Scagnelli, Jeffrey Mark. 2013. "What the Crowd Yields: Considerations When Crowdsourcing." *Survey Practice* 6 (3). https://doi.org/10.29115/SP-2013-0018.

ARTICLES

What the Crowd Yields: Considerations when Crowdsourcing

Jeffrey Mark Scagnelli[1]

[1] Nielsen

Keywords: enumeration, mobile, crowdsourcing

https://doi.org/10.29115/SP-2013-0018

Survey Practice
Vol. 6, Issue 3, 2013

In this age of emerging technologies and fragmented populations, the ability to obtain cost-effective data from a broad sample is more elusive than ever. Crowdsourcing is a potentially 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). 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. We have built on that approach, while also including a training task as (Le et al. 2010) have to accelerate the learning process. Crowdsourcing allows you to quickly reach a wide array of potential respondents and there is a need to ensure that you include these quality controls to reduce data quality issues. In many cases the respondents to these services will be taking part in multiple studies at once, often driven by the advertised incentives. While you cannot ethically deny payment of an incentive when a respondent participates in good faith, it is justifiable to monitor the quality of returns for unusable submissions to protect the integrity of your process. In this research we will share some results of a pilot study conducted by Nielsen between September 7th 2012 and October 17th 2012 within Hyderabad India.

Background

In this age of emerging technologies and fragmented populations, the ability to obtain cost-effective data from a broad sample is more elusive than ever. Crowdsourcing is a potentially 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). 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 and Facebook 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. We have built on that approach, while also including a training element as Le et al. (2010) and Dow et al. (2012) have suggested to accelerate the learning process. Crowdsourcing allows you to quickly reach a wide array of potential respondents, and there is a need to ensure that you include these quality controls to reduce data quality issues. In many cases, the respondents to these services will be taking part in multiple studies at

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	0.0	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 (master's, 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

once, often driven by the advertised incentives. While you cannot ethically deny payment of an incentive when a respondent participates in good faith, it is justifiable to monitor the quality of returns for unusable submissions to protect the integrity of your process. In this research, we will share results of a pilot study conducted by Nielsen between September 7, 2012, and October 17, 2012, within Hyderabad India.

Sample

Respondents were recruited through a non-probability sample from an online panel. A total of 264 unique respondents participated in the pilot, 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%. The sample composition was skewed to males 18–24 years of age, representing 70% of the total user base. The sample also included mostly college educated respondents, with 64% attending college or having a degree. Table 1 presents the panel demographics for these respondents, which are gathered through their self-reported Facebook profile data. This data is accessed when a user registers for the panel which includes consent to the collection of this information is provided during this process.

Methods

Previous testing of this approach had demonstrated the ability to gather information; however, data quality was an issue. In this test, respondents were asked to identify cosmetic stores within the city limits of Hyderabad. They

[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.

were asked to submit contact information for the store, along with a photograph from their verified mobile device. Data submitted was reviewed through a series of filters. First, an automated review performed a geo-location and duplicate file check. Next, the photograph submitted 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 other cases afterward. Incentives for participation were awarded once a record passed the quality review process. The panel vendor provided this in the form points, redeemable for mobile airtime minutes, assigned to a respondent's validated mobile device. Records which were deemed by these quality checks were further reviewed through telephone audits to the stores. These audits were performed by Nielsen directly and sought to validate against the Nielsen cosmetic store definition.[2] A total of 395 stores were contacted with audits being completed for 125, representing a 32% contact rate.

There were three primary measures of success for this test:

1. Achieve a cost per completion of between $0.50–$1.00 USD per record.
2. Gain a quality level of 80% or greater for approved records.
3. Identify approximately 1000 cosmetic stores.

Results

A total of 1775 completed responses were submitted during the pilot period, from the 264 unique respondents referenced earlier. This task was not targeted to any specific demographic groups; however, high unit level nonresponse can lead to issues with the representation of data. When they did respond, 71% of the time an individual completed five or fewer submissions. The crowd review process lead to 46% of the submissions coming from individuals completing five or more submissions. An individual can more easily sit at his or her computer and complete multiple photo identification tasks than physically visit a store. The burden of the task that you present to the crowd will have a direct correlation to the response you receive as you can see here.

Data Quality

Data quality is a key consideration when utilizing this methodology for survey data collection. Previous testing within Nielsen had proven the ability to gather data, but the level of quality was lacking. This pilot aimed to improve upon the existing methodology by incorporating quality checks and best practices from other applications. Le et al. have demonstrated that the inclusion of a

[2] Stores primarily handle female cosmetics and do not stock any food products.

