Case Study Comparing Data Collected via Crowdsourcing vs. Trained Data Collectors for Tobacco Retail Audits

Annice E. Kim, PhD[1], Alicea J. Lieberman, MPH[2], Daniel Dench, BA[1]

[1]Public Health Policy Research Program, RTI International, Research Triangle Park, NC, USA
[2]Community Health Promotion Research Program, RTI International, Research Triangle Park, NC, USA

1. INTRODUCTION

Crowdsourcing is an "online, distributed problem solving and production model" and leverages online networks to: 1) gather information; 2) distribute large-scale tasks that are easier for humans rather than machines to process (e.g. analyzing photos); or 3) solicit ideas or solutions to existing problems as a challenge that can also be vetted by peers (Brabham, 2013). Crowdsourcing has been applied to a wide range of health topics, (e.g. Bow et al., 2013; Bradley et al., 2009; Lessl et al., 2011; Mavandadi et al., 2012; Swan et al., 2012), one being the collection of local-level data (Merchant et al, 2013). Crowdsourcing has been used to analyze large-scale pieces of information that require many hours of human cognitive processing (e.g., coding video/photo content, coding sentiment of Tweets) via online platforms like Amazon Mechanical Turk. Researchers use crowdsourcing for studies of wide-ranging topics from drug discovery (Lessl et al., 2011) to medical analyses and diagnoses (e.g., Mavandadi et al., 2012), development of education tools (e.g., Bow et al., 2013; Bradley et al., 2009), and public health issues, such as drug use and genomics (Swan, 2012).

Crowdsourcing has also been used to collect local-level data. In the past, data at the local level have been collected by hiring and training a core set of data collectors and paying for their transportation costs, which can be time-consuming and cost-prohibitive. Leveraging the power of local laypersons to collect data has the potential to lower costs and data collection time. In 2012, the University of Pennsylvania enlisted residents of Philadelphia in a crowdsourcing project to locate automated external defibrillators (AEDs)—life-saving devices used during cardiac arrest—that are mounted in public spaces throughout the city, but are often forgotten and neglected because few agencies maintain a database of AED locations (Winslow, 2012). The MyHeartMap Challenge, utilized the power of local laypersons to locate and map 1,429 AEDs across Philadelphia within 6 weeks, demonstrating the time and cost efficiency potential of crowdsourcing (Merchant et al., 2013).

Crowdsourcing has the potential to be useful for collecting data in the tobacco retail environment. Thousands of licensed tobacco retailers (LTRs) are dispersed across the United States, making it challenging to collect data from the entire census of retailers in short time-frames or cost-effectively. As part of our annual evaluation of the Bureau of Tobacco Free Florida (BTFF), RTI International conducts an audit of a random sample of LTRs in Florida; however, because it is time-intensive to train data collectors, the audits are only conducted in about 4% to 10% of LTRs annually.

In an ever-changing retail environment, with new tobacco products, shifting advertising practices, increasing use of promotions, and governments attempting to regulate these practices, the traditional method may not be effective to obtain data that requires a quick response. Crowdsourcing could be used for several purposes, for example, to set a baseline before a new regulation takes effect and later test whether or not this regulation had a meaningful impact on the retail environment.

RTI's study is the first study to our knowledge that has used crowdsourcing to collect point of sale (POS) tobacco data. In addition, no other studies we are aware of have compared the quality of data gathered by untrained local residents identified via crowdsourcing to data collected by trained data collectors. The purpose of this pilot study is to examine the feasibility of crowdsourcing Florida Retail Advertising Tobacco Study (RATS) data collection and compare the quality of data collected through crowdsourcing to that collected by trained data collectors. Results may help BTFF and other agencies determine whether crowdsourcing may be a viable option for conducting surveillance of POS tobacco marketing practices.

2. METHODS

In 2012, RTI International conducted the annual Retail Advertising Tobacco Study (RATS) for the Bureau of Tobacco Free Florida (BTFF). RTI trained surveyors from Retail Diagnostics Inc. (RDI) to conduct in-person audits of licensed tobacco retailers (LTRs) in Florida to document the extent of tobacco advertising, promotions, product availability, and placement. Over 2 days, surveyors were trained on study measures and data collection protocol using standardized codebooks, audit forms, and sample training datasets to assess interrater reliability. Surveyors' responses were compared to RTI coding (gold standard), and extensive feedback was provided for follow-up training sessions to review areas of weaknesses and to retrain on measures as needed. Surveyors then conducted in-field audits on a subset of retailers located in Miami and Tampa, Florida (N = 194), during a 3-week period (August 31 to September 19, 2012). During the same time period and at the same stores, we implemented a crowdsourcing study at the same store to assess how the answers of untrained local residents identified through crowdsourcing would compare to trained data collectors.

