Assessing Quality of User-Submitted Need Statements From Large-Scale Needfinding: Effects of Expertise and Group Size

Cory R. Schaffhausen
Department of Mechanical Engineering,
University of Minnesota,
Minneapolis, MN 55455
e-mail: schaf390@umn.edu

Timothy M. Kowalewski[1]
Department of Mechanical Engineering,
University of Minnesota,
Minneapolis, MN 55455
e-mail: timk@umn.edu

Collecting data on user needs often results in a surfeit of candidate need statements. Additional analysis is necessary to prioritize a small subset for further consideration. Previous analytic methods have been used for small quantities (often fewer than 75 statements). This study presents a simplified quality metric and online interface appropriate to initially screen and prioritize lists exceeding 500 statements for a single topic or product area. Over 20,000 ratings for 1697 need statements across three common product areas were collected in 6 days. A series of hypotheses were tested: (1) Increasing the quantity of participants submitting needs increases the number of high-quality needs as judged by users; (2) increasing the quantity of needs contributed per person increases the number of high-quality needs as judged by users; and (3) increasing levels of self-rated user expertise will not significantly increase the number of high-quality needs per person. The results provided important quantitative evidence of fundamental relationships between the quantity and quality of need statements. Higher quantities of total needs submitted correlated to higher quantities of high-quality need statements both due to increasing group size and due to increasing counts per person using novel content-rich methods to help users articulate needs. Based on a multivariate analysis, a user's topic-specific expertise (self-rated) and experience level (self-rated hours per week) were not significantly associated with increasing quantities of high-quality needs. [DOI: 10.1115/1.4031655]

Keywords: user, needs, preferences, problems, quality, rating, assessment, crowd, expertise, needfinding

1 Introduction

Understanding unmet user needs is a critical first step in the development process for new products and services. This process is often a divergent task, and a number of methods exist to generate a list of potential unmet needs for further consideration. Interviews are commonly used during this phase [1,2] as well as observational methods [3] and other empathy-building tools [4]. Open innovation methods have been previously used to collect needs-based data from large groups of users [5]. Existing methods to collect user needs often rely on in-depth research with a small sample of users, and often users are targeted for specific skills or expertise. These subgroups might be experts [6] or "lead users" [7] or "extreme users" [8,9]. In each of these cases, users exceed the knowledge or skill of novice or average users. Collecting input from a wide range of users has been suggested [9]; however, this approach has been reported as less common for some specialized product areas, such as medical technology [10]. A narrow subgroup might not adequately reflect the diversity of the entire user group, and little quantitative evidence exists to guide the desirable characteristics of groups for large-scale user research.

Existing methods to assess the quality of need statements and converge on a subset of prioritized needs are poorly suited for the large numbers of needs that may be collected. A novel, simplified method to assess quality using crowd ratings of crowd-submitted need statements is an important tool to facilitate understanding of the needs of large, diverse groups. This study rated previously submitted need statements related to common topics of cooking, cleaning, and travel. The need statements were rated for quality with two user-rated criteria: Importance of the problem and satisfaction with existing solutions. These criteria have been previously described for prioritizing user needs using either a single criterion or both together [2,11–14]. Additional factors (e.g., market size or regulatory environment) may be useful after this rapid initial screening with two user-rated criteria.

Many factors including characteristics of the need statements themselves might influence quality ratings; however, only characteristics of the groups (e.g., size and user demographics) submitting and rating needs were considered here. Understanding quantitative fundamental relationships between user demographics and the quantity and quality of need statements can be of significant value to inform selection of participants for user research.

1.1 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 [13]. In this case, need statements are generally specific attributes expected for a new or incremental future product [12,13] or product family [15]. Others suggest the identification of product affordances as a method for capturing user needs [16,17]. The need can then be paired with a product requirement, indicating a quantitative metric to achieve in order to satisfy the customer [18,19]. 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 s.

[1]Corresponding author.

