Uncovering the Ground Truth: Using Crowdsourcing to Reach the Masses

Jeff Scagnelli, Nielsen
Michael W Link, Nielsen
Justin T Bailey, NPD Group

Paper Presented at the American Statistical Association International Conference on Methods for Surveying and Enumerating Hard–to-Reach Populations, New Orleans, LA

Abstract

Crowdsourcing, or recruitment of a large, relatively diverse group of potential respondents through an open call (Howe, 2006) has recently surfaced as method for potentially reaching otherwise hard-to-reach populations to complete survey data or provide meaningful information on a variety of tasks. By leveraging mobile technologies, researchers have better opportunities to reach these populations than ever before. Questions remain, however, about how the effectiveness and reliability of crowdsourcing data. For example, can researchers rely on the crowd to provide accurate data, and to do it at an equal or lower cost than more traditional survey methods? In this paper, we discuss recent research that utilizes social media (i.e., Facebook), mobile phones, and crowdsourcing to collect information about local consumer behaviors in India. While our information was specific to our business needs, we demonstrate how crowdsourcing can be used to collect data by creating a preliminary validation to ensure quality information and obtaining insights and sentiments of hard-to-reach populations. We focus on the practical uses of crowdsourcing as a survey research tool that, while effective, might not be included in the standard research "toolkit". We show that crowdsourcing, in the right circumstances, can be a useful alternative for gaining response from underrepresented populations.

Keywords: technology, crowdsourcing, data quality

1.0 Introduction

Crowdsourcing offers the survey research community with an attractive solution to the challenge of recruiting hard to reach participants. Crowdsourcing is defined as the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call (Howe 2006). This ability to quickly mobilize respondents across a wide area in a cost effective manner can extend the reach of products which traditionally require face to face contact for recruitment and training. Previous research has shown that this method can be effective at gathering reliable data, while enjoying the benefits discussed above (Behrend et al. 2011). While the ability to acquire data through open call web sources such as Amazon Mechanical Turk has been demonstrated, the quality of the data is a key concern. Wais et al. 2010 attempted to address this issue with their work on filtering low-quality results to improve quality. A key to their approach was to design a qualification task in order to ensure that respondents who participated in the desired task were competent and motivated. Their process resulted in a 35.6% qualification rate and the series of process reviews they implemented resulted in a 1.69% final participation rate.

In this paper we will discuss the task response rates, user engagement and data quality during a pilot study conducted within Hyderabad India to determine the impact of quality control measures implemented as part of the research.

2.0 STUDY DESIGN

A one month Pilot study was conducted between September 7th 2012 and October 17th 2012 to determine the viability of this approach.

2.1 Sample

Respondents were recruited in from within the geography of Hyderabad India. They were recruited through a non-probability sample from within the vendors' online panel. A total of 264 unique respondents participated in the pilot. In general the sample composition was skewed to males 18-24 years of age with them representing 70% of the total user base. The sample was also composed of mostly college educated respondents, with 64% attending college or having a degree. Panel demographics for the respondents are gathered through their self-reported Facebook profile data, which is accessed when a user provides consent during the panel vendor enrollment process.

Table 1: Respondent demographic Profile

		Female		Male		Total
		(n)	%	*(n)*	%	*(n)*
Age						
	Under age 17	1	9.1	10	90.9	11
	18-24 years of age	39	17.4	185	82.6	224
	25-34 years of age	1	4.0	24	96.0	25
	35-54 years of age	1	33.3	2	66.7	3
	55+	0	.0	1	100.0	1
Education						
	High School or High School Degree	3	15.0	17	85.0	20
	Some College or College Degree (BA, BS)	29	17.3	139	82.7	168
	Higher Degree (Masters, PhD, etc.)	3	5.5	52	94.5	55
	Unknown	7	33.3	14	66.7	21
	Total	42	15.9	222	84.1	264

2.2 Methods

This pilot was designed to evaluate the accuracy and cost of using crowdsourcing to identify cosmetic stores in Hyderabad India. Previous testing had demonstrated the ability to gather information of this type; however data quality was an issue which required further review.

