Vaping the Web: Crowdsourcing and Web Scraping for Establishment Survey Frame Generation

Bryan B. Rhodes[1], Annice E. Kim[1], Brett R. Loomis[1]

1. RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709

Introduction

Establishment survey sampling frames have typically come either from commercial lists of establishments (e.g., Dun & Bradstreet), government lists based on licensure or registration information, or specific establishment organizations or trade groups. These lists, however, do have some inherent shortcomings that may lead to under-coverage bias and impact survey costs. These shortcomings include missing new establishments, including establishments no longer in business, incomplete or inaccurate contact information, inadequate information about the establishment (e.g., type, size), high costs, and lack of easy access for researchers. For certain studies that are particularly interested in new businesses or a very specific kind of business these shortcomings could introduce unreasonable bias. For this study we developed a new methodology for creating an establishment sample frame that attempts to counter some of these biases. This methodology uses freely available information from the internet and crowdsourcing techniques in an attempt develop a sample frame that is both current and provides sufficient coverage for one particular type of establishment.

Electronic nicotine delivery systems (ENDS) such as e-cigarettes, are battery-powered devices that heat liquid in a cartridge to deliver to the user an inhaled dose of nicotine and other additives, including a humectant (propylene glycol or glycerol) and flavorings. ENDS are sold in traditional retail outlets that are licensed to sell tobacco products such as convenience stores and grocery stores. A recent national study found that ENDS were available in more than 30% of licensed tobacco retailers sampled nationwide[1] and in 71% of licensed tobacco retailers in Florida.[2] At licensed tobacco retail stores, a handful of top ENDS brands are sold alongside traditional tobacco products and other consumer goods. Availability and sales of ENDS at these outlets are being monitored through existing retail audit systems and retail scanner sales data; however, these do not capture data from the increasing number of vape stores that primarily sell ENDS.[3] These vape stores can sell a wider selection of ENDS devices than traditional retail outlets, including higher end tank devices, accessories and liquids/juices that until now were primarily available online. A study of online consumer reviews of vape stores suggests that consumers considered the availability of flavors, devices, and accessories as particularly important in their evaluation of such stores.[4] Some vape stores (e.g., Henley Emporium (http://www.thehenley.com)) also offer tasting bars and 'lounges' designed to encourage socializing and entertainment, thereby, reinforcing positive social norms around ENDS use. Understanding how ENDS are advertised and sold in these specialty vape stores is important given the influence of retail tobacco marketing on behaviors; an extensive body of research shows that retail advertising of tobacco products increases youth susceptibility to smoking and underage retail tobacco purchases, and increases craving and unplanned tobacco purchases among adults.[5] Studies have also shown that retail stores selling tobacco products are disproportionately located in ethnic minority and low-income communities.[6] The extent to which ENDS vape stores

---

[1] Rose SW, Barker DC, D'Angelo H, et al. The availability of electronic cigarettes in U.S. retail outlets, 2012: results of two national studies. Tob Control 2014;23(Suppl 3): iii10–16.
[2] Loomis BR, Hebert C, Dench D, et al. Highlights from the 2013 Florida Retail Advertising of Tobacco Survey: Report submitted to the Florida Department of Health. 2013.
[3] Lee YO, Kim AE. 'Vape shops' and 'E-Cigarette lounges' open across the USA to promote ENDS. Tob Control 2015;24:410–12.
[4] Sussman S, Garcia R, Cruz TB, et al. Consumers' perceptions of vape shops in Southern California: an analysis of online Yelp reviews. Tob Induc Dis 2014;12:22.
[5] Paynter J, Edwards R. The impact of tobacco promotion at the point of sale: a systematic review. Nicotine Tob Res 2009;11:25–35.
[6] Rodriguez D, Carlos HA, Adachi-Mejia AM, et al. Predictors of tobacco outlet density nationwide: a geographic analysis. Tob Control 2013;22:349–55.

are geographically concentrated in certain communities, and whether exposure to these stores influences youth and adult behaviors is largely unknown. The first step in any research around vape stores is determining how many vape stores exist and where they are located.

For this project we looked to develop a list of all vape stores in the state of Florida. Our methodology first uses software to pull information from various sources on the internet about the name and location of potential vape stores. Next we use crowdsourced workers to contact these establishments and determine if they are in fact vape stores. This paper will focus on explaining this methodology and technical details on how it was implemented.

Methods

Web scraping methods

OutWit Hub is a computer software tool used to automatically collect and organize data from websites.[7] This process of using software to collect information from the internet is often called web scraping. Web scraping software can be used to quickly collect standard information from a large number of web pages, and provide structure to this information for later use and analysis.

Web scraping software can visit a user provided list of webpages, or attempt to discover pages organically by attempting to simulate human web browsing. Web scraping software provides a relatively cheap method for collecting and organizing large amounts of data from the web. Because the process is automated it also allows the user to collect data from the web on a regular basis (even as often as daily or hourly).