2.1 Crowdsourcing Data Collection

Gigwalk is a crowdsourcing mobile application, launched in 2011, that allows employers to post temporary job opportunities located anywhere across the Unites States. Postings are accessed by Gigwalk's workforce of more than 230,000 people (93% of whom are college educated), via a smartphone application. Gigwalk workers can view job posts in their area and apply to a posting of interest. Gigwalk automatically selects workers based on their skills and experience.

We posted 194 jobs spanning a 3-week period (August 31 through September 19, 2012) to the stores that the trained RDI surveyors were visiting. Each job ("gig") required the worker to visit an LTR in Miami or Tampa and to collect information on tobacco product advertising, promotions, and products sold. Because Gigwalk is intended for brief data collection efforts, we selected a subset of questions from the Florida RATS instrument and posted two different gigs: *Conduct a Cigarette Audit* (N = 99 stores) and *Audit of NEW SMOKELESS tobacco products!* (N = 95 stores). For cigarette audits, workers were asked to assess the presence of advertisements, products, and sales/specials for Marlboro Reds. For the smokeless audit, workers were asked to assess e-cigarette advertisements and sales of snus, e-cigarettes, and dissolvables. We included the instructions that trained RDI surveyors were given in their paper audit forms along with several example photographs of smokeless tobacco products from the training manual. No additional instructions or training were provided to Gigwalk workers. Retail location addresses were uploaded in an Excel spreadsheet, and each location served as a separate gig. Audits were due 1 week after posting.

Upon completion of a job, workers uploaded their survey answers, photos, and current location. We reviewed the survey responses and provided payment if work was complete. If questions were unanswered or unclear, we requested clarification from the worker. Workers received $7 for each completed gig. If a worker provided unsatisfactory work, RTI had the option to have them blocked from claiming future jobs on this study. We only had to use this option for one worker who posted unclear survey responses (possibly due to a language barrier) that required multiple communication exchanges to clarify their survey responses.

2.2 Measures

The two job postings included a subset of items taken from the trained data collectors' audit protocol. Workers were asked to take and upload photos (e.g., exterior marketing advertisements) with their smartphones. Comments/questions could be uploaded and sent directly to RTI. These comments were monitored for questions and addressed throughout data collection. Additionally, Gigwalk provided information on when gigs were posted, claimed, and completed, worker ID, and GIS coordinates of retail locations. Workers were instructed to collect the following measures:

- **"Conduct a Cigarette Audit" (N = 99 stores)**
  Workers were asked to record whether the store had any exterior cigarette ads (i.e., signage, portable or free-standing displays, or functional items affixed or not affixed to building) and, if so, the number of ads and to provide a photograph of the ads. Workers were also asked whether the stores sold Marlboro Red