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received March 30, 2015; final manuscript received September 9, 2015; published online October 15, 2015. Assoc. Editor: Carolyn Seepersad.

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" [20], 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 [2,20], which is consistent with a focus on problems rather than desires [21].

A very broad sense use of the word "needs" is assumed for this study, 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 [3,9]. Ma et al. take this broad approach, presenting short storyboards to online users. Storyboards combine an example need with a potential solution in order to collect needfinding validation data; however, this includes feedback on both components [22]. As described in Sec. 2.3, 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). These statements reflect problems or unmet needs users face when performing common tasks or using common products.

1.2 Existing Quality Assessment Metrics. Assessing the quality of *ideas* generated during later phases of development has been extensively studied and previously summarized [23]; 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.

1.2.1 Kano Model. The Kano model is a framework developed in the 1980 s for classifying different types of user requirements [11]. 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 [12]. The model describes three types of desirable requirements or attributes: a basic requirement (also called a dis-satisfier 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 [11,12,21,24,25].

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 [26] or Taguchi methods [24], Monte Carlo simulation [25], or as an element of quality function deployment or house of quality [21,27]. Reports of these analytic methods often limit the quantities of statements (75 or fewer) [24,28].

1.2.2 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 [13]. Features are arranged hierarchically, consistent with voice of the customer methods [1]. The set of features used can be a subset of the total with a preference for those where importance is nonobvious. For example, obvious critical features for a product to function can be omitted. The suggested survey uses two questions: a rating of importance 1 (undesirable) 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 [13]. While quantifying the excitement from a feature might imply the degree existing products satisfy the particular need, a more explicit question might be beneficial.

1.2.3 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 [2]. The author states that an unfocused reliance on statements representing product solutions or benefits is a reason why voice of the customer methods continues to produce unpredictable results [20].

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 [20].

Once the list of outcomes is complete, it is distributed to a large number of potential users (often between 180 and 600 people), 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 criterion 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 [2,20]. These metrics share similarities with those used in quality function deployment [21], 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 the following equation, where "Imp" is the rated importance:

$$\text{Opportunity} = \text{Imp} + \max((\text{Imp} - \text{Satisfaction}), 0) \quad (1)$$

The current study used a quality criteria derived from the opportunity calculation by Ulwick but with important changes. The opportunity equation used by Ulwick [2] 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 very 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 this study.

1.3 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 differing degrees on population overviews or segments of a population [1,27]. Other methods are intended to inform requirements on specific products or a product category [2,13]. Critical areas where current methods might 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 attributes. Features relevant to a particular solution were explicitly discouraged.
(2) *Population Overview*: The output of quality ratings was not necessarily used to target a specific population segment (e.g., "soccer moms"). The list of highest-rated statements

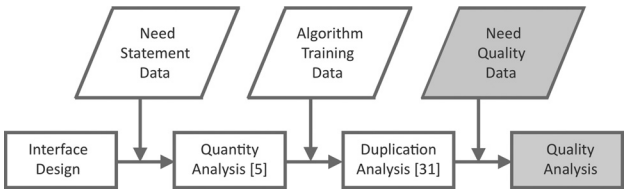

Fig. 1 Overview of previous (unshaded) and current (shaded) data collection and analysis

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 pursue divergent user needs research, but without an explicit focus on quantity. A summary of one systematic approach to need quantity and comparable ideation quantity research is available elsewhere [5].

1.4 Contributions and Hypotheses Tested. The results provided a quantitative foundation to identify effective quality metrics and the desirable characteristics of groups for large-scale user research. A series of hypotheses were tested: (1) Increasing the quantity of participants submitting needs increases the number of high-quality needs as judged by users; (2) increasing the quantity of needs contributed per person increases the number of high-quality needs as judged by users; and (3) increasing levels of self-rated user expertise will not significantly increase the number of high-quality needs per person.