For this project we used OutWit Hub to scrape information from the web about vape stores in the state of Florida. We identified three main sources to scrape for vape store information: Yelp, Google Maps, and YellowPages.com. Google Maps contains a list of over 95 million businesses and places of interest pulled from Google's web-crawl and searchable using Google's search function. Yelp is a web-based directory of businesses, built primarily by crowdsourcing from users' information and reviews on local businesses. YellowPages is an online directory of businesses modelled on the traditional Yellow Pages listings. Although other sources of information are available on the web these three were chosen due the kind of information they provide, their extensive geographic reach, their perceived complementary nature, and the structure of their webpages.

All three provide business contact information as well as information about the business itself, in particular, for our purposes, each attempts to identify a business specifically as a vape shop. All three also attempt to cover the entire state of Florida (and indeed the entire nation). The websites are also complementary. Google pulls information on businesses from its crawl of the web. Yelp relies on the public to add information on businesses to its site and generate much of its information (rather than crawling the web) and focuses primarily on local brick-and-mortar businesses that may or may not have a presence on the web (in this way it complements Google Maps). The information on YellowPages.com is typically provided by phone companies or the businesses themselves, thereby complementing the other two sources.

In addition to the advantages of the content provided by each source, all three also provide their business information in a semi-structured way. While OutWit Hub and other web scraping software can collect unstructured data from websites, because each of these sites provides a predictable structure to their webpages, capturing and organizing the data was a more efficient process.

Using Outwit Hub we were able to develop a program to scrape information from each source directory. Outwit Hub provides several options that allow the scraping process to be effective and efficient. To use Outwit Hub, however, we have to first tell it how to find the information we want on the webpages (web scraping) and then tell the program what pages we want it to visit to get this information (web crawling).[8]

---

[7] Outwit Technologies. https://www.outwit.com.

[8] Although we used OutWit Hub for this project, the same general methods would be applied using any number of web scraping software programs.

Outwit Hub has several pre-programmed scrapers that look for things like contact information, lists, or tables on the webpage. These can be helpful, but to get all the specific information we wanted in a useful format we needed to develop our own scraper. Below we describe how we told OutWit Hub to find the contact information we are looking for from a page of search results on Yelp.com (a very similar process was used for Google Maps and YellowPages.com).

To scrape the search results webpages we have to tell the scraper how to identify the information we want from the webpage HTML source code. Looking at the source code for the webpage we first searched for HTML code that might indicate the data we want to capture. In the example below, from the website source code, we want to capture the business name (in this case "407 Vape"), so we tell Outwit Hub to capture everything on the webpage between the two highlighted portions.

```
<span class="indexed-biz-name">3.          <a data-
hovercard-id="31YsMSvh1Q3PZqWsCRQS9Q" href="/biz/407-vape-
orlando"class="biz-name">     <span class=
"highlighted">407 Vape</a>
```

Outwit Hub will then look for this specific code sequence (…) anywhere on the page and capture that information (the result is it will collect all business names listed on the page in this format). We then repeat the process for the business address and other contact information we are interested in capturing. The exact code will change for each site, but is usually consistent across all pages on a particular website, so we don't have to set this up for each individual page (i.e., the code is the same on all Yelp.com pages).

Next we need to tell the program what specific webpages to visit to scrape this information. We wanted to capture information from all the search results using several different search terms in combination with locations across Florida. Here we will show you how we were able to breakdown and manipulate the Yelp search URL (the same methods apply to both Google Maps and Yellowpages.com). Since we were interested in pulling information from Yelp search results we first looked at the URL of a typical search result. Below is the URL for a search of the word "vape" in "Orlando, FL."[9]

```
http://www.yelp.com/search?find_desc=vape&find_loc=Orlando%2C+FL&ns=1
```

There are three key features of this URL which we will be able to manipulate so we are able to search multiple terms across multiple cities effectively. The first is the location of the search term in the URL. In the URL below we've highlighted the location of the "vape" search term.

http://www.yelp.com/**search?find_desc=vape**&find_loc=Orlando%2C+FL&ns=1

Next we look for the location that we entered. The highlighted section in the URL below shows this.

http://www.yelp.com/search?find_desc=vape&**find_loc=Orlando%2C+FL**&ns=1

[9] It is important to try several different searches to determine if the URL for the search result pages takes on the same general format.