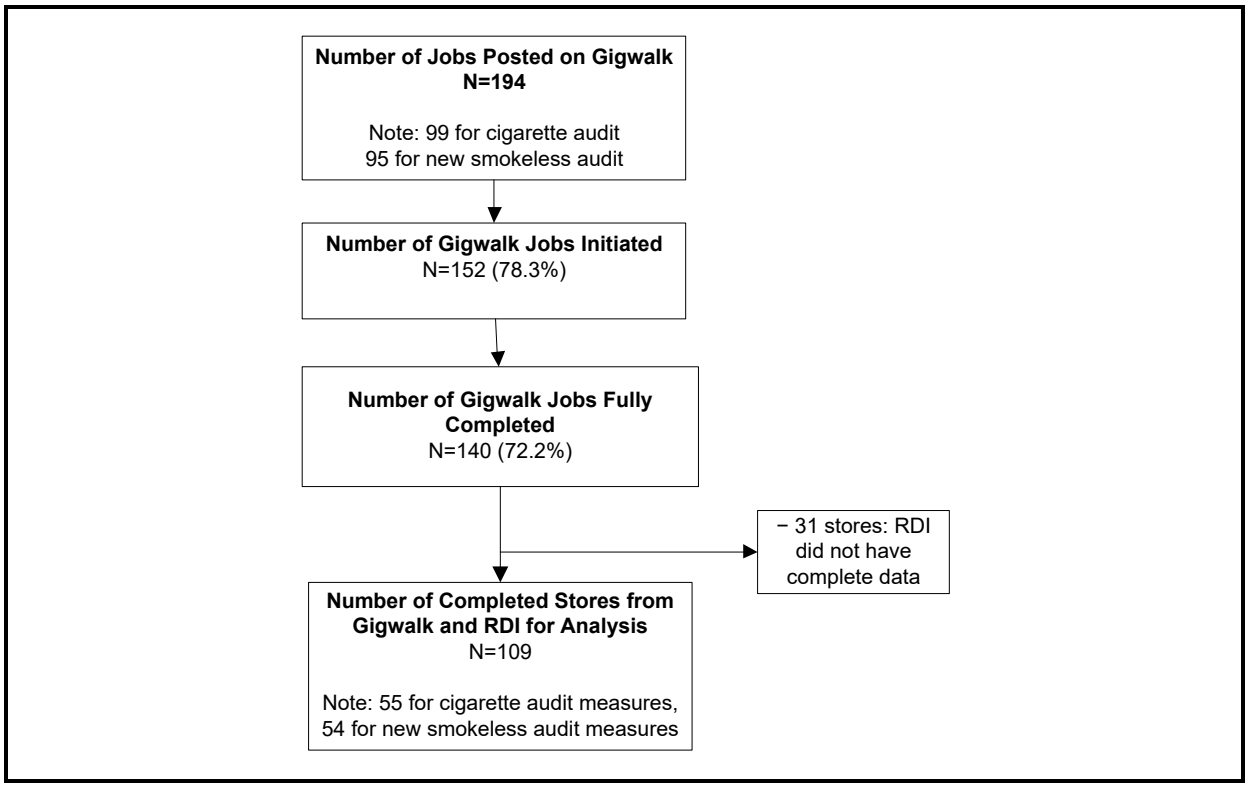
cigarettes (yes/no) and if so, whether any of the following promotions were available: sales prices, multipack discount, mail-in rebate, coupon, or free pack with purchase.

- **"Audit of NEW SMOKELESS tobacco products!" (N = 95 stores)**
  Workers were asked to record whether the store had any interior electronic cigarette (e-cigarette) ads; whether the store sold e-cigarettes and, if so, the brands of the e-cigarettes, and whether any were flavored and to provide a photograph of e-cigarettes; and whether the store sold snus or dissolvable products and, if so, the brands of each.

2.3 Analysis

Of the 194 stores posted only 109 total comparisons could be made with trained data collectors. Figure 2-1 illustrates the loss in sample for our analysis sample.

Figure 2-1. Sample Disposition of Gigwalk Jobs Posted, Completed, and Analyzed



Data were exported from Gigwalk and linked with the trained collectors' data at the store level. Using Stata 12.0, we computed percent agreement, and Cohen's kappa statistics (Cohen, 1960). For presentation in the results section we interpreted kappa values as follows: <0 poor, 0–0.20 *slight*, 0.21–0.40 *fair*, 0.41–0.60 *moderate*, 0.61–0.80 *substantial*, and 0.81–1 *almost perfect agreement* (Landis and Koch, 1977). The proportions of positive and negative agreements are also reported as kappa statistics do not take into account prevalence of the attribute and bias (Sim and Wright, 2005).

The pictures submitted by the data collectors were reviewed and verified by a co-author (AL) and a research assistant who were familiar with the common brand, packaging, and look of ecigarettes, allowing them to more accurately spot and confirm these products in the photos. For all ecig pictures, they reviewed each photograph and confirmed that an image of the product (e.g., ecigs) was present and shown in the photo. In some photos, the data collectors removed the product from the shelf and took a close-up photograph of that product, resulting in very quick and straightforward confirmation of the photo. In other circumstances, the data collector took a photo of an entire wall of products, increasing the need for the photo reviewer to be familiar with the look of the product packaging.

For any cases that verification was not explicit, the picture was reviewed and confirmed by an additional analyst. If a data collector claimed that a product was present, but it was not clearly visible in the submitted photographs, then the product's presence was not confirmed. The process to confirm the exterior ads was similar. In Figure 2-2 you can see an example of the type of pictures uploaded by Gigwalk workers that would be validated.

