Assessing quality of unmet user needs: Effects of need statement characteristics

Cory R. Schaffhausen and *Timothy M. Kowalewski*, Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN 55455, USA

We demonstrate a front-end, user-centered method to prioritize unmet needs previously generated from large groups. Several hypotheses were tested: (1) Needs submitted first will be less likely to be high quality than needs submitted after a sustained period of time; (2) Semantically similar need statements will be rated as equivalent in quality; (3) Need statements will be rated as higher quality if a detailed description of the need context was available. Over 20 000 ratings for 1697 statements across three common product areas were analyzed. The results showed needs that first come to mind are not lower quality than needs that come to mind later and can inform early design phases to balance in-depth research and size of user groups.

Keywords: design methods, research methods, user centered design, user participation, needfinding

Developing successful products and services is unpredictable, but the process is aided by a thorough understanding of the needs of target users. The needs of any group can be a complex combination of technical, personal, and emotional content, and this is especially evident for large, diverse groups. Common needfinding methods often rely on in-depth or immersive interactions (e.g. interviews or observations) with a small sample of users, and often users are experts. In environments such as health care, where immersive study is particularly challenging, this approach remains commonly used in spite of limitations for reflecting the diversity of the user group (Martin & Barnett, 2012; Money et al., 2011). While soliciting needs from large groups might better reflect user diversity, one disadvantage is the subsequent steps to review and prioritize long lists of candidate needs.

This paper summarizes a web-based process to rapidly screen for high-quality needs out of a data set from previous web-based user needs collection. The high-quality needs from this screening may be candidates for further refinement during the project definition phase of the design process. A series of analyses measure the effects of need statement characteristics on quality ratings. Understanding what characteristics might facilitate capturing high-quality needs (directly articulated by users) will be valuable guidance for needfinding methods

Corresponding author:
Cory R. Schaffhausen
schaf390@umn.edu

www.elsevier.com/locate/destud
0142-694X *Design Studies* **44** (2016) 1–27
http://dx.doi.org/10.1016/j.destud.2016.01.002

and may ultimately reduce uncertainty in the project definition. However, little quantitative evidence exists for this process. The quality-rating studies included participants discussing common topics of cooking, cleaning, and travel. The analyses herein test new hypotheses to fill existing gaps pertaining to need statement characteristics and the results validate novel, web-based methods as a means to implement early-stage needfinding or user research phases.

1 Background

1.1 Overview of needfinding

Needfinding is a process of capturing input for unmet needs of product and service users. The needs input can inform early development phases and subsequently be translated into requirements for features (Bayus, 2008; Patnaik, 2014; Patnaik & Becker, 1999; Ulrich & Eppinger, 2004). This early phase of design is described in many varying terms, partly dependent on the discipline of origin. Generally this can be described as the 'establishing a need' phase (Howard, Culley, & Dekoninck, 2008). There is not consensus on terms and bounds of each phase, so 'needfinding' is used here and elements of this process are described below.

In order to minimize bias originating from the design team, needfinding should go straight to the group of users itself. Product failures can often be traced to a faulty over-reliance on input from the design team or company managers rather than information directly validated with users (Kelley & Littman, 2001). Validating these assumptions often requires prolonged engagement to develop a deep understanding of the users' actual behavior, because actions can differ from what is said. While validating user statements and needs can rely on qualitative observational data, additional quantitative methods have been evaluated to contribute to the early needs assessment and prioritization (Schaffhausen & Kowalewski, 2015a; Ulwick, 2002, Ulwick, 2005). The objective of needfinding is to be purposefully agnostic of solutions. A need statement can be more beneficial if it does not include embedded solutions. In this case, an embedded solution might be an invention, but this invention might be one of many alternatives to solve an underlying unmet need. Describing a solution too early can short-circuit the process of carefully defining a problem and thoroughly evaluating potential solutions (Patnaik & Becker, 1999; Zenios et al., 2010).

Any engagement with users also facilitates empathy for users, and empathy is critical for recognizing the needs and differing perspectives of users (Alkaya, Visser, & De Lille, 2012; Herriott & Jensen, 2013; Johnson et al., 2014; Kelley & Littman, 2001; Kouprie & Visser, 2009). Direct observation can have a particularly lasting influence on empathy in the observer (Patnaik, 2009), yet information on user needs can come from many sources. Direct

statements from users is one source. Users might struggle to articulate their own needs; however, improving the outcome for user-articulated needs, in particular for large groups, is a promising area for research. While the present work focuses on unmet needs, others have similarly argued for an increase in large-scale research of product use in general (Margolin, 1997). Large-scale needs and product use research may find the same success as open innovation has shown for ideation (Boudreau & Lakhani, 2013; Brabham, 2008; Enkel, Gassmann, & Chesbrough, 2009; Poetz & Schreier, 2012). Faste explicitly states that advances in online knowledge management 'could be applied to crowdsourced needfinding research' (Faste, 2011, 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.' (Faste, 2011, p. 4).

The analyses herein relate to the quality of need statements, but definitions of need statements often vary. While engineering texts often define a need statement as it relates to defining a final product attribute or requirement (Ulrich & Eppinger, 2004), Ulwick captures the current variability in the term 'requirements' and points out that companies discuss requirements and include 'needs, wants, solutions, benefits, ideas, outcomes, and specifications, and they often use these terms synonymously' (Ulwick, 2005, 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 (Ulwick, 2002, 2005). This is consistent with a focus on problems rather than desires (Matzler & Hinterhuber, 1998) and is likely to decrease the uncertainty of Voice of the Customer research outcomes (Ulwick, 2005).

A very broad sense use of the word 'needs' is assumed for this work, 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 (Patnaik, 2014; Patnaik & Becker, 1999). As described in Section 2, the need statements included in these analyses were collected with an explicit instruction to describe problems users face when performing common tasks or using common products, and to avoid embedded solutions (e.g. a new feature or invention).

1.2 Quantity focus

Increasing the size of a group of users can increase the likelihood of identifying a rarely articulated need. Griffin and Hauser performed consumer products-based studies focused on the quantity of needs collected during in-depth interviews and analyzed effects of increasing 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 (Griffin & Hauser, 1993). This study was limited to interviews

and discusses the quality of resulting needs primarily based on importance. The same quantity focus has been used when providing users with a variety of stimulus types to increase need quantity and quality. When calculating quality based on user ratings of importance and satisfaction, the number of high-quality needs increases with quantity for both individuals and groups (Schaffhausen & Kowalewski, 2015a).

