checked*

  An equivalent pair of sentences would meet one of the following conditions:

  - Describes the same problem, even in different words.
  - Enough similarity to suggest each person submitting the sentence might have been thinking about the same problem.

Appendix C

Images Collection Pilot

C.1 Pilot 2 Additional Material

The exact content of the second pilot is included in this section. Actual content is reflected in normal font, and editorial information is *provided in italics.* Not all text formatting used in the final user interface is reflected in this section. Much of the content is identical to Pilot 1, and is noted as such.

C.1.1 Amazon Mechanical Turk Interface for Pilot 2

- *Title for the HIT*

  Upload a photo (cell phone is OK) or screen capture of products you use in your home.

- *Summary description shown to AMT workers*

  First Launch of Pilot 2 This is a research study where you create and upload a new digital image (photograph or screen capture) of a product or website you use in your home. You will be given a specific topic area when you begin (such as cooking).

  Second Launch of Pilot 2 This is a research study where you create and upload a new digital image (photograph or screen capture) relating to a common household topic. You will be given a specific topic area when you begin (such as cooking).

- *Keywords entered into AMT*

  study survey research image photograph product software

- *Main body of the HIT displayed above the box to paste in the completion code*

  First Launch of Pilot 2 This research will collect visual information about the types of products people use in the home. The HIT should take approximately 5-10 minutes.

  Second Launch of Pilot 2 This research will collect visual information about common household activities. The HIT should take approximately 5-10 minutes.

  The steps include:

  - Review instructions, including the type of product for your image (such as cooking), and formatting requirements
  - Create a new image (not downloaded from the internet)
  - Format the image to be less than 1 MB
  - Upload the image

  Warning: Do not include identifiable information in the image, such as names, addresses, or people. You will receive basic instructions on image size formatting,

but if you are not able to create an image less than 1MB, you cannot complete the HIT.

There will be more than one HIT available (The topic will be different each HIT).

Make sure to leave this window open as you complete the survey. When you are finished, you will return to this page to paste the code into the box.

Survey link: Open survey in new tab

C.1.2 Zoho Creator App for Pilot 2

- *Page 1*

  See Pilot 1, Section B.1 content for Page 1, changes are included below.

  The steps include:

  - Review instructions
  - Take a digital photograph of a product(s) or object in your surroundings
  - Upload the digital image

  You will earn a payment of $ 0.60 for completing the study. Submitting an out-of-focus or illegible image or a photograph you did not take may not be approved.

- *Page 2*

  Page 2 begins with calculated text describing the assigned topic area.

  You will need to upload an image relating to: [topic area].

  [topic area description]

  Your photograph should show something you dislike about [topic area].

  The image you upload must meet the following criteria:

  - Take a close-up photograph related to your topic.
  - A digital image can be a photograph of an object or screen capture of a website you use, not an image downloaded off the internet.
  - Do not include any identifying information (i.e. a person, name, address).
  - Use landscape orientation (instead of portrait).
  - The image must be in focus and bright enough to clearly see what is shown.
  - The image file size must be less than 1MB (a 1280 x 1024 or similar size preferred).

  Click to see instructions on changing the image file size.

  Mac users:

- Open the image using "Preview" software.
- On the top menu select Tools ¿ Adjust Size.
- Make sure the box for "Scale Proportionally" is checked.
- Select "Fit Into" 1280 x 1024 or similar size. This will be much less than 1 MB.

PC users:

- Open the image using "Paint" software.
- On the Home tab, in the Image group, click Resize.
- In the Resize and Skew dialog box, Choose "Pixels" and select the Maintain aspect ratio check box.
- Enter a new value for width (or height), such as 1280.

Click to see instructions on creating a screen capture of a website.

Mac users:

- Open the "Grab" software.
- On the top menu select Capture ¿ Selection.
- Click your mouse and drag a box around the website window.
- Save the file (.tiff). Converting to .jpeg is optional. Screen captures are usually less than 1 MB.

PC users:

- Open the "Snipping Tool" software.
- On the "New" tab, select "Rectangular Snip".
- Click your mouse and drag a box around the website window.
- Save the file (.jpeg). Screen captures are usually less than 1 MB.

Use the button below to navigate to the image, or drag and drop the image in the box.

[File upload box]

Appendix D

Part 3: Quality

D.1 Quality Study Additional Material

The exact content of the Part 3: Quality study is included in this section. Actual content is reflected in normal font, and editorial information is *provided in italics.* Not all text formatting used in the final user interface is reflected in this section. Much of the content is identical to previous studies, and is noted as such.