2 Methods

This study analyzed data collected using a novel method of large-scale needfinding [5]. The data consisted of sentence-length need statements and paragraph-length stories providing additional context and detail. Previous need statement data were combined with new quality ratings data as shown in Fig. 1. Previous studies have been limited by the exclusion of quality data. Earlier work is summarized below and only the methods and data used to evaluate the above hypotheses are described in detail.

2.1 Need Statement Data Description. Need statements were previously collected using an interactive, content-rich custom web application developed using Zoho Creator (Zoho Corp., Pleasanton, CA). The Amazon Mechanical Turk (AMT) worker site[1] was used for recruiting participants, and a 95% approval rating was required. AMT worker demographics have been previous characterized and found to be comparable to traditional recruiting methods [29]. Data integrity from AMT workers can be maintained, in particular when targeting high reputation workers [30]. The objective was to collect as many need statements as possible. The instructions were intentionally framed to motivate a focus on quantity and to create a process comparable to brainstorming ideas.

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 that 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, and there was no limit to the number of selections. The complete details of the interface design as well as analysis of effects of interface content and rates for duplication have been previously described [5,31].

2.2 Need Statement Quality Rating Data Set. The raw data from the needs collection described in Sec. 2.1 were analyzed using a specialized type of natural language processing algorithm referred to as semantic textual similarity (STS) [32]. This automated algorithm rates the similarity of two statements on a scale of 0–5. After analyzing the raw data using STS algorithms, a small number of potential duplicates were identified. Potential duplicates are defined here as a pair of statements with a similarity score greater than 4. These statements were excluded from the analysis. Additional examples and details of the analysis are available elsewhere [31]. 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. The quantity of need statements (500+ per topic) significantly exceeds most prior work as described in Sec. 1.2.

2.3 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 such as: "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" and "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 Sec. 2.5.

Each participant was shown a random selection of ten problem statements related to the assigned topic. If a statement included a

Table 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	1735	1246
STS duplicates	N/A	−38	−30
Phase 1	341	1697	1216

[1]https://www.mturk.com/

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 ten statements. Repeat participants automatically bypassed demographics questions and proceeded to rate ten 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 ten ratings in that set would be labeled as an attention "fail." These were omitted from analysis.

2.4 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 1. Subsequent phases began with a modified set after preliminary analysis as shown in Table 2. All statements were initially rated by a minimum of five 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.

2.5 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 Eq. (2). 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

$$\text{Quality} = \text{Importance} + (6 - \text{Satisfaction}) \tag{2}$$

2.6 Data Analysis Methods. The effects of user group size on overall need quality (hypothesis 1) were 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 2 shows a schematic representation of analyzing one permutation. Each user and each need statement were 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.

Table 2 Overview of exclusion criteria for phases

	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

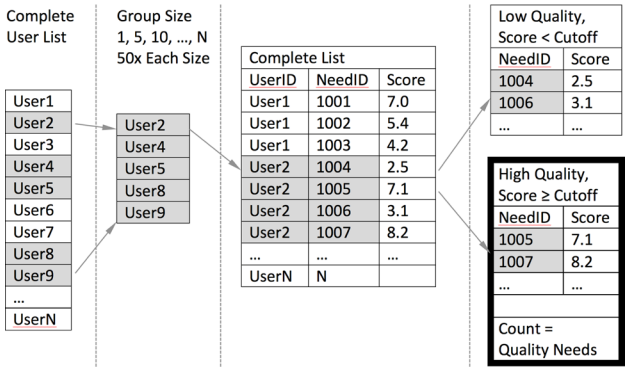

Fig. 2 Process for analysis of one group size permutation

In Fig. 2, shaded cells represent data included in the single permutation, and nonshaded 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 Fig. 2 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 cutoff (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 2, 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 3, the effects of user demographics (submitter or rater). Count data were not a normal distribution and were 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) was used on the generalized linear model to test pairwise combinations of user demographic groups [33]. For each demographic included in the analysis, if the response was blank, the quality data were excluded.