Of note, this result is consistent with a significant body of research on the correlations of quantity and quality for ideation. In the years since Osborn's hypothesis on quantity to achieve quality (Osborn, 1953), the preponderance of evidence supports this hypothesis, namely that there is a correlation between quantity and quality of ideas during brainstorming. This correlation has been affirmed both for cumulative group quantity (Diehl & Stroebe, 1987; Leggett Dugosh & Paulus, 2005; Paulus, Kohn, & Arditti, 2011) and also individuals within a group (Kudrowitz & Wallace, 2013).

The methods used in this study, and similar previous work, imply contributions from open innovation or 'crowds'. The objective of seeking crowd input may vary depending on the application. Large, non-expert crowds have repeatedly been shown to provide a degree of accuracy in aggregated input at least equivalent to small numbers of experts (Surowiecki, 2005). These tasks are typically quantitative assessment tasks. The central limit theorem suggests the distribution of responses would allow for the calculation of a mean value with increasing confidence as the sample size increases. On the other hand, when soliciting input from a crowd for qualitative tasks, the benefit lies in capturing the full diversity (high variance) of the potential space. Figure 1 is a schematic representation of this distinction. In the area of needfinding, the initial qualitative task of collecting needs can benefit from the this divergent process of capturing crowd input since a large number of unique needs may likely result in finding some high-quality needs. However, after the list of needs is collected, a quantitative process can also be used to perform a preliminary ratings-based assessment for prioritization.

1.3 *Users articulating needs*

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 shows a schematic to represent different approaches to this limitation. In the case of immersive methods, such as interviews used by Griffin and Hauser, 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

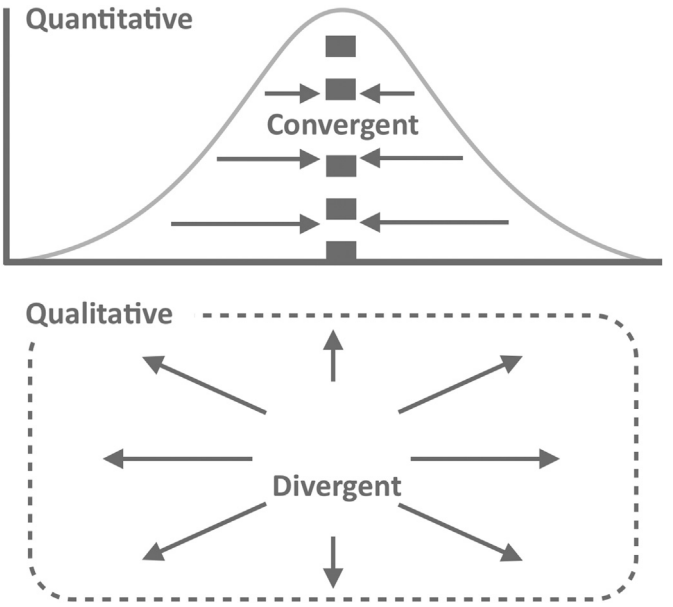

Figure 1 Convergent and divergent uses of crowd input for quantitative and qualitative study

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 very high, even if individual users have little time. This is typically the case in crowd-sourcing scenarios.

A majority of research using stimuli in creative tasks has focused on ideation, and stimuli are predominately visual (e.g. sketches, images) (Jansson & Smith, 1991, Viswanathan & Linsey, 2013). However, textual stimuli have been effective for design tasks in architectural design (Goldschmidt & Sever, 2011). 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 (Bayus, 2008; Reich, Konda, Monarch, Levy, & Subrahmanian, 1996), and preliminary studies support the use of empathy tools to aid users articulating their own needs (Lin & Seepersad, 2007). Studies evaluating stimuli of contextual images and also example need statements from previous users show the mean quantities of needs per person can increase by 50%–80% relative to a control group (Schaffhausen & Kowalewski, 2015).

1.4 *Managing large data sets to filter duplication*

When a large number of users (e.g. over 100) are able to submit need statements, the list can be very long (over 500). Determining rates of duplication

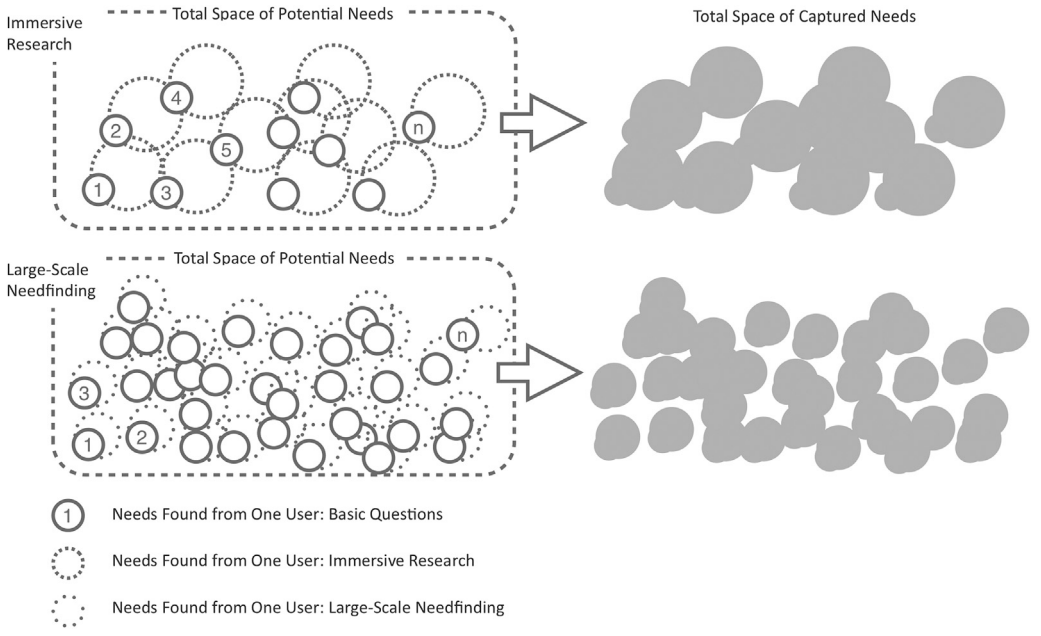

Figure 2 Multiple approaches to capturing a large portion of the needs space from user research

for very long lists can be aided by automated processes (e.g. finding duplicates with similar semantic meaning but not identical text like 'trying to stay cool' or 'not getting too hot'). In particular, natural language processes (NLP) research has generated a wealth of computational algorithms using machine learning to improve the assessment of the semantic similarity between two textual passages. Previous research has evaluated state-of-the-art Semantic Textual Similarity (STS) algorithms and identified top performers in international competitions of algorithm performance (Agirre, Diab, Cer, & Gonzalez-Agirre, 2012, 2013).