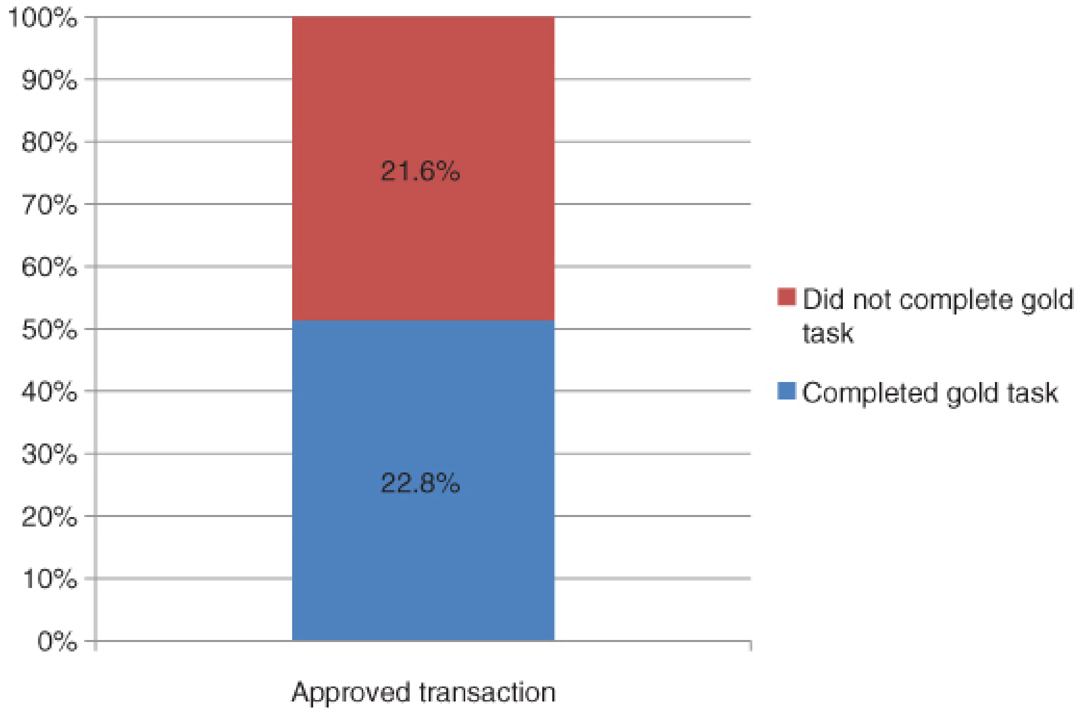

Figure 1 Gold Task participation and record approval.

training process can be effective in improving data, so a training task (Gold Task) was included here. The task replicated the process for store enumeration, sending respondents to known cosmetic stores within the geography based on Nielsen data. Programming issues prevented us from making this mandatory, though 83% of the final respondents participated in the Gold Task nonetheless. We expected to see those who completed the Gold Task provide higher quality results for the primary task. This was not the case for this pilot, and there was no significant difference between the two groups as Figure 1 illustrates.

While we did not implement the kind of dynamic feedback system that Le et al. had designed for their work, we did expect to see a positive result from the Gold Task It is possible that the panel based nature of this sample lead to respondents already being familiar with this type of work.

Once records were approved by the panel quality checks, the stores were phoned to further validate the data output. The audit results were judged on two levels: (1) Was it a valid store (name, address, and telephone number) and (2) Did it meet the Nielsen cosmetic store definitionn. Figure 2 displays the results at these levels, with the cosmetic store identification explored by gender. These are mixed results; while the crowd was able to identify valid store locations at a high rate of accuracy, the majority did not meet the business definition.