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

3 Results

The data collection process generated a total of 25,837 ratings across the three phases for the total set of 1697 need statements. Table 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 4 shows a summary of the counts of ratings collected for all phases and the number of

Table 3 Summary of need statement data sets

Criteria	Phase 1	Phase 2	Phase 3
Rated statements	1697	1168	289
E: flagged[a]	−66	−5	0
E: bottom quartile(s)[a]	−463	−874	N/A
After exclusions	1168	289	289

[a]E represents exclusions.

Table 4 Summary of need statement quality ratings

	Phase 1	Phase 2	Phase 3	Total
Ratings submitted	9739	11,854	4244	25,837
E: failed attention question[a,b]	−658	−859	−340	−1857
E: marked as flag[a]	−940	−925	−273	−2138
Ratings analyzed	8141	10,070	3631	21,842

[a]E represents exclusions.
[b]All ten survey ratings were omitted for a failed attention question.

individual ratings excluded because of flags or the participant failed the attention question as described in Sec. 2.3.

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.

3.1 Need Quality Distribution. Figure 3 shows the distribution of mean quality ratings for all need statements (aggregated topics) included in each phase. The quality equation is described in Sec. 2.5. 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.

3.2 Need Quality for Varying Group Sizes. The results for the group size permutation analysis (hypothesis 1) for each topic are shown in Fig. 4. The entire population (e.g., all segments) is included. Curves for high-quality needs versus group size are repeated for a range of cutoff values representing varying degrees

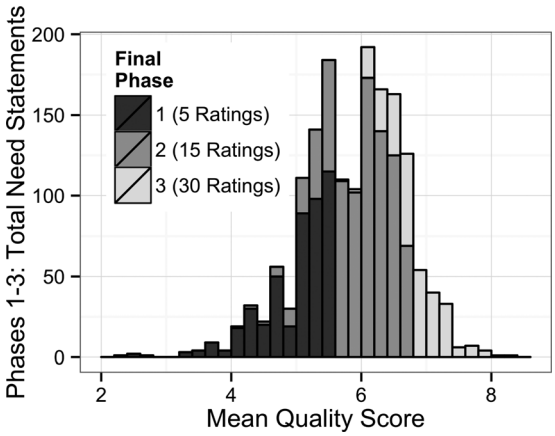

Fig. 3 Stacked distribution of quality scores (all phases)

of quality (7, 7.25, 7.5, and 8). Each point represents the mean of 50 random subsamples as described in Sec. 2.6. 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 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.

3.3 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 2). The trend of greater top quartile needs with increasing total counts is shown in Fig. 6. The data represent integer values; however, overlapping points are offset for clarity.

3.4 High Quality and User Expertise. Figures 7 and 8 descriptively represent the effects of user demographics on need quality (hypothesis 3). Figure 7 summarizes the number of top quartile need statements submitted by users in each self-rated expertise group. Figure 8 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 were less than 3% for both expertise and experience.

A Poisson regression analysis was used for testing difference of log(means) for expertise and experience based on likelihood-ratio