In Figure 2, overlapping circles represents duplicate or redundant information. There is limited research to understand the specific degree of scatter or clusters of users across the space for generalized user groups. Previous assessments for exact duplicates, rather than equivalent meaning, indicated rates of less than 1% (Schaffhausen & Kowalewski, 2015a).

1.5 Assessing quality of need statements

The application of large-scale needfinding requires a simplified metric appropriate for a rapid screening of 500 or more need statements. Common metrics for preliminary quantitative prioritization include variations on the importance of the need and whether existing solutions are satisfactory. The relative importance of needs has been commonly used and suggested in product development texts (Ulrich & Eppinger, 2004). The Kano model and analytical

methods rely on assessing satisfaction or dissatisfaction (Kano, Seraku, Takahashi, & Tsuji, 1984; Mikulic & Prebezac, 2011). Combinations of importance and satisfaction have been reported in descriptions of quality functional deployment (Matzler & Hinterhuber, 1998) as well as for prioritizing lists of outcome-based need statements (Ulwick, 2002).

1.6 Affects of need statement characteristics

Any particular need statement could be defined by a large number of characteristics. Several characteristics have been previously studied within the domain of problem finding or creativity, often in psychology and education disciplines. Examples include position on the spectrum of ill-structured to well-structured (Lee & Cho, 2007; Yeo, 2015) or whether the problem was presented or discovered (Runco & Okuda, 1988). Additional characteristics are particularly relevant to large-scale needfinding; however, prior research is more limited. Several examples are the effects of the sequence of the need relative to the complete list an individual has provided, the effects of the need statement originality or novelty, and the level of detail present for the need statement.

The effects of sequence can be introduced with reference to comparable research from ideation. Here, Kudrowitz and Dippo show the originality of ideas submitted during divergent thinking exercises increases as individuals submit more ideas. This means that the first entries submitted per person are typically listed by many people and are therefore not adding significant value to the pool of ideas (Kudrowitz & Dippo, 2013). While originality is typically a component of quality of ideas, this work does not independently rate quality of ideas.

In needfinding, the question arises whether it is better to increase the quantity of needs submitted by each individual or increase the number of individuals. In other words, would one representation shown in Figure 2 result in higher quality than the other. If the first needs submitted by an individual are frequently duplicated, a larger group may quickly result in diminishing returns. While previous work has based originality on a measure of rates of occurrences using manual sorting, the availability of automated-algorithm similarity ratings as described in Section 1.4 allows for a new metric of originality. This paper provides an analysis of effects of originality based on the presence of many or few similar statements as scored by a state-of-the-art NLP algorithm.

Previous research has considered the content of statements in problem finding studies, such as whether the topic relates to real-world problems (Okuda, Runco, & Berger, 1991). However, little attention has been directed towards the level of detail provided to explain a need statement. When performing a quantitative prioritization of many needs statements using a quality metric,

users rating the quality could be presented with only summary statements or additional full-length content and descriptions. This impacts the resources required to prioritize many needs as reviewing full-length stories takes more time.

1.7 Contributions and hypotheses tested

The results provided a quantitative foundation to understand effective quality-rating methods and the desirable characteristics of need statements submitted by users. The quality-rating methods described in Section 2 relate to new web-based rankings of need statements. The actual need statements were submitted by crowds of users during a previous study employing a different web-based application. A series of hypotheses were tested. In addition to these primary hypotheses, a number of additional descriptive analyses were performed.

H1: Needs submitted first will 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.

H2: Semantically similar need statements will 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.

H3: Need statements will 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.

2 Methods

This study analyzed quality ratings data collected using a web-based application applied to large user groups. Quality ratings assessed the quality of previously generated need statement data consisting of sentence-length need statements and paragraph-length detailed stories. The data collection and

analysis methods used to evaluate effects of need statement characteristics on rated quality are described here.

2.1 Need statement data description

Need statements were previously collected using an interactive, content-rich custom web application with an objective to collect as many need statements as possible. Participants were recruited from Amazon Mechanical Turk (AMT). Previous work has demonstrated the validity of data from AMT workers when restricting participation to high-reputation workers (Peer, Vosgerau, & Acquisti, 2014). Workers with lower than a 95% rating were excluded.

The instructions were intentionally framed to motivate a focus on quantity and to be analogous to brainstorming. Participants for data collection were randomly assigned to one of three general consumer product topic areas: preparing food and cooking, doing housecleaning and household chores, and planning a trip. The process of generating need statements and contextual stories was aided by three types of stimuli consistent with previous descriptions (Schaffhausen & Kowalewski, 2015). This stimulus information could be viewed simultaneously while entering need statements. The three stimulus types included: a narrative prompt, a group of previously submitted need statements (by other participants), and a group of images related to the topic. Participants could choose any stimulus type in any sequence and without limits on repetition. Time durations varied from under 5 min to nearly 3 h of information entry. The median duration was 11 min. Figure 3 shows a sample image during the need collection process. 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.

2.2 Need statement quality rating data set

The raw data from the needs collection described in Section 2.1 was analyzed using automated STS algorithms as described in Section 1.4. This automated algorithm rates the similarity of two statements on a scale of zero to five (Agirre et al., 2012). Need statement data was processed to remove contractions and non-standard characters. All possible pairwise combinations of need statements were analyzed to identify those pairs with a score greater than a pre-defined cutoff score. For these pairs, the algorithm rating suggests an equivalent semantic meaning, and these are potential duplicates. The cutoff score for identifying duplicates was based on analyses of best-case accuracy for false-positives and false-negatives over a range of scores compared to human raters (Schaffhausen & Kowalewski, 2015b). The raw data included a small number of potential duplicates, defined here as a pair of statements with a similarity score greater than four. There was no evidence of malicious copying, and a majority of