The last portion of the URL that is worth noting is at the end. This portion of the URL indicates what page of the search results. Often times search results are listed across multiple pages. The highlighted portion in the URL below, shows the search result page number.

http://www.yelp.com/search?find_desc=vape&find_loc=Orlando%2C+FL&**ns=1**

Using this information we can then create a program in Outwit Hub that will load all the pages where we would like to scrape information. This program is represented by a long URL that specifies each of the criteria and where it would fit in the URL, an example looks like this:

www.yelp.com/search?find_desc=**[vapin;vape;vapor]**&find_loc=**[Orlando;Miami;Tampa]**%2C+FL&ns=**[1-10]**

The highlighted portions show where we have entered the various criteria we want to explore (we have reduced the list for this example). Outwit Hub will look at all combinations of these criteria to create essentially a list of URLs to visit. Outwit Hub also has several options including "fast scraping" and the ability to automatically navigate through search result pages. These options can be helpful, but may or may not work on a particular website and so thorough testing is recommended.

Outwit Hub now knows what webpages to visit and what information to pull from each page. We then execute the program and Outwit Hub is able to capture the information into an MS Excel file. Depending on the website and the information being scraped it can take anywhere from a few minutes to a few hours for the program to execute.

For this project we used ENDS or vape-related search terms ('ecig,' 'e-cigarette,' 'vape,' 'vapor,' 'vaper,' 'vapin') on all three sites, and searched in the 409 cities/towns in Florida.[10]

Crowdsourcing with Mechanical Turk

Scraping the three online directories resulted in a list of potential vape stores in the state of Florida. We next cleaned the list by removing duplicates and establishments that were obviously not vape stores (e.g., churches, city health departments). Our final list consisted of 1,459 potential vape stores in Florida. Since not all of these were likely to meet our definition of a vape store (e.g., a store may be a traditional retailers that happens to sell some ENDS products, rather than a specialty vape store), we next wanted to determine which businesses were actual ENDS vape stores of interest. Traditionally, this might be done by reviewing administrative records or by having professional data collectors call or visit the establishments. However, given the relatively recent rise of ENDS vape stores, administrative records for ENDS vape stores are limited. Furthermore, given the large number of search results, having professional data collectors visit these establishments would be time consuming and potentially cost prohibitive. To overcome these challenges, we used a crowdsourcing platform, Amazon Mechanical Turk (MTurk), to verify the information we had gathered.

MTurk allows users to post discrete jobs or tasks called Human Intelligence Tasks (HITs) to be completed by an online network of MTurk workers. These tasks are typically small jobs that can be quickly completed by a human worker but are often difficult for a computer, such as identifying objects in a digital image, gathering location data

[10] City and town list from the Florida Department of State website: http://dlis.dos.state.fl.us/fgils/cities.html.

on retailers[11, 12] and sentiment analysis of social media posts.[13]

For this project we submitted successive jobs to help us determine which of the establishments in our list was truly a vape store of interest. For the first job we asked MTurk workers to call each establishment (n=1,459) and determine (1) whether the business sells ENDS (yes/no), (2) whether the business primarily sells ENDS and/or ENDS-related juices/fluids or accessories (yes/no), and (3) whether the business sells other tobacco products (e.g., cigarettes, cigars) (yes/no). These questions were asked to decipher the ENDS vape stores of interest from businesses that either did not sell ENDS or sold ENDS along with other tobacco products (e.g., tobacco shop, convenience store).

In the second job, we submitted the list of businesses that were positively confirmed as primarily selling ENDS (n=442) and asked MTurk workers to call the establishment and determine (1) whether the business sells ENDS primarily to individual customers or other businesses like a distributor (customer/distributor), and (2) whether the business has a storefront that is open to the public (yes/no). These questions were asked to decipher the ENDS vape stores of interest from businesses that were distributors/resellers or web-only retailers with no physical brick-and-mortar storefront for selling directly to individual consumers.

As a quality control measure each HIT was completed by three separate MTurk workers for each establishment, so each establishment is called three times and asked the same questions. This allows us to compare answers across workers and determine a consensus answer. If no consensus is reached we re-submitted the HIT to be completed by three new MTurk workers.

MTurk workers were paid $0.25 per completed HIT, and workers could work on multiple HITs (meaning a worker could call multiple establishments). This resulted in a total cost of $1,568.33:

- HIT 1: 1,459 establishments called by 3 workers each at $0.25 each ((1,459*3)*$0.25 = $1,094.25
- HIT 2: 442 establishments called by 3 workers each at $0.25 each ((442*3)*$0.25 = $331.50
- Amazon Mechanical Turk charged a 10% fee on top of all payments to workers = $142.58.[14]

HITs were submitted to MTurk at the start of the work day so workers would be more likely to reach an open business when they called. Each set of HITs was completed in under 8 hours.

As an additional layer of quality control MTurk provides information that can help identify workers who may be falsifying answers. For example using data provided by MTurk we can determine if a single worker is found to be in the minority (i.e., providing a different answer than two other workers calling the same establishment) on a regular basis. This allows us to determine if a worker is potentially falsifying data or not completing the HIT properly. We can also examine how long it took for a worker to complete a HIT, if a worker is completing a HIT much more quickly than other workers that can warrant further investigation. If there is any doubt about the validity of answers provided by a particular worker all answers are thrown out and new HITs are submitted to capture this information again.