Two unique tasks were presented to the potential respondents:

1. Pre-qualification (Gold Task)
2. Store identification

The gold task was an optional pre-qualification for users which would participate in the primary task of identifying new cosmetic stores. The inclusion of a training process has been shown to increase data quality within crowdsourcing literature (Le et al. 2010). The Gold Task provided users with the address of a known cosmetic store and asked them to visit the address and submit the same information as with the primary task. The store address list was based on a subset of known retail establishments within the area from the Nielsen database.

Respondents in both tasks were required to submit the following information from their mobile device:

1. Photograph of the store front
2. Name
3. Address
4. Phone number

In order to gauge and improve the quality of these submissions a series of quality control processes were implemented through the panel vendor.

The process consisted of three stages:

1. Automated Review
2. Crowd Review
3. Manual Review

The automated review process evaluated characteristics of the photograph to verify that the photograph came from within the expected geography (Hyderabad India), as well as verify that it was not a duplicate submission. Once a record passed the crowd review process it was sent to a separate group of users as a photo review task. Those users would review the respondent submitted photograph and validate it against the store criteria provided by the respondent. A subset of store submissions was also reviewed manually by the panel vendor. In some cases the review took place prior to the crowd review and in some cases afterwards.

Incentives for participation were awarded once a record passed the quality review process and were approved. They were awarded by the panel vendor in the form of mobile airtime minutes, assigned to a respondent's validated mobile device. The vendors' incentive structure awards the respondents points for task completion, which equate to mobile airtime based on their structure. We will evaluate these in the results based on their equivalent US dollar amount.

2.3 Analysis

This pilot study had two primary measures of success which were stated in the design:

1. Achieve a cost per completion of between $0.50 - $1.00 USD per verified store.
2. Gain a Quality level of 80% or greater for approved records.
3. Identify approximately 1,000 cosmetic stores within Hyderabad India.

These measures were evaluated in the following ways:

1. **Cost:**
   a. Average cost per completion across pilot period: Total cost
   b. Effect of cost on accuracy rate: Does a lower incentive provide similar results?
   c. Effect of incentive level on activity: Does a lower incentive reduce activity?
2. **Quality:**
   a. Evaluate the record approval rates by each stage to gauge individual effectiveness.
   b. Gold Task: Did pre-qualification result in higher quality?
   c. Respondent tenure: Do more experienced panelists provide higher quality data?
   d. Respondent Activity: Do more active respondents provide higher quality data?
   e. Demographic representation: Did certain demographics provide better data?
3. **Completion Rates**:
   a. Unique Users That Saw Task (Reach)
   b. Unique Users That Clicked On Task (Conversion Rate)
   c. Unique Users That Made At Least One Submission (Response Rate)

3.0 Results

A total of 1,775 complete responses were submitted by respondents during the pilot period. The results were processed through the aforementioned quality checks and analyzed for accuracy based on the Nielsen store criteria used to enumerate stores within our retail sample.

3.1 Response Rates

In total 264 unique respondents completed and submitted at least one store identification task within the pilot period. This was out of a total of 21,466 unique users who were presented with the task, leading to an overall response rate[1] of 1.23%. Table 2 below displays the daily task reach, while table 3 deals with task response.

[1] Response rate is computed by using the total number of respondents who submitted at least one entry over the total number of respondents who viewed the task.

Table 2: Task Reach (Daily)