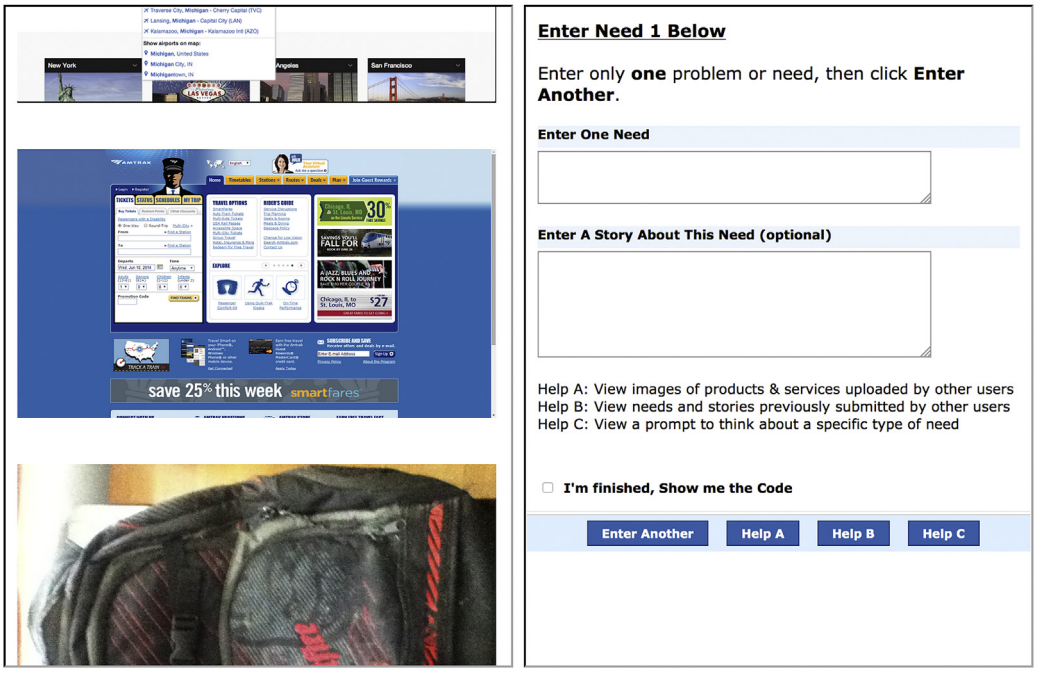

Figure 3 User interface with display of stimulus information and need statement entry

equivalent statements originated from different individuals. These statements were excluded from the analysis. Table 1 includes a breakdown of need statements for each topic area, the proportion submitted with an optional story, and potential duplicates removed from analysis.

2.3 Quality rating data collection

All quality ratings were collected using a custom online survey interface. Participants were recruited from AMT. Each participant was randomly assigned to one of the same topics originally used for need collection. Each participant answered optional demographics questions and selected descriptions to identify potential population segments. The instructions then provided details for the two quality criteria as described in Section 2.4.

Each participant was shown a random selection of 10 need 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. Statements could be flagged if they were unclear or represented a description of a solution rather than a need. Flagged statements were not rated for quality.

Table 1 Summary of need statements and topics (italics indicate exclusions)

Topic	*Users*	*Need statements*	*Including stories*
Cooking	104	568	439
Cleaning	121	650	422
Travel	116	517	385
Original subtotal	341	1735	1246
STS Duplicates	*N/A*	*−38*	*−30*
Rated for quality	341	1697	1216

The quality ratings for need statements were collected in three separate 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 1. Subsequent phases began with a modified set after a preliminary analysis to remove flagged statements and to exclude bottom quartile(s) of rated statements. The final phase included need statements with the quality rating in the top quartile and used a target sample size of 30 ratings per statement. An 'attention question' was included to check if raters were reading questions completely. Data from participants who failed the attention question was omitted.

2.4 Quality metric

The two criteria in this study were: how important the problem was to the participant, and how satisfied the participant 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').

The final quality rating was a linear combination of the two criteria scores as defined by Equation (1). The value of Satisfaction is inverted by subtraction from 6. This value contributes to the final rating such that high quality is a combination of a need with high importance and high 'Unsatisfaction'. Rating for Satisfaction was ultimately used in the study to use a more familiar phrasing.

$$Quality = Importance + (6 - Satisfaction) \tag{1}$$

This metric incorporates minor changes to the quality criteria as described by Ulwick. For the 'Opportunity' equation used by Ulwick, satisfaction is subtracted from importance, but cannot go below zero (Ulwick, 2002). This was not used because this formula loses fidelity when the satisfaction is high and importance is low. Statements might be rated the same opportunity but have very different satisfaction scores.

2.5 Data analysis methods

The effects of need statement sequence on overall need quality (hypothesis 1) were analyzed with graphical summaries of quality data. Two metrics were used to represent quality for groups of need statements. The first metric was a median of quality ratings per sequence group (represented by box plots) where the progression of groups describe all needs submitted first by users, all needs submitted second by users, etc. This has a benefit of capturing very 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 (e.g. if a participant submits 5 high-quality needs and 20 low quality, the median 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). 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, trend lines were plotted using scatter plots.

Hypothesis 1 was also tested using two data sets, further described in Section 3. 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 very 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 2, only need statement pairs previously identified as potential duplicates based on STS algorithms were analyzed. The difference in rated quality for statements in 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 2.2. Pairs were created using the average quality of the first submitted need, the 'baseline', and the average quality of each STS duplicate. The baseline statement and each duplicate were not rated by the same groups, these were each independent, randomized groups. A two-sided t-test was

performed using quality scores for the paired data. The trend line was plotted for the STS similarity scores and corresponding differences in rated quality.

Hypothesis 3 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. Pairs were created using the average quality of the 'story' need and the average quality of each corresponding 'no story' need. Different randomized user groups rated the 'story' and 'no story' needs. 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, 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 in this study.

A metric of originality can be calculated based on STS scores. This assumes if a need statement has very 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 indicate similar meaning. For each baseline need (submitted first), the count of pairs including this statement was calculated and plotted with corresponding quality ratings.

3 *Results*

Each analysis included only a portion of the total data collected. For example, the data collected for hypothesis 1 included a total of 25 837 ratings across the three phases for the total set of 1697 need statements. Table 2 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 as described in Section 2.3. All analysis sets in Table 2 are after exclusions.