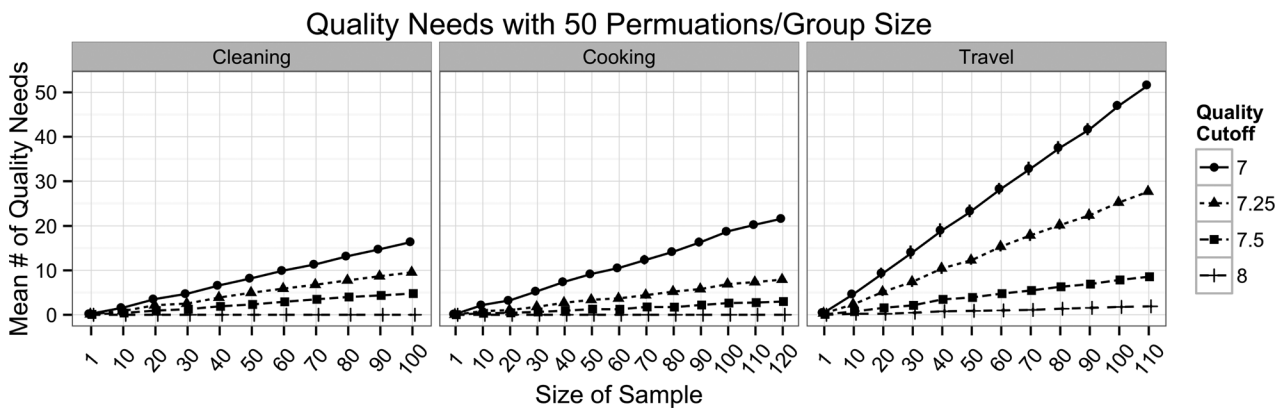

Fig. 4 High-quality needs for increasing group sizes (error bars indicate standard errors)

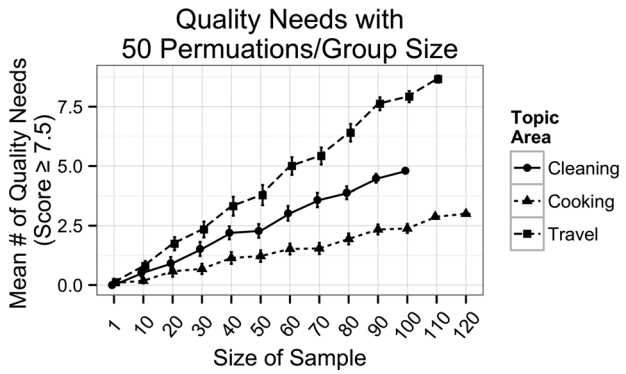

Fig. 5 High-quality needs (cutoff score = 7.5) for all topics and group sizes (error bars indicate standard errors)

test results and goodness of fit tests. The model was preferred because the additional parameter of the negative binomial model did not improve fit. 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 $p < 0.05$). 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 hrs).

3.5 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 3, 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 (group 1) and rated by a most-experienced user (group 4). Figure 9 shows the mean quality score for each level of submitter–rater difference for experience groups. There is no trend indicating that the degree of similarity of submitter and rater demographics (e.g., experience) affects the quality rating.

3.6 Highest-Rated Need Statements. Top-rated need statements both overall and for a selection of segments are listed in Table 5 for a single representative topic of cleaning. The top-rated overall need includes ratings from all users. Ratings for

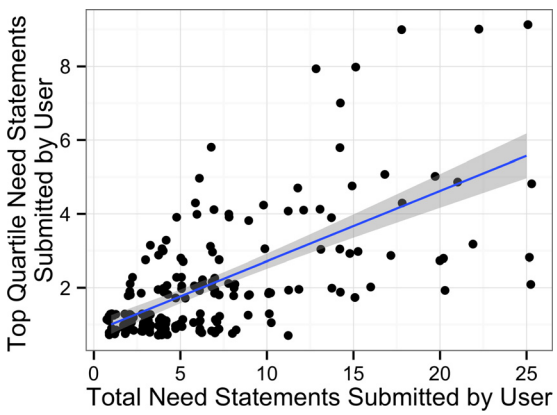

Fig. 6 Top quartile needs for users with increasing total need counts (shading indicates 95% CI)

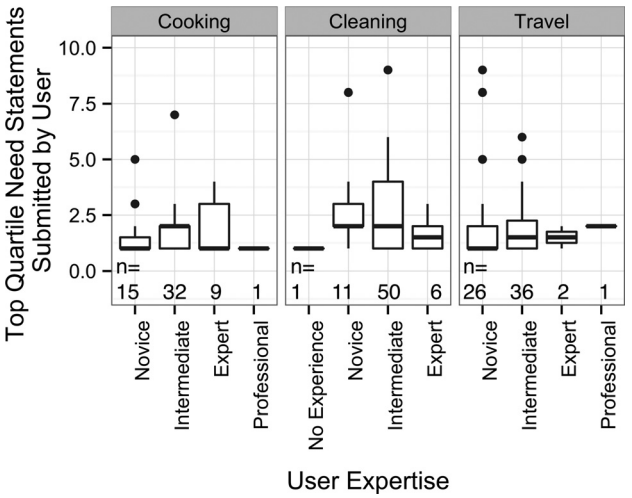

Fig. 7 Top quartile needs for all topics and expertise groups (group size, *n*, shown)

population segments include only those raters identifying with the user description shown as described in Sec. 2.3 (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.