D.1.1 Amazon Mechanical Turk Interface for Quality Study

- *Title for the HIT*

  Rate the importance of problems with common products and services

- *Summary description shown to AMT workers*

  This is a research study where you rate if problems submitted by other people are important to you. Topics may include areas such as cooking, cleaning, or travel. Answer 2-3 questions about 10 problems for $.50.

- *Keywords entered into AMT*

  needs, problems, study, research, survey, product, rate, important

- *Main body of the HIT displayed above the box to paste in the completion code*

  Rate if problems submitted by other people are important to you. Topics may include cooking, cleaning, or travel.

  The steps include:

  - Review instructions
  - Answer 5 questions about yourself [no identifying information]
  - Rate the importance of a list of problems you might face doing common activities (2-3 questions ea.)

  You are eligible for a payment of $.50 for rating 10 problem statements.

  You are allowed to complete multiple HITs if available.

  Warning: Do not use browser reload or back buttons as these may cause errors and prevent a code from displaying.

  Survey link: Open survey in new tab

D.1.2 Zoho Creator App for Quality Study

- *Page 1*

  See Study 1, Section A.1 content for Page 1, changes are included below.

  Consent Information:

The information you provide going forward will be stored and used for the research study. Review the information below and indicate you agree to participate if you would like to continue. You are invited to be in a research study to rate the importance of a list of problems you might face doing common activities. You were selected as a possible participant because you selected this HIT on Amazon Mechanical Turk. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

Procedures:

If you agree to be in this study, we would ask you to do the following things:

- Review instructions
- Answer 5 questions about yourself [no identifying information]
- Rate the importance of a list of problems you might face doing common activities (2-3 questions ea.)

You are eligible for a payment of $.50 for rating 10 problem statements.

You are allowed to complete multiple HITs if available.

- *Page 2*

  Step 1: Review Instructions

  You will be rating the importance of a list of problems submitted by other people. The ratings help prioritize which problems could be solved to help the most people. There are no right or wrong answers. The final list will relate to the following topic: planning a trip. This includes any travel beyond daily routines. Trips might be work-related, vacations, local, abroad, by yourself, or with others.

  1: Read the problem statement and the additional full description, if one is provided.

  2: Check if statements meet basic requirements. You should flag a statement if it:

  - Already describes a SOLUTION, not a problem or need.
  - Has unclear meaning.

  Examples (these relate to cooking, your final list might be a different topic): Ex. A. I need vegetables to be pre-chopped when I buy them. This is a solution - the statement should have focused on a problem, such as Chopping vegetables is too time consuming.

  Ex. B. I need a way to perform multiple steps easier. If there is not enough detail, it is unclear and would be a poor use of time to rate.

  Ex. C. Milk overflows so fast when being boiled. OK. This is not a solution, and should be clear even if the problem does not apply to you.

3: Rate remaining problems on two criteria:

- How important is this problem to you? [1 = Unimportant, 5 = Very Important]
- How satisfied are you with existing solutions to this problem? [1 = No Solutions or Very Unsatisfied, 5 = Very Satisfied]

[Continue]

- *Page 3*

  One new demographics question was added. Different answer choices were presented to participants in each topic.

  Check any descriptions that apply to you.

  Cooking

  - Family member with diet restrictions
  - Cook for small children
  - Cook for large family
  - Enjoy healthy cooking
  - None

  Cleaning

  - Pet owner
  - Small children at home
  - Teenagers at home
  - Carpeted flooring
  - Wood/linoleum flooring
  - None

  Travel

  - Business traveler
  - Travel for out-of-town family
  - Travel for vacations
  - Travel with children
  - International travel
  - None

- *Page 4*

  Step 3: Rate the importance of problem statements for planning a trip.

  [checkbox] Click to display complete instructions.

Flag statements if they do not meet basic requirements. Rate remaining problems on two criteria:

- How important is this problem to you? [1 = Unimportant, 5 = Very Important]
- How satisfied are you with solutions to this problem that already exist? [1 = No Solutions or Very Unsatisfied, 5 = Very Satisfied]

Need Statement [X]:

Full length need statement and story (if available). Labels for "X" were itemized 1 to 10

[checkbox] Flag problem statement [X] if unclear or a solution

If flagged, the flag detail question was displayed

Select the reason for a flag.

- Provides a Solution, not a problem or need
- Unclear meaning

If NOT flagged, importance and satisfaction questions were displayed

Rate the importance of Problem [X] to you.

- (5) Very Important
- (4) Important
- (3) Moderately Important
- (2) Of Little Importantce
- (1) Unimportant

Rate your satisfaction with existing solutions to Problem [X].

- (5) Very Satisfied
- (4) Satisfied
- (3) Neutral
- (2) Unsatisfied
- (1) No Solutions or Very Unsatisfied
- *Page 5*

Thank you for your participation! We are interested in your feedback for this task. Please share your comments below.