Table 2 Summary of need statement data sets

Analysis description	*Need statements*	*Quality ratings*
H1, Raw data	1697	25 837
H1, All users analysis (Figure 4)	1626	21 717
H1, All users top quartile analysis (Figure 5)	406	8492
H1, Seven needs per user analysis (Figure 6)	144	1867
H1, Seven needs per user top quartile analysis (Figure 7)	37	780
H2, STS Pairs, Raw data	66	1146
H2, STS Pairs analysis (Figures 8 and 9)	64	992
H3, Story/No Story pairs, Raw data	90	2759
H3, Story/No story pairs analysis (Figure 10)	84	2181
Stimulus type analysis (Figure 11)	1626	21 717
Statement uniqueness analysis (Figure 12)	191	2674

Table 3 Examples of highest rated need statements overall and individual metrics (bold indicates the metric used for selection)

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.	**4.43**	4.14	6.28

Tables 3 and 4 include selected examples of high and low rated need statements. Table 3 includes highest rated statements (out of any topic) including overall score, or individual importance and satisfaction metrics. Table 4 includes corresponding low rated examples. Due to the study design of omitting

Table 4 Examples of low rated need statements with 10 or more ratings (bold indicates the metric used for selection)

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

lowest quality statements from final data collection phases, the selected examples also met a criterion of at least 10 ratings each.

3.1 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 1. Hypothesis 1 was not confirmed. Rather, results support the null hypothesis where needs submitted first will be equivalent in quality to needs submitted after a sustained period of time. The best 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 Figure 5 and trend up in Figure 7, 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 2. Figures 4 and 5 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 6 and 7 show the same analysis including only the approximately 20 users who submitted 7 needs. Figures 4(a) and 6(a) show median values of quality scores over the respective sequence ranges. Figures 4(b) and 6(b) show scatter plots with linear best fit lines and 95% confidence intervals. Figures 5 and 7 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, 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.

3.2 Quality of duplicate statements

Results support Hypothesis 2. The paired t-test results showed no significant difference between quality scores of duplicate pairs (p-value = 0.37). The difference in quality score was calculated for each pair. Figure 8 shows the distribution of differences in quality scores.

As shown in Table 1, there were 38 pairs of statements scored by STS algorithms as potentially duplicate. These 38 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 average quality score of the baseline statement and also the average quality score of the STS duplicate statement.

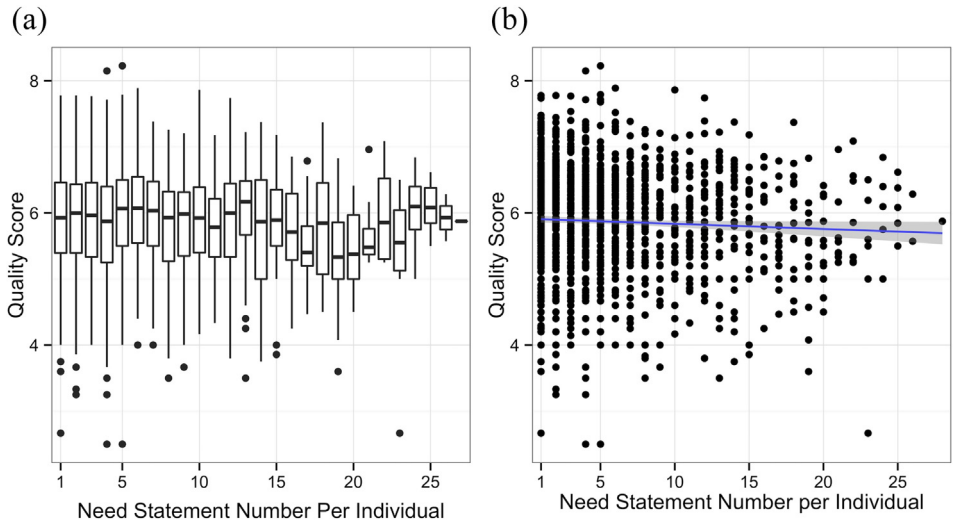

Figure 4 (a) Left, (b) Right: Quality of need statements for the sequence of needs per user [Shading indicates 95% CI, all users]

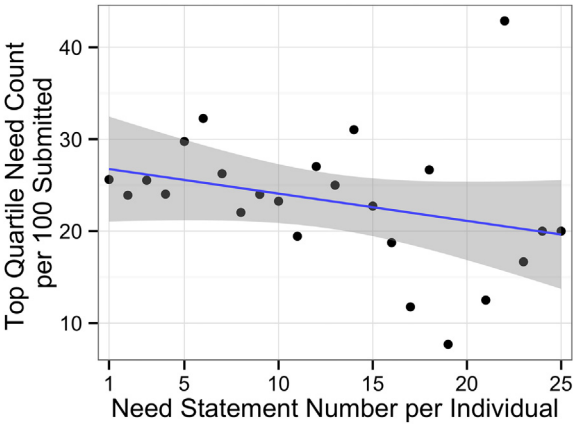

Figure 5 Count of top quartile need statements for the sequence of needs per user [Shading indicates 95% CI]

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 9 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 provides examples of representative points on the axis extremes in Figure 9. One example combines the lowest STS score (4.0) with greatest

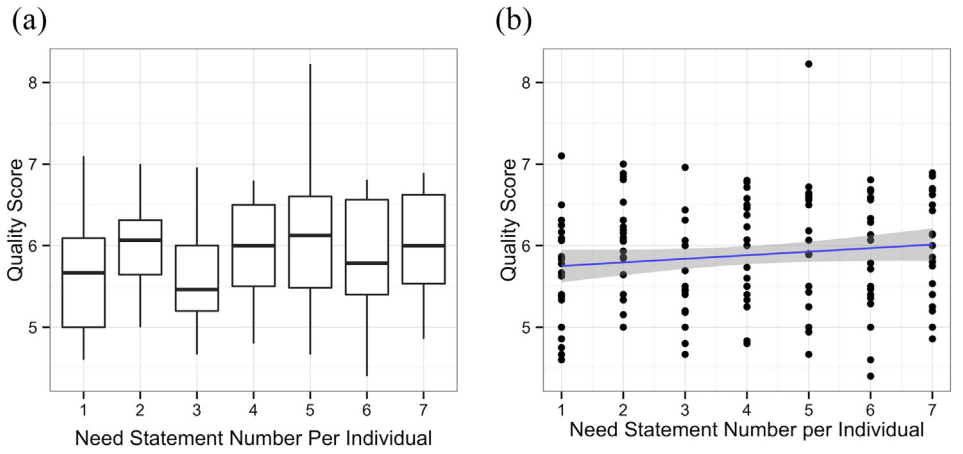

Figure 6 (a) Left, (b) Right: Quality of need statements for the sequence of needs per user [Shading indicates 95% CI, only users submitting 7 needs]

difference in quality score (2.4) The second example combines the highest STS score (5.0) with the lowest difference in quality score (0.1). Full text need statements are shown for each example.