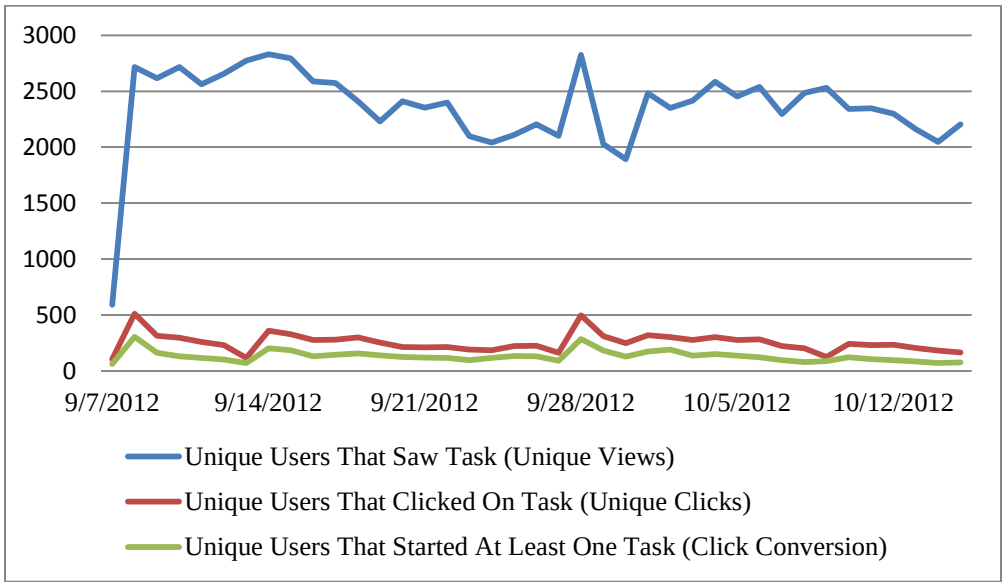
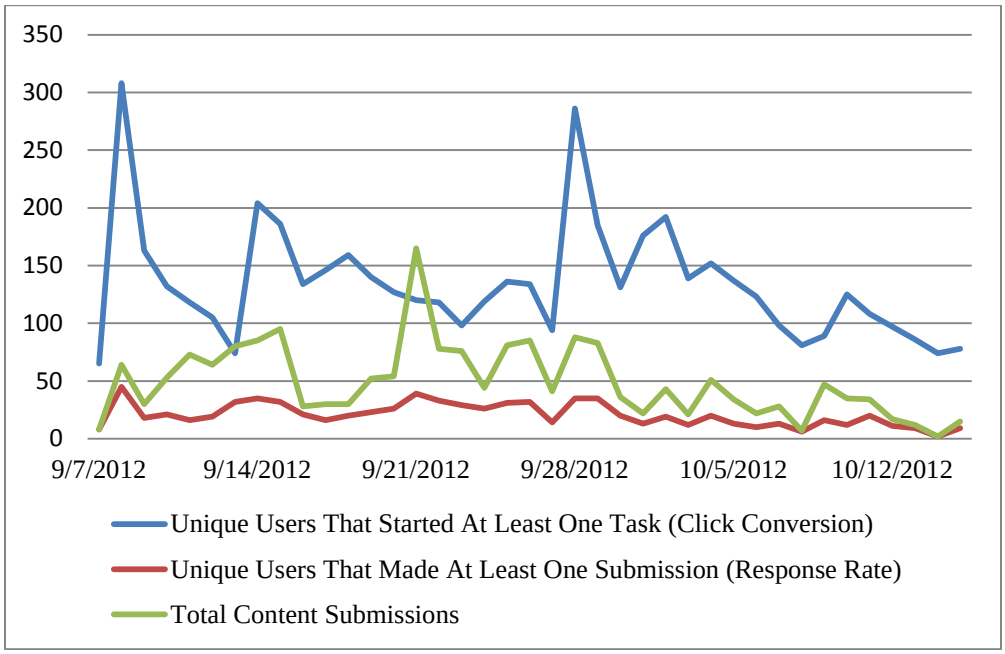
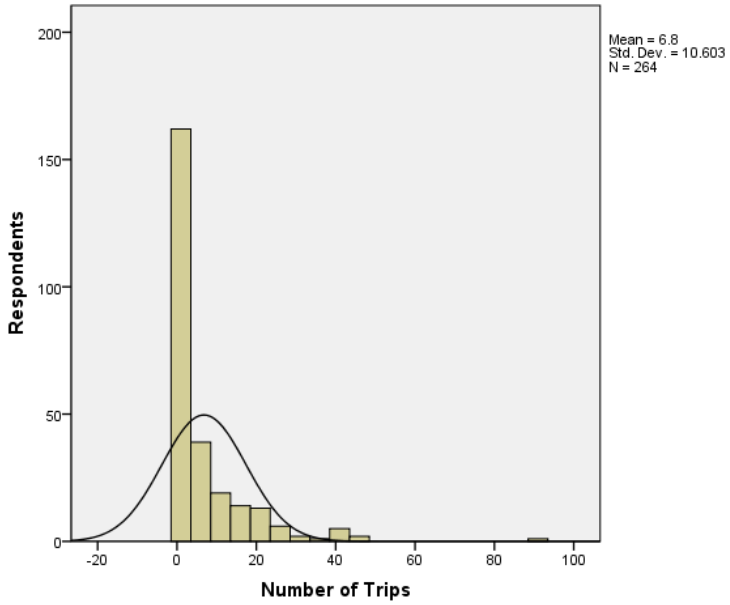