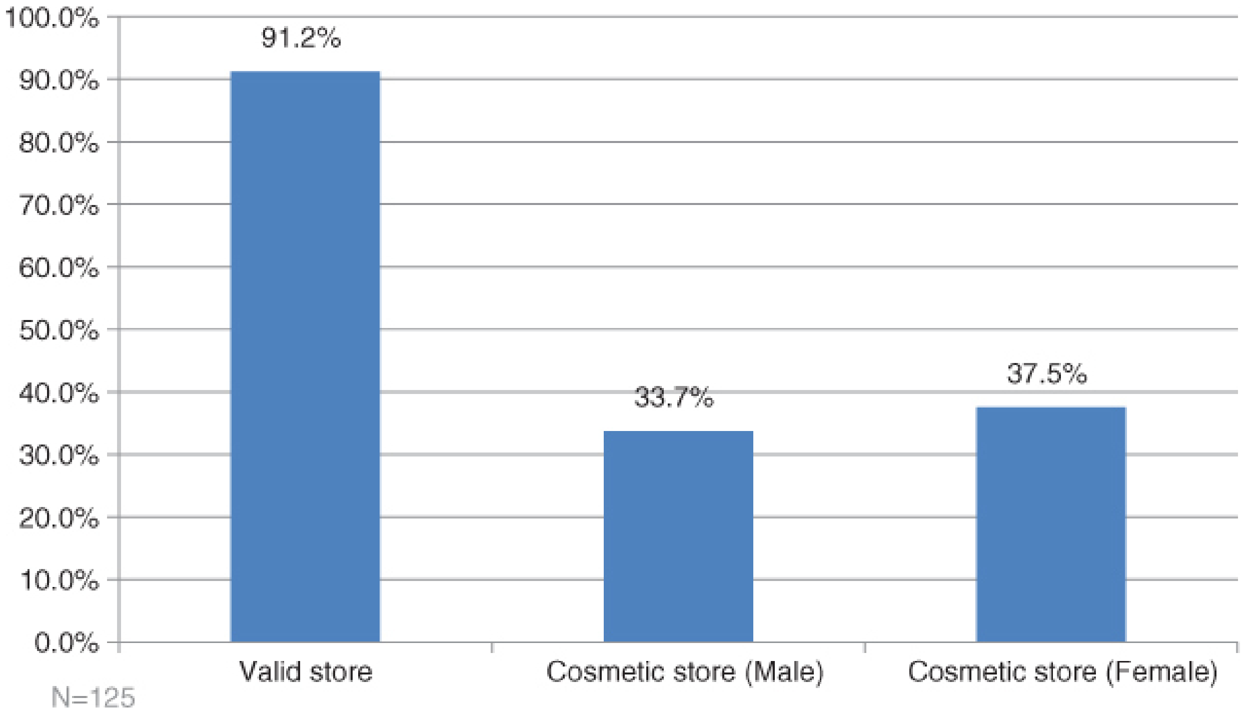

Figure 2 Response quality audit verified (gender).

We explored the impact of gender on the results as the majority of the sample was composed of younger males (18–24 years). Gender did not lead to response bias; however, we cannot quantify the impact nonresponse bias may have played.

Discussion

Crowdsourcing can yield a wealth of cost effective data in a short amount of time, but quality must be a key consideration in your design. This includes building the appropriate systematic checks and tailoring your task for the intended respondents. In this case, we experienced similar results to that of Wais et al. and the desired throughput was not achieved. There are over 3000 cosmetic stores in the city,[3] and the crowd sourced panel only correctly identified 39. While the final yield was low, the quality control process did successfully remove 78% of the invalid responses keeping the cost down. We did not enjoy the full benefit of a training task, but future applications will include expanded rigor to maximize the effect. Regardless of how you choose to employ this methodology, having the appropriate metrics in place to measure quality will lead to more reliable conclusions.

[3] Based on Nielsen Retail establishment data.

REFERENCES

Behrend, T.S., D.J. Sharek, A.W. Meade, and E.N. Wiebe. 2011. “Theviability of Crowdsourcing for Survey Research.” *Behavior Research Methods* 43 (3): 800–813.

Dow, S., A. Kulkarni, S. Klemmer, and B. Hartmann. 2012. “Shepherding the Crowd Yields Better Work.” In *Proceedings of the Association for Computing Machinery 2012 Conference on Computer Supported Cooperative Work*, ACM: 1013-1022.

Howe, J. 2006. “The Rise of Crowdsourcing.” *Wired Magazine* 14 (6): 1–4.

Le, J., A. Edmonds, V. Hester, and L. Biewald. 2010. “Ensuring Quality in Crowdsourced Search Relevance Evaluation: The Effects of Training Question Distribution.” In *Special Interest Group on Information Retrieval (SIGIR) 2010 Workshop on Crowdsourcing for Search Evaluation*, 21–26.

Wais, P., S. Lingamneni, D. Cook, J. Fennell, B. Goldenberg, D. Lubarov, and H. Simons. 2010. “Towards Building a High-Quality Workforce with Mechanical Turk.” In *Proceedings of Computational Social Science and the Wisdom of Crowds Neural Information Processing Systems Foundation (NIPS)*, 1–5.