Figure 2-2. Example Photos showing identification of a validated presence of an item

3. RESULTS

Table 3-1 summarizes characteristics of the Gigwalk data collection. Of the 194 jobs that were posted online, 78.3% were initiated and 72.2% (70 cigarettes audit and 70 new smokeless audit) were fully completed with all questions answered. Twenty-five different data collectors completed an average of 3 gigs ranging from 1 to 25 stores. The median time to complete and submit assignments was 18.1 hours after posting the job online, with 54.7% of the gigs completed within 24 hours or less. Approximately 80.0% of completed audits included photos of tobacco products sold or advertised, and all initiated audits included latitude and longitude for the retail location. Of the initiated audits, 67% were within 100 meters of the mapped location when data was uploaded, 17% were between 100 meters and 1000 meters of the mapped location, and 16% were more than 1000 meters of the mapped location.

Table 3-1. Gigwalk Data Collection Summary

Measure	Statistic
Number of jobs posted	194
Number and percentage of jobs initiated	152 (78.3%)
Number and percentage of jobs completed	140 (72.2%)
Number of unique data collectors responding to jobs	25
Average number of jobs completed by data collector	Median = 3 (range = 1 to 25)
Average time to complete and submit assignment − (time submitted − time posted by RTI)	Median = 18.1 hours (range = 1.2 to 129.7)
Time of job completion after being posted	
Less than 24 hours	84 (55.2%)
1 to less than 2 days	23 (15.1%)
2 to less than 4 days	23 (15.1%)
4 or more days	22 (14.5%)
Number of jobs that provided latitude/longitude coordinates data	152 (100%)
Less than 100 meters within mapped location	102 (67%)
100 to less than 1000 meters within mapped location	27 (17%)
More than 1000 meters within mapped location	23 (16%)
Number of jobs that provided photos	121 (79.6%)

Table 3-2 summarizes agreement between crowdsourced and trained data collectors. There was substantial agreement on exterior cigarette advertisements (85.5%, k=0.71, $p<0.01$) but only fair agreement on interior e-cigarette advertising (88.9% agreement, k=0.21, $p=0.037$).

Table 3-2. Interrater Reliability of POS Measures Collected by Crowdsourcing vs. Trained Data Collectors

Measure	Crowdsourced Data Collectors % (n)	Trained Data Collectors % (n)	%Agreement for Present, Absent % (n)	% Agreement overall, kappa statistic, p-value
Tobacco Advertising				
Exterior cigarette advertising (N = 55 stores)				
Present	50.9%, n=28	54.6%, n=30	86.2%, n = 25	85.5%, k = 0.71, p = 0.000
Absent	49.1%, n= 27	45.4%, n = 25	84.6%, n = 22	
Interior e-cigarette advertising (N = 54 stores)				
Present	3.7%, n=2	11.1%, n=6	25.0%, n = 1	88.9%, k = 0.21, p = 0.037
Absent	96.3%, n=52	88.9, n=48	94.0%, n = 47	
Tobacco Promotions (N = 55 stores)[a]				
Sale offer				
Present	29.1%, n=16	3.6%, n=2	0.0%, n = 0	67.3%, k = −0.07, p = 0.822
Absent	70.9%, n = 39	96.4, n = 53	80.4%, n = 37	
Multi-pack discount				
Present	16.4%, n=9	10.9%, n=6	53.3%, n = 4	87.3%, k = 0.46, p = 0.000
Absent	83.6%, n= 46	89.1%, n = 49	92.6%, n = 44	
Mail-in rebate				
Present	0.0%, n=0	0.0%,n= 0	N/A	100.0%, k = N/A, p = N/A
Absent	100.0%, n = 55	100.0%, n = 55	100.0%, n = 55	
Coupon attached to pack				
Present	3.6%, n=2	0.0%, n=0	0.0%, n=0	96.4%, k = 0.00, p = N/A
Absent	96.4%, n = 53	100.0%, n = 55	98.2%, n=53	
Free pack(s) with purchase				
Present	1.8%, n=1	0.0%, n = 0	0.0%, n = 0	98.2%, k = 0.00, p = N/A
Absent	98.2%, n = 54	100.0%, n = 55	99.1%, n=54	
Product Availability (N = 54 stores)				
Store sells snus				
Present	22.2%, n=12	22.2%, n = 12	66.7%, n=8	85.2%, k = 0.57, p = 0.000
Absent	77.8%, n =42	77.8%, n = 42	90.5%, n=38	
Store sells any brand of dissolvable (e.g., Camel Orbs, Camel Sticks, Ariva)				