4 Discussion

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 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 user and group characteristics provided several important observations to inform large-scale needfinding.

4.1 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 1) and individuals (hypothesis 2). Figure 4 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

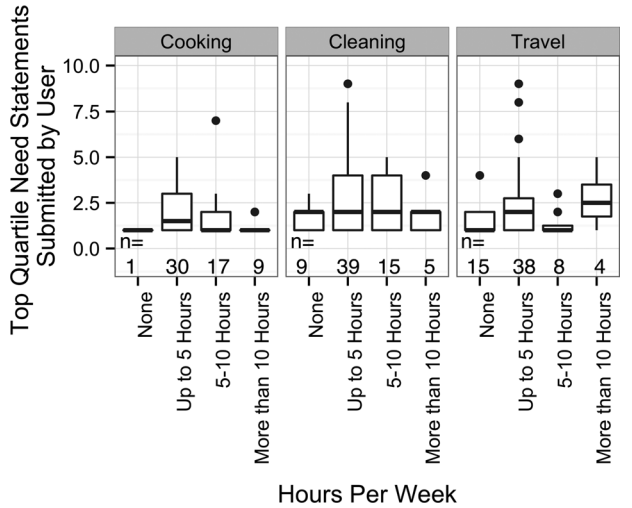

Fig. 8 Top quartile needs for all topics and experience groups (group size, *n*, shown)

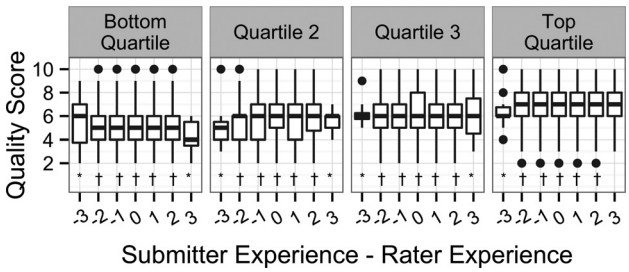

Fig. 9 Mean ratings for differences in submitter and rater experience (negative difference: need from low-experience user-rated by high-experience user), group sizes: *for $n < 20$; †for $n > 100$

from larger groups have previously been shown to include increasing counts of unique needs [31]. The increase in unique statements also results in a higher quantity of top-rated needs. Figure 6 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 [5] 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 suggests that diversity of other demographics (e.g., ethnic or socioeconomic) may 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 [1]. While previous studies have shown an asymptotic curve with diminishing returns for groups larger than 30 [1], the specificity of the topic might influence this outcome. To the best of 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 [34–36] and also individuals within a group [37].

While there was an effect for topic area on the number of high-quality needs submitted, previous results do not reflect this effect for total quantity [5]. This suggests that the final number of users needed to capture a specified number of high-quality needs will vary by topic.

4.2 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 [10], these results support a contrary hypothesis (No. 3) that increasing levels of self-rated user expertise will not significantly increase the number of high-quality needs per person. This is consistent with previous results that experts do not submit a higher quantity of total needs [5]. This is not confirmation that experts generally will be equivalent to nonexperts, as the data do 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 support 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.

4.3 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 [13], 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) related to computer keyboards, but this was not the case when isolating single segments. The examples of actual 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. Comparisons to existing methods are challenging due to insufficient data for existing methods.

4.4 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

Table 5 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 is not 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 is 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 cannot trust that the little splinters will be visible, or that they will get picked up	3.88	2.06	7.81

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 were 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 [38]. 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 [31]. 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.