Table 3: Task Response Rates by stage of participation (Daily)



Respondents who participated in the store enumeration task were likely to complete multiple entries. It was expected during the study design process that this task would lend itself to repeatability. Table 4 below displays the distribution of responses by unique respondents. While the average number of submissions per respondent was almost 7, most respondents submitted 5 or fewer responses (71%).

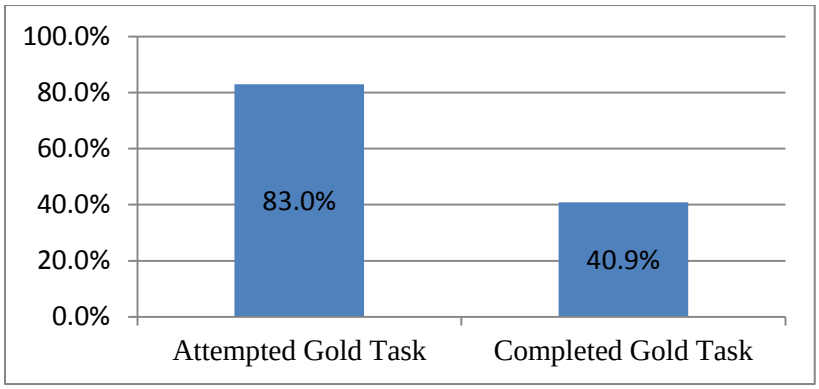
Table 4: User Level Response



3.2 Data Quality

Data quality was one of the key measures of success, previous testing had proven the ability to gather data through this methodology. Crowdsourcing literature has demonstrated that the inclusion of a training process can be effective in improving data quality (le et al. 2010). In this pilot we included a training task which utilized known cosmetic stores from the Nielsen retail database within the Hyderabad India. Due to programming issues, we were not able to make the task required as desired. It was presented as an optional task for the respondents, though the majority of them participated in the task nonetheless. A respondent who clicked on the task and begun the process was deemed to attempt the task, while the submission of complete data was required to complete it. Table 5 below illustrates the participation and completion rate for the gold task.

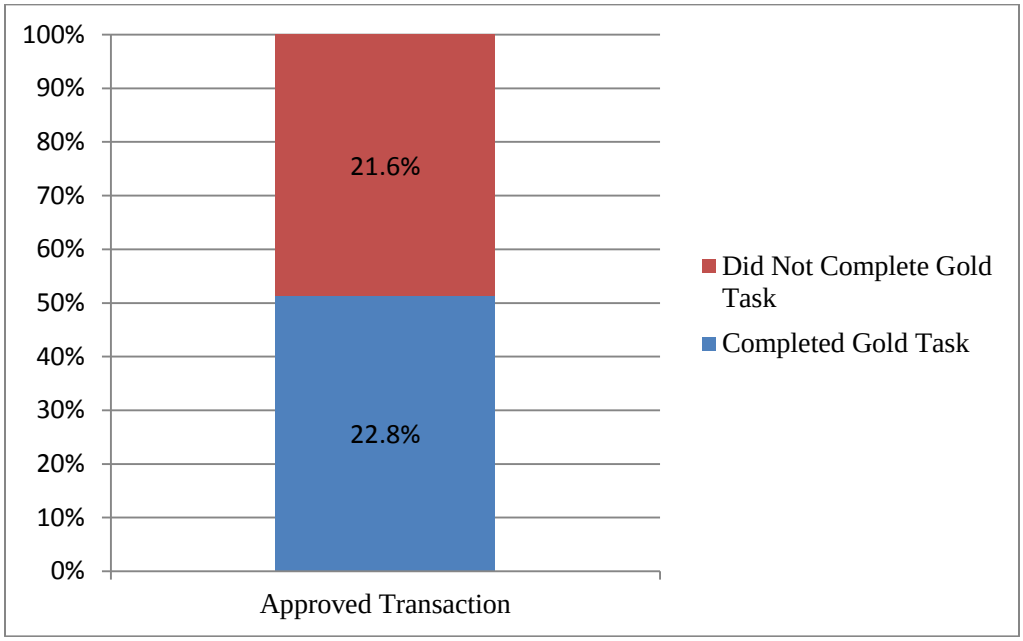
Table 5: Gold Task (Training) Participation