3.3 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 average quality scores of each statement with and without the original detailed story. Results support a null hypothesis that need statements will be rated as the same quality if a detailed description is provided or omitted, and do not support hypothesis 3. The difference in quality score was calculated for each pair. The paired t-test results showed no significant difference between quality scores of duplicate pairs (p-value = 0.33). Figure 10 shows the distribution of differences in quality scores.

3.4 Quality of need statements after viewing a stimulus

Figure 11 shows a secondary analysis evaluating if a particular type of stimulus affects the rated quality of the need statement that follows. 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 very little change for different types, and Travel shows the Images ratio as less than 50% of others.

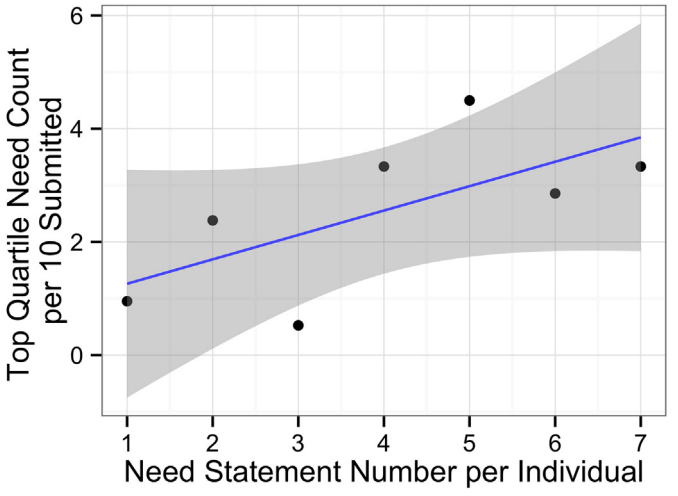

Figure 7 Count of top quartile need statements for the sequence of needs per user [Shading indicates 95% CI, only users submitting 7 needs]

3.5 Need statement quality and statement uniqueness

Figure 12 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 very 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). The best-fit linear

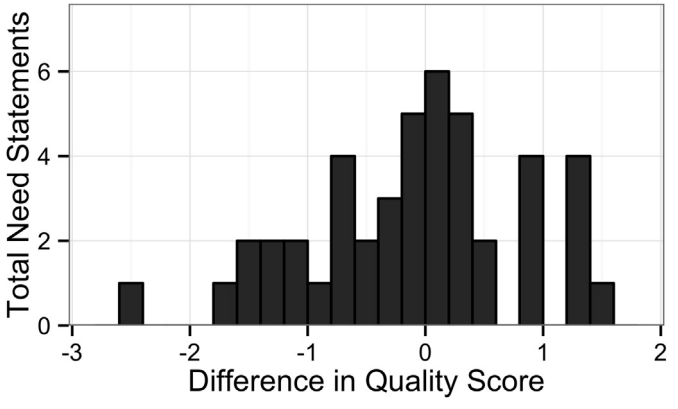

Figure 8 Distribution of variation in quality for STS duplicates

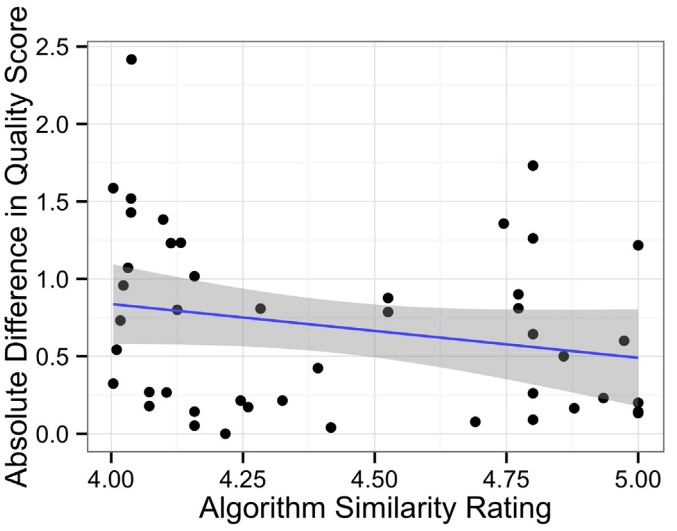

Figure 9 Different in quality for duplicate needs for varying similarity [Shading indicates 95% CI]

trend line does not indicate the number of similar need statements is correlated to a change in quality.

4 Discussion

The overall goal of this study was to evaluate effective characteristics of need statements. The results provided additional evidence that 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 need statement characteristics provided several important observations to inform large-scale needfinding.

The interpretation of results includes evaluating line fits in Figures 4b, 5, 6b, 7, and 9. The slopes of these lines are generally slightly positive or negative but within a confidence interval that may include a flat or reversed slope. In

Table 5 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

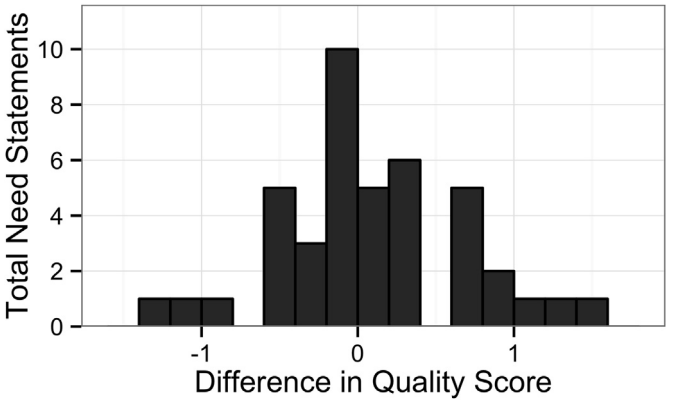

Figure 10 Distribution of variation in quality for omitted-story duplicates

the case of Figure 7, the pronounced slope was not replicated using additional analysis approaches, including those with a larger data set. For these reasons, conclusions state there was no effect in spite of this described variation of sloped line fits.