Present	7.4%, n=4	0.0%, n = 0	0.0%, n = 0	92.6%, k = 0.00, p = N/A
Absent	92.6%, n=50	100.0%, n = 54	96.2%, n=50	
Store sells e-cigarettes				
Present	42.5%, n=23	33.3%, n = 18	73.2%, n = 15	79.6%, k = 0.57, p = 0.000
Absent	57.5%, n = 31	66.7%, n =36	83.6%, n=28	
Store sells flavored e-cigarettes				
Present	37.0%, n=20	20.4%, n = 11	51.6%, n=8	72.2%, k = 0.34, p = 0.003
Absent	63.0%, n=34	79.6%, n=43	80.5%, n=31	

[a] *Promotions were assessed for Marlboro cigarettes based on posted advertisements and display and not actual purchase*

Crowdsourced and trained data collectors had overall high agreement on coding mail-in rebate (100%), free pack(s) with purchase (98.2%), and coupons (96.4%). However, kappa was either zero or could not be computed because of either perfect agreement (i.e. no variability) or low prevalence rate (i.e. small cell sizes). There was moderate agreement for multi-pack discounts (87% agreement, k=0.46, p=0.000) and poor agreement for sales offers (65.2% agreement, k=−0.06, p=0.837).

When coding product availability, there was moderate agreement for snus (85.2% agreement, k=0.57, p<0.01) and e-cigarettes (79.6% agreement, k=0.57, p<0.01), but only fair agreement for flavored e-cigarettes (73.2% agreement, k=0.35, p<0.01). There was high agreement for dissolvables (92.6%), but kappa was zero due to low prevalence rates.

Photos were examined to validate exterior advertisements and e-cigarette availability. Crowdsourced workers noted that 28 stores displayed exterior advertisements whereas trained data collectors noted 30 stores. We reviewed the photos submitted by crowdsourced workers and confirmed that of the 28 stores, 27 had exterior advertisements, while 1 was miscoded. We also examined the photos of other stores that crowdsourced workers classified as *not* having exterior advertisements and found that 2 stores did in fact have exterior advertisements. If these 2 stores had been coded correctly, the agreement between crowdsourced workers and trained surveyors would be higher. Crowdsourced workers noted that 23 stores sold e-cigarettes and photos were confirmed for 21 stores but 2 stores could not be verified because of low-quality photos. Trained surveyors noted e-cigarette availability in only 18 stores.

.

4. DISCUSSION

Results suggest that crowdsourcing may be a promising form of data collection for some measures in the tobacco retail environment. For most measures related to tobacco product availability, tobacco promotions, and presence of exterior ads, the untrained crowdsourced workers had high agreement with trained data collectors. Agreement was lower for sales offers and interior advertisements, which can be challenging to code without the extensive training that field data collectors received. Crowdsourced workers were only given definitions of what a sales price is, but without extensive training and testing, they may not have interpreted the instructions as we intended. In this study, photos of e-cigarettes and exterior advertisements served to validate the presence but not the absence of products and advertising. We found that crowdsourced workers were more accurate than trained data collectors in coding e-cigarette availability. By asking clerks directly about whether certain tobacco products were available for sale, crowdsourced data collectors may have identified products that were out of view or simply missed by trained data collectors who were attempting to collect data inconspicuously on an extensive audit form. Since Gigwalk workers were instructed to take photos of e-cigarettes if the store sold them, we were able to review the photos submitted to validate their answers.

This study had several limitations. First, store environments could have changed between the time that crowdsourced and trained data collectors visited the same stores, which may explain some disagreements. Second, we could not directly compare the cost between the two approaches because the contractor that managed the data collection did not specify the proportion of the $60 cost per store audit allocated to the trained data collector's wage vs. overhead costs. Third, we were unable to examine whether factors like photo submission, GPS tracking, survey mode (phone app vs. pen and paper), or procedure in store (collecting data inconspicuously vs. interacting with clerks) may have influenced differences in the quality of data collected by crowdsourced vs. trained workers. Fourth, we were unable to validate responses for all measures due to cost constraints. This is a general challenge for the field given that validity assessments were only reported in 6% of tobacco retail audit studies (Lee et al., 2013). Our results suggest that photos may be useful for assessing validity.

An additional limitation was that some uploaded geo-coordinates varied substantially from our mapped geo-coordinates. There are several explanations for why this could have occurred. Inaccuracies in geocoding addresses – which may explain those who were within 1000 meters, but not likely the outlier distances; workers may have been outside of cellular data range within the store, in which case the data would not have uploaded until they came back into data range (e.g. back at home); or, the worker may have visited the wrong store. One strategy would be to ask workers to submit a photo of the store sign with the address number visible so that we can verify they visited the proper location. We can then use geocoordinate data as additional verification.