While the participation rate for this task was high, the intention was for this activity to drive higher quality in the store enumeration task. In order to evaluate the approval rate of store identification tasks evaluated between respondents who completed the Gold Task

versus those who did not. The expectation here was that respondents who completed the Gold Task would yield a higher rate of approved records, which ultimately pass all quality checks. Table 6 illustrates the difference in approved transactions between these two groups.

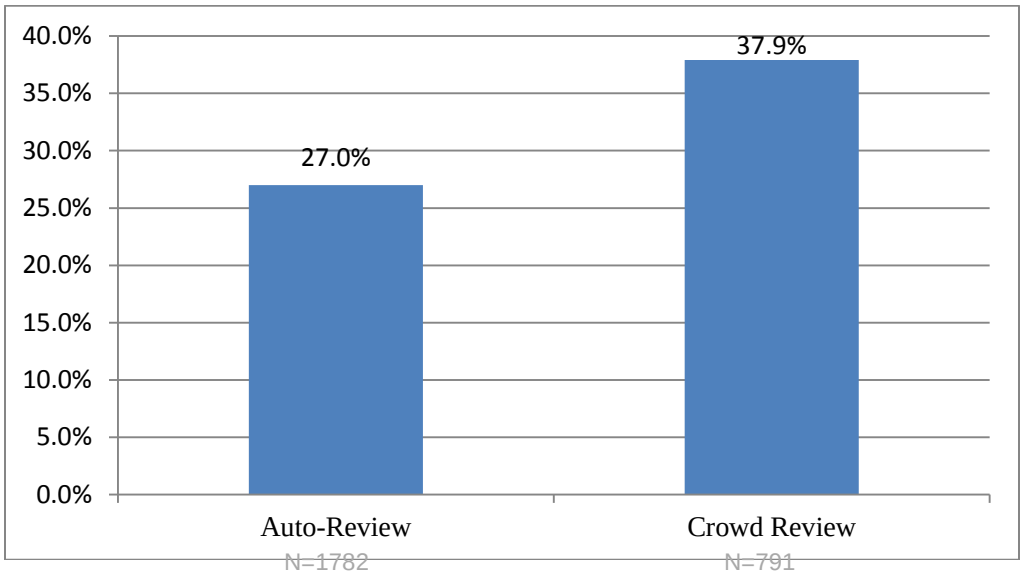
Table 6: Gold Task (Training) Participation and Store Data Quality



The presence of no statistical difference between the two groups was a surprise. The reason for this is unknown, though it may be due to both the panel based nature of these users as well as the complexity of the task they were asked to perform. We are not able to determine if the low approval rates are due to the submission of bad data or a misinterpretation of the task itself. This is something to investigate more closely in future research.

The inclusion of the separate review stages for submissions was also designed to raise the overall quality of approved results. Table 7 displays the percentage of records which went through each quality check and were ultimately approved.

Table 7: Quality Review Effectiveness (Approved Transaction Rate)



Note that the records which received a crowd review are not inclusive of records which failed the auto-review process or manually reviewed prior. Each stage built on the results of the previous step and was effective in removing invalid records. As previously stated, respondents were only incented when a record passed all stages of approval so each record removed in this process amounted to a cost savings.

During the course of this pilot the incentive level was adjusted in order to manage the flow of responses. The initial incentive level for an approved record was the equivalent of $1 US, this was adjusted as high as $2 and as low as $0.30. As the vendor managed this component to deliver the desired volume of responses the results cannot be interpreted as a controlled test of incentives. The results can however provide directional feedback on the impact of incentives on quality and quantity of responses. Table 8 displays the volume of responses by each incentive level while table 9 shows the rate of approved responses by the different levels.