5 Conclusion

These results 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 group 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 hrs/week) could be rated for quality by the same experience group or any different experience group (e.g., 5–10 hrs/week) without significantly effecting quality scores. The results support the future work of evaluating the effects of need statement characteristics on need quality and also for need collection and rating using specialized topics and specialized crowds.

Acknowledgment

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, Ph.D. for early guidance on experimental design.

References

[1] Griffin, A., and Hauser, J. R., 1993, "The Voice of the Customer," Mark. Sci., **12**(1), pp. 1–27.
[2] Ulwick, A. W., 2002, "Turn Customer Input Into Innovation," Harv. Bus. Rev., **80**(1), pp. 91–97.
[3] Patnaik, D., 2014, *Needfinding: Design Research and Planning*, 3rd ed., CreateSpace Independent Publishing Platform, Lexington, KY.
[4] Lin, J., and Seepersad, C. C., 2007, "Empathic Lead Users: The Effects of Extraordinary User Experiences on Customer Needs Analysis and Product Redesign," ASME Paper No. DETC2007/35302.
[5] Schaffhausen, C. R., and Kowalewski, T. M., 2015, "Large-Scale Needfinding: Methods of Increasing User-Generated Needs From Large Populations," ASME J. Mech. Des., **137**(7), p. 071403.
[6] Vredenburg, K., Mao, J., Smith, P., and Carey, T., 2002, "A Survey of User-Centered Design Practice," SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves, pp. 471–478.
[7] Von Hippel, E., 1986, "Lead Users: A Source of Novel Product Concepts," Manage. Sci., **32**(7), pp. 791–805.
[8] Bayus, B. L., 2008, "Understanding Customer Needs," *The Handbook of Technology and Innovation Management*, S. Shane, ed., Wiley, West Sussex, UK, pp. 115–141.
[9] Patnaik, D., and Becker, R., 1999, "Needfinding: The Why and How of Uncovering People's Needs," Des. Manage. J. (Former Ser.), **10**(2), pp. 37–43.
[10] Money, A., Barnett, J., Kuljis, J., Craven, M., Martin, J., and Young, T., 2011, "The Role of the User Within the Medical Device Design and Development Process: Medical Device Manufacturers' Perspectives," BMC Med. Inf. Decis. Making, **11**(15), pp. 1–12.
[11] Kano, N., Seraku, N., Takahashi, F., and Tsuji, S., 1984, "Attractive Quality and Must-Be Quality," J. Jpn. Soc. Qual. Control, **14**(2), pp. 147–156.
[12] Mikulic, J., and Prebezac, D., 2011, "A Critical Review of Techniques for Classifying Quality Attributes in the Kano Model," Managing Serv. Qual.: Int. J., **21**(1), pp. 46–66.
[13] Ulrich, K. T., and Eppinger, S. D., 2004, *Product Design and Development*, 3rd ed., McGraw-Hill/Irwin, New York.
[14] Takai, S., and Ishii, K., 2010, "A Use of Subjective Clustering to Support Affinity Diagram Results in Customer Needs Analysis," Concurrent Eng., **18**(2), pp. 101–109.
[15] Simpson, T. W., Bobuk, A., Slingerland, L. A., Brennan, S., Logan, D., and Reichard, K., 2012, "From User Requirements to Commonality Specifications: An Integrated Approach to Product Family Design," Res. Eng. Des., **23**(2), pp. 141–153.
[16] Cormier, P., Olewnik, A., and Lewis, K., 2014, "Toward a Formalization of Affordance Modeling for Engineering Design," Res. Eng. Des., **25**(3), pp. 259–277.
[17] Ciavola, B. T., Wu, C., and Gershenson, J. K., 2015, "Integrating Function- and Affordance-Based Design Representations," ASME J. Mech. Des., **137**(5), p. 051101.