4.1 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

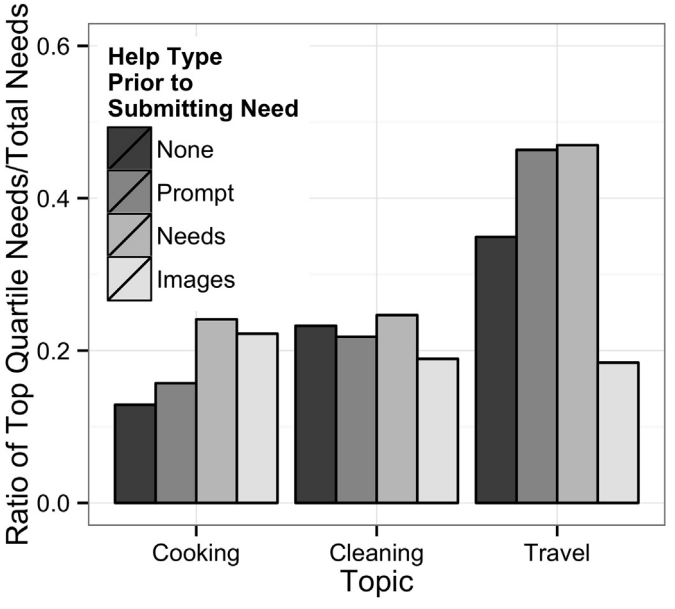

Figure 11 Quality of need statement for users viewing different stimulus types

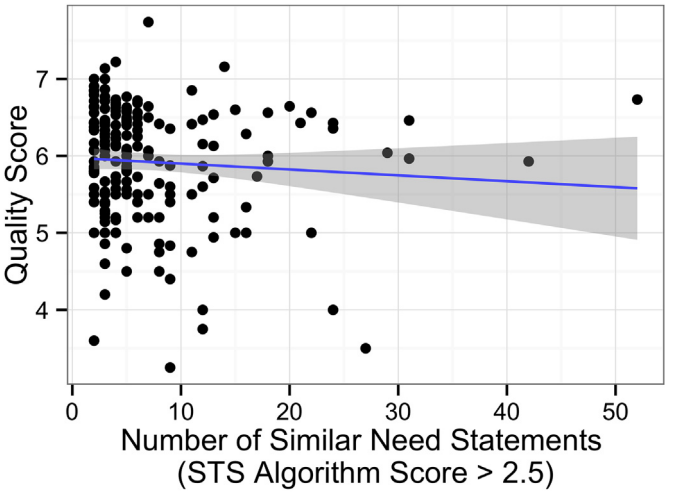

Figure 12 Quality of need statements and algorithmically-rated uniqueness [Shading indicates 95% CI]

time reviewing images and examples and listing a need arising fifth or tenth or twentieth. 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 using these methods as it indicates that there may not be a penalty in quality for using a higher quantity of individuals rather than longer per-person engagement with the online needfinding interface. 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.

4.2 *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 very 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 Figure 8 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 (Schaffhausen & Kowalewski, 2015b). Overall the result confirms that need statements scored as semantically equivalent will typically have equivalent quality ratings.

4.3 *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.

4.4 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.

4.5 Limitations and future work

There were several limitations to this work that should be considered. First, 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. Previous work has summarized the use of specific analyses for population segments (Schaffhausen & Kowalewski, 2015a). Only the overall population metric was considered here in order to assess broad preferences and to maintain higher sample sizes.

The results described here apply only to our needfinding methods. This relates to a content-rich, interactive online method where large groups of individuals can supply need statements using an interface specifically designed to help users articulate needs. Common methods, such as in-depth interviews, focus groups, or other interactions, might have different results. Additional research would be required to evaluate the similarities and differences of different methods.

All need statements were verbatim content supplied directly by users. In some cases, the statement might have been more clear or could potentially have been interpreted to reflect a similar, higher-quality need, but this step was intentionally not performed. The current work provides an important baseline of evidence and any future work to evaluate the merits of standardizing or rephrasing need statements should be compared to this baseline to understand whether the additional resources improve results. If so, future work can

address the user interface to encourage specific standardized grammar and formats and can also evaluate the application of previous methods to systematically rephrase existing statements (Vernon & Hocking, 2014).

The needs statements used in these studies do not represent lists encompassing the entire need space. Even with group sizes exceeding 100 per topic, there was not evidence of saturation of the qualitative data in this study. This may be attributed to the intentionally broad topics. A more specific topic may have higher rates of duplicates, demonstrating saturation sooner within a smaller group. This might suggest fewer unarticulated needs remain. As rates of duplication increase, the quantity of omitted data also increases; therefore, rates of high-quality needs may be different.

5 Conclusion

The study measured quality of need statements from user ratings and tested three hypotheses regarding the effects of different need statement characteristics on rated quality. The results show the needs that come to mind first are not lower quality than needs that come to mind later. This supports a balanced approach to recruiting new users and retaining existing users for immersive and in-depth input. 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. Quality ratings of semantically similar statements were equivalent, providing support for the use of automated algorithms to identify duplicate statements. These results support the use of large-scale needfinding methods for front-end design research and support future work of using less general topic areas and evaluating additional need statement characteristics.

Acknowledgements

This work was funded in part by internal University of Minnesota grants. We thank the Statistical Consulting Service at the University of Minnesota, and in particular Felipe Acosta, helped with the analysis of these experiments. We also thank William Durfee, PhD for early guidance on experimental design.

References

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., & Guo, W. (2013). 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* (pp. 32–43). Atlanta, Georgia, USA: Association for Computational Linguistics.

Agirre, E., Diab, M., Cer, D., & Gonzalez-Agirre, A. (2012). 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, Montre'al, Canada* (pp. 385–393).

Alkaya, M., Visser, F. S., & De Lille, C. (2012). 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* (pp. 1–8). Design Management Institute.

Bayus, B. L. (2008). Understanding customer needs. In S. Shane (Ed.), *The Handbook of Technology and Innovation Management* (pp. 115–141). West Sussex, England: John Wiley & Sons.

Boudreau, K. J., & Lakhani, K. R. (2013). Using the crowd as an innovation partner. *Harvard Business Review, 91*(4), 60–69.

Brabham, D. C. (2008). 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.

Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: toward the solution of a Riddle. *Journal of Personality and Social Psychology, 53*(3), 497–509.

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

Faste, H. (2011). Opening "Open" innovation. In*Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces, DPPI'11, 54* (pp 1–54). New York, NY, USA: ACM, 8.