Table 8: Response Volume by Incentive Levels
Incentive Levels

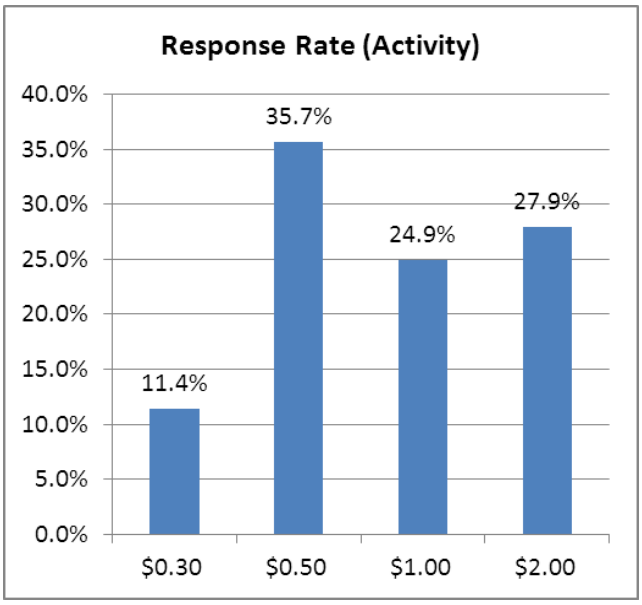
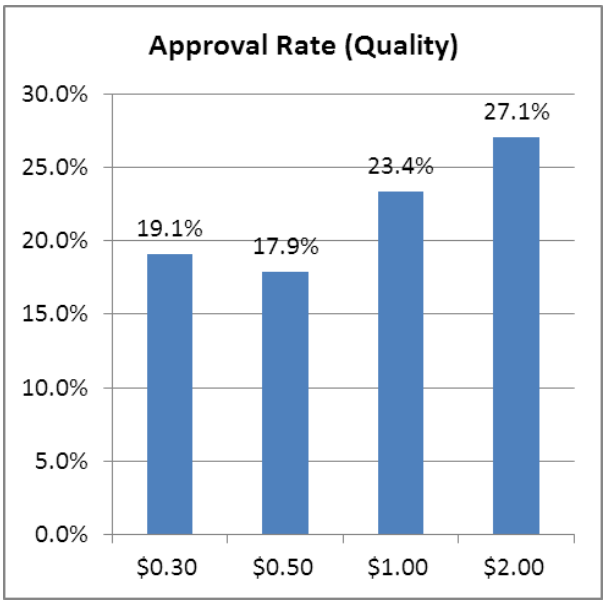