[18] Green, M. G., Rajan, P. K. P., and Wood, K. L., 2004, "Product Usage Context: Improving Customer Needs Gathering and Design Target Setting," ASME Paper No. DETC2004-57498.
[19] Scaravetti, D., Nadeau, J.-P., Pailhès, J., and Sebastian, P., 2005, "Structuring of Embodiment Design Problem Based on the Product Lifecycle," Int. J. Prod. Dev., **2**(1–2), pp. 47–70.
[20] Ulwick, A. W., 2005, *What Customers Want: Using Outcome-Driven Innovation to Create Breakthrough Products and Services*, Vol. 71408673, McGraw-Hill, New York.
[21] Matzler, K., and 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), pp. 25–38.
[22] Ma, X., Yu, L., Forlizzi, J. L., and Dow, S. P., 2015, "Exiting the Design Studio: Leveraging Online Participants for Early-Stage Design Feedback," 18th ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 676–685.
[23] Dean, D. L., Hender, J. M., Rodgers, T. L., and Santanen, E. L., 2006, "Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation," J. Assoc. Inf. Syst., **7**(10), pp. 646–698.
[24] Chen, C.-C., and Chuang, M.-C., 2008, "Integrating the Kano Model Into a Robust Design Approach to Enhance Customer Satisfaction With Product Design," Int. J. Prod. Econ., **114**(2), pp. 667–681.
[25] Sharif Ullah, A. M. M., and Tamaki, J., 2011, "Analysis of Kano-Model-Based Customer Needs for Product Development," Syst. Eng., **14**(2), pp. 154–172.
[26] Zahedi, F., 1986, "The Analytic Hierarchy Process: A Survey of the Method and Its Applications," Interfaces, **16**(4), pp. 96–108.
[27] Duhovnik, J., Kušar, J., Starbek, M., and Tomaževič, R., 2006, "Development Process With Regard to Customer Requirements," Concurrent Eng., **14**(1), pp. 67–82.
[28] Agouridas, V., Winand, H., McKay, A., and de Pennington, A., 2006, "Early Alignment of Design Requirements With Stakeholder Needs," Proc. Inst. Mech. Eng., Part B, **220**(9), pp. 1483–1507.
[29] Paolacci, G., Chandler, J., and Ipeirotis, P., 2010, "Running Experiments on Amazon Mechanical Turk," Judgment Decis. Making, **5**(5), pp. 411–419.
[30] Peer, E., Vosgerau, J., and Acquisti, A., 2014, "Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk," Behav. Res. Methods, **46**(4), pp. 1023–1031.
[31] Schaffhausen, C. R., and Kowalewski, T. M., 2015, "Large Scale Needs-Based Open Innovation Via Automated Semantic Textual Similarity Analysis," ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (in press).
[32] Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A., 2012, "Semeval-2012 Task 6: A Pilot on Semantic Textual Similarity," 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, Montréal, Canada, Association of Computational Linguistics, Stroudsburg, PA, pp. 385–393.

[33] Bretz, F., Hothorn, T., and Westfall, P., 2010, *Multiple Comparisons Using R*, CRC, Boca Raton, FL.
[34] Diehl, M., and Stroebe, W., 1987, "Productivity Loss in Brainstorming Groups: Toward the Solution of a Riddle," J. Pers. Soc. Psychol., **53**(3), pp. 497–509.
[35] Dugosh, K. L., and Paulus, P. B., 2005, "Cognitive and Social Comparison Processes in Brainstorming," J. Exp. Soc. Psychol., **41**(3), pp. 313–320.
[36] Paulus, P. B., Kohn, N. W., and Arditti, L. E., 2011, "Effects of Quantity and Quality Instructions on Brainstorming," J. Creat. Behav., **45**(1), pp. 38–46.
[37] Kudrowitz, B. M., and Wallace, D., 2013, "Assessing the Quality of Ideas From Prolific, Early-Stage Product Ideation," J. Eng. Des., **24**(2), pp. 120–139.
[38] Vernon, D., and Hocking, I., 2014, "Thinking Hats and Good Men: Structured Techniques in a Problem Construction Task," Thinking Skills Creat., **14**, pp. 41–46.