Goldschmidt, G., & Sever, A. L. (2011). Inspiring design ideas with texts. *Design Studies, 32*(2), 139–155.

Griffin, A., & Hauser, J. R. (1993). The voice of the customer. *Marketing Science, 12*(1), 1–27.

Herriott, R., & Jensen, B. G. (2013). Students' responses to inclusive design. *Design Studies, 34*(4), 438–453, (Special issue: Articulating design thinking).

Howard, T. J., Culley, S. J., & Dekoninck, E. (2008). Describing the creative design process by the integration of engineering design and cognitive psychology literature. *Design Studies, 29*(2), 160–180.

Jansson, D. G., & Smith, S. M. (1991). Design fixation. *Design Studies, 12*(1), 3–11.

Johnson, D. G., Genco, N., Saunders, M. N., Williams, P., Seepersad, C. C., & Holtta-Otto, K. (2014). An experimental investigation of the effectiveness of empathic experience design for innovative concept generation. *Journal of Mechanical Design, 136*(5). 051009–1–9.

Kano, N., Seraku, N., Takahashi, F., & Tsuji, S. (1984). Attractive quality and must-be quality. *Journal of the Japanese Society for Quality Control, 14*(2), 147–156.

Kelley, T., & Littman, J. (2001). *The Art of Innovation*. Number ISBN-13: 978–0385499842. New York, NY: Crown Business.

Kouprie, M., & Visser, F. S. (2009). A framework for empathy in design: stepping into and out of the user's life. *Journal of Engineering Design, 20*(5), 437–448.

Kudrowitz, B., & Dippo, C. (2013). When does a paper clip become a sundial? Exploring the progression of originality in the alternative uses test. In *Proceedings of the National Conference on Undergraduate Research (NCUR), University of Wisconsin La Crosse 2013, WI* (pp. 3–18).

Kudrowitz, B. M., & Wallace, D. (2013). Assessing the quality of ideas from prolific, early-stage product ideation. *Journal of Engineering Design, 24*(2), 120–139.

Lee, H., & Cho, Y. (2007). Factors affecting problem finding depending on degree of structure of problem situation. *The Journal of Educational Research, 101*(2), 113–123.

Leggett Dugosh, K., & Paulus, P. B. (2005). Cognitive and social comparison processes in brainstorming. *Journal of Experimental Social Psychology, 41*(3), 313–320.
Lin, J., & Seepersad, C. C. (2007). 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* (pp. 289–296).
Margolin, V. (1997). Getting to know the user. *Design Studies, 18*(3), 227–236.
Martin, J., & Barnett, J. (2012). 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.
Matzler, K., & Hinterhuber, H. H. (1998). 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.
Mikulic, J., & Prebezac, D. (2011). A critical review of techniques for classifying quality attributes in the Kano model. *Managing Service Quality: An International Journal, 21*(1), 46–66.
Money, A., Barnett, J., Kuljis, J., Craven, M., Martin, J., & Young, T. (2011). 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.
Okuda, S. M., Runco, M. A., & Berger, D. E. (1991). Creativity and the finding and solving of real-world problems. *Journal of Psychoeducational Assessment, 9*(1), 45–53.
Osborn, A. F. (1953). *Applied Imagination*. New York, NY: Scribner's.
Patnaik, D. (2009). *Wired to Care: How Companies Prosper When They Create Widespread Empathy*. Upper Saddle River, New Jersey, USA: Rearson Education, Inc.
Patnaik, D. (2014). *Needfinding: Design Research and Planning* (3rd ed.). CreateSpace Independent Publishing Platform.
Patnaik, D., & Becker, R. (1999). Needfinding: the why and how of uncovering People's needs. *Design Management Journal (Former Series), 10*(2), 37–43.
Paulus, P. B., Kohn, N. W., & Arditti, L. E. (2011). Effects of quantity and quality instructions on brainstorming. *The Journal of Creative Behavior, 45*(1), 38–46.
Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon mechanical turk. *Behavior Research Methods, 46*(4), 1023–1031.
Poetz, M. K., & Schreier, M. (2012). The value of crowdsourcing: can users really compete with professionals in generating new product ideas? *Journal of Product Innovation Management, 29*(2), 245–256.
Reich, Y., Konda, S. L., Monarch, I. A., Levy, S. N., & Subrahmanian, E. (1996). Varieties and issues of participation and design. *Design Studies, 17*(2), 165–180.
Runco, M. A., & Okuda, S. M. (1988). Problem discovery, divergent thinking, and the creative process. *Journal of Youth and Adolescence, 17*(3), 211–220.
Schaffhausen, C., & Kowalewski, T. (2015). Large-scale needfinding: methods of increasing user-generated needs from large populations. *Journal of Mechanical Design, 137*(7), 071403.
Schaffhausen, C., & Kowalewski, T. (2015a). Assessing quality of user-submitted need statements from large-scale needfinding: effects of expertise and group size. *Journal of Mechanical Design, 137*(12), 121102.

Schaffhausen, C., & Kowalewski, T. (2015b). Large scale needs-based open innovation via automated semantic textual similarity analysis. In *Proceedings of the International Design Engineering Technical Conference IDETC 2015*. V007T06A045. http://dx.doi.org/10.1115/DETC2015-47358.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Toronto, Canada: Anchor Books.

Ulrich, K. T., & Eppinger, S. D. (2004). *Product Design and Development* (3rd ed.). New York, NY: McGraw-Hill/Irwin.

Ulwick, A. W. (2002). Turn customer input into innovation. *Harvard Business Review, 80*(1), 91–97.

Ulwick, A. W. (2005). *What Customers Want: Using Outcome-Driven Innovation to Create Breakthrough Products and Services*. New York, NY: McGraw-Hill.

Vernon, D., & Hocking, I. (2014). Thinking hats and good men: structured techniques in a problem construction task. *Thinking Skills and Creativity, 14*, 41–46.

Viswanathan, V. K., & Linsey, J. S. (2013). Design fixation and its mitigation: a study on the role of expertise. *Journal of Mechanical Design, 135*(5), 051008.

Yeo, J. (2015). Promoting students ability to problem-find. In *Motivation, Leadership and Curriculum Design* (pp. 215–224). Springer.

Zenios, S. A., Makower, J., Yock, P. G., Brinton, T. J., Kumar, U. N., Denend, L., et al. (2010). *Biodesign: The Process of Innovating Medical Technologies*. New York, NY: Cambridge University Press.