Crowdsourcing retail audits may have several benefits. First, data collection can be deployed quickly with minimal lead time providing rapid data to policymakers. Second, locals' familiarity with their neighborhood could facilitate data collection. However, this could also be a potential hindrance if assessing retailers' compliance with regulations. Finally, because crowdsourced platforms like Gigwalk have an app that can be downloaded on mobile devices, researchers can easily collect photographs and geolocation data without having to build these capabilities. Photographs can serve as validity assessments or as data sources for coding.

Despite these limitations, the study has several strengths. First, by deploying the crowdsourcing study at the same time that trained data collectors were in the field, we were able to compare the quality of data of crowdsourced workers against the current gold standard of trained data collectors. Second, by having crowdsourced workers report on a range of measures from product availability to sales price, we were able to explore what type of items may be more amenable to crowdsourced than to trained data collection. Third, by having crowdsourced workers take photos of e-cigarette items, we were able to validate the accuracy of their responses.

In conclusion, our results suggest that crowdsourcing may be a viable option for collecting data on the retail tobacco environment. Future studies should examine which POS measures are most amenable to crowdsourcing, in what instances crowdsourced data collectors may be able to collect more accurate data than trained data collectors, and strategies to optimize crowdsourced data collection (e.g., more detailed instructions in the form of a video). Researchers need to also quantify the cost-effectiveness of crowdsourcing compared to traditional trained data collectors to help inform decisions to use crowdsourcing. Finally, studies should be conducted to see whether our results are replicable across locations, crowdsourcing services, and workers.

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. doi: 10.3758/s13428-011-0081-0

Bow, H. C., Dattilo, J. R., Jonas, A. M., & Lehmann, C. U. (2013). A crowdsourcing model for creating preclinical medical education study tools. *Academic Medicine, 88*(6), 766-770. doi: 10.1097/ACM.0b013e31828f86ef

Brabham DC (2013). Crowdsourcing. Cambridge, Massachusetts; MIT Press.

Bradley, J. C., Lancashire, R. J., Lang, A. S., & Williams, A. J. (2009). The Spectral Game: Leveraging Open Data and crowdsourcing for education. Journal of Cheminformatics, 1(1), 9. doi: 10.1186/1758-2946-1-9

Cohen J (1960). A coefficient of Agreement for Nominal Scales. *Journal of Educational and Psychological Measurement* 1960;**20**(1):37–46.

Kim, A. E., Murphy, J., Richards, A., Hansen, H., Powell, R., and Haney, C. (in press). Can Tweets replace polls? A U.S. health care reform case study. In C. Hill, E. Dean, & J. Murphy (Eds.), *Social media, sociality, and survey research*. Hoboken, NJ: Wiley & Sons.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–117.

Lessl, M., Bryans, J. S., Richards, D., & Asadullah, K. (2011). Crowd sourcing in drug discovery. *Nature Reviews Drug Discovery, 10*(4), 241–242. doi: 10.1038/nrd3412

Lee JG, Henriksen L, Myers AE*, et al.* A systematic review of store audit methods for assessing tobacco marketing and products at the point of sale. *Tob Control* 2013 Jan 15. [Epub ahead of print]

Mavandadi, S., Dimitrov, S., Feng, S., Yu, F., Sikora, U., Yaglidere, O., . . . Ozcan, A. (2012). Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study. *PLoS One, 7*(5), e37245. doi: 10.1371/journal.pone.0037245

Merchant, R. M., Asch, D. A., Hershey, J. C., Griffis, H. M., Hill, S., Saynisch, O., . . . Becker, L. B. (2013). A crowdsourcing innovation challenge to locate and map automated external defibrillators. *Circulation: Cardiovascular Quality and Outcomes, 6*(2), 229–236.

Sim J and Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy.* 2005; 85:257-268.

Swan, M. (2012). Crowdsourced health research studies: An important emerging complement to clinical trials in the public health research ecosystem. *Journal of Medical Internet Research, 14*(2), e46. doi: 10.2196/jmir.1988

Winslow, R (2012). The device that saves lives, but can be hard to find. *Wall Street Journal.* Retrieved July 3, 2013, from http://online.wsj.com/article/SB10001424127887324073504578115051054664668.html