Table 9: Response Quality by

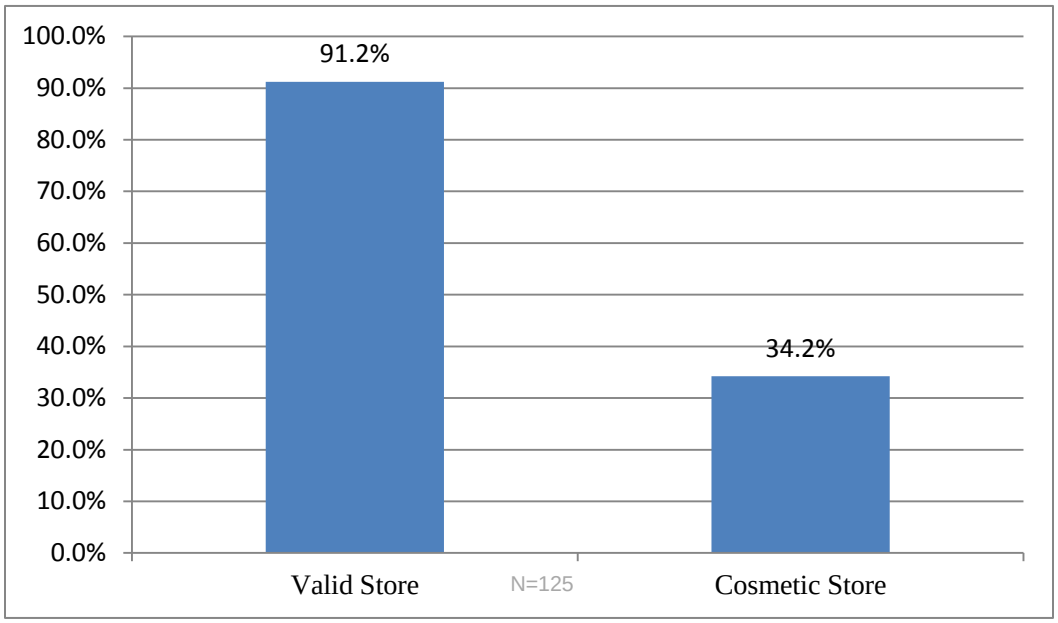


The data in tables 8 and 9 directionally indicate that a higher incentive level can lead to higher itme quality, but it may not nescessarily drive a higher volume of responses. While a higher incentive level did lead to higher quality responses, the differences may be mitigated through a higher volume of overall responses. Future testing should look to explore these impacts in a controlled setting to better understand the levels of impact and how to optimize for efficiency.

Once records passed all of the panel vendor's quality checks they were deemed approved results and were provided to Nielsen as valid store information. Nielsen then performed a phone audit verification process to further validate the results. A total of 395 stores were phoned with audits being completed for 125 of them, representing a 32% contact rate. The audits sought to validate that the stores met the Nielsen cosmetic store definition[3].

The audit results werte judged on two levels; 1) Was it a valid store (Name, address, phone number and 2) Did it meet the cosmetic store definition. Table 10 displays the results of this review.

Table 10: Response Quality Audit Verified



The results above demonstrate the ability of crowd to identify valid store locations with a high rate of accuracy; however the majority did not meet the store type criteria desired. This demonstrates both the promise and the challenges of this approach. It is unclear from our results whether the issues were related to misinterpretation of the task or data gathered during the phone audits. With the majority of respondents being younger males (18-24 years), it is possible that they were not best suited for this particular task. Future efforts utilizing this methodology should take into consideration the particular task and tailor a user group best suited for the process. These results were similar to Wais et al. in that the desired throughput was not achieved, while the screening process was effective at removing invalid data.

[3] Stores which primarily handle female cosmetics and do not stock any Food products.

While the results of the store identification task were disappointing the crowd review process did display promise for future work. Nearly 5,000 unique photo reviews were completed by the crowd review group, with 46% completed by power users (5 or more completions). On average the review was completed within 7 minutes and the incentive payout was equal to $0.01. The majority of the completions came from within the local area which points to the potential of leveraging local knowledge for an identification task such as this.

4.0 Conclusions

While we were able to develop a list of accurate stores at a low cost by leveraging a crowd sourced panel, the throughput goals were not achieved. There are over 3,000 cosmetics stores in the city[4]; however, the crowd sourced panel only identified 395 of these stores. This limits the scalability of the process for this type of task, which is critical for this product. The quality of the results indicates a challenge with respondents' ability to perform tasks that require judgment to complete the task accurately (e.g., be able to identify a Nielsen-defined Cosmetics store). This may have been due to the majority of responses being submitted by Males 18-24 years of age, who are not the primary customers for these store types. This is something that is important to plan for in future sample designs as it may introduce bias into your results. In this case the throughput was a critical metric and response and incentives were managed to drive response. The fact that the respondents who completed the qualification task did not submit higher quality responses was a surprise. The key consideration for future research on this topic is to ensure that the correct metrics are in place to allow you to measure the impact of these factors.

[4] Based on Nielsen Retail establishment data.

References

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior research methods*,*43* (3), 800-813.

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM*, *54*(4), 86-96.

Dow, S., Kulkarni, A., Klemmer, S., & Hartmann, B. (2012, February). Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 1013-1022). ACM.

Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010, July). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation* (pp. 21-26).

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, *14*(6), 1-4.

Huberman, B. A. (2008). Crowdsourcing and attention. *Computer*, *41*(11), 103-105.

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010, April). Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems* (pp. 2863-2872). ACM.

Wais, P., Lingamneni, S., Cook, D., Fennell, J., Goldenberg, B., Lubarov, D., & Simons, H. (2010). Towards building a high-quality workforce with mechanical turk. *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS)*, 1-5.