Results

In that state of Florida ENDS retailers are required to register with the state as tobacco retailers. For this project we were able to compare the results of our methodology to the state list of tobacco retailers. When we compared our

---

[11] Keating MD, Rhodes BB, Richards A. Crowdsourcing: a flexible method for innovation, data collection, and analysis in social science research. In: Hill CA, Dean E, Murphy J, eds. Social media, sociality, and survey research. New York: Wiley, 2013:179–202.

[12] Keating MD, Furberg RD. A methodological framework for crowdsourcing in research. Federal Committee on Statistical Methodology Research Conference.
Washington DC 2013.

[13] Kim AE, Lee YO, Shafer P, et al. Adult smokers' receptivity to a television advert for electronic nicotine delivery systems. Tob Control 2015;24:132–5.

[14] This has recently been increased by Amazon to 20%.

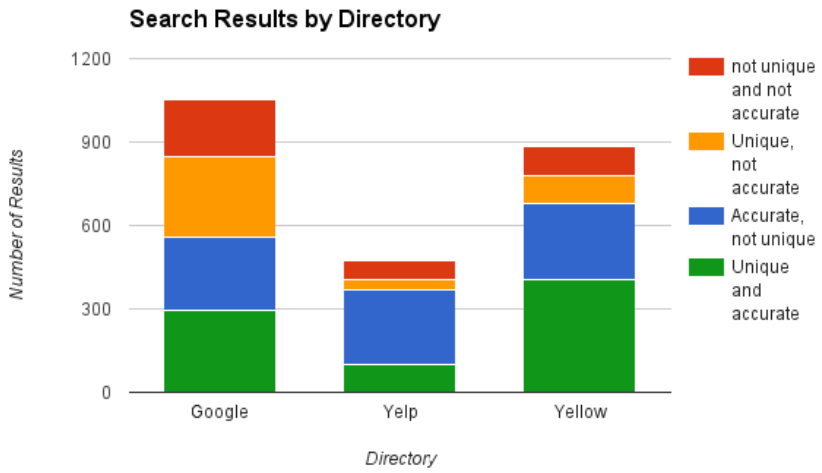
final list of vape stores from the online search methodology (n=403) to the full tobacco licensure list (n=29,039), we found that 131 of the 403 stores were on the licensure list (32.5%), while 272 stores were not (67.5%). This analysis is further discussed in Kim et al. (2015).[15]

In addition to comparing the results of this methodology to a more standard list of retailers, we also wanted to measure the effectiveness of this particular protocol. This way we could identify ways of refining the protocol to improve results.

We first looked at the three online directory sources: Yelp, Google Maps, and Yellowpages.com. If one of the sources was found to be duplicative of the other two or provided a large number of inaccurate results we would want to consider if we could refine our search of the source or look for a better source of information.

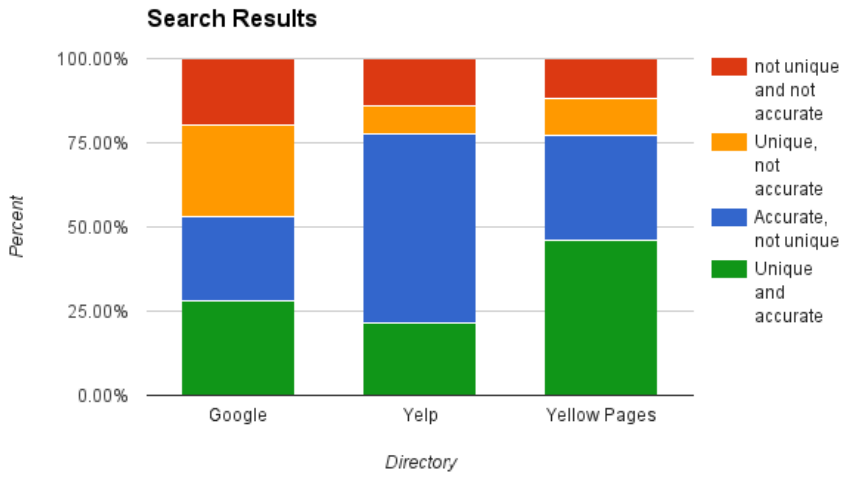
Figures 1 and 2 below provide an overview of the results from each source. Each result was examined to determine if 1) it was not found by either of the other two sources (i.e., it was a unique result), and 2) if the result was obviously not a vape shop (i.e., it was an accurate result). A source that provides a high proportion of accurate results is valuable, as inaccurate results are not useful and require additional labor to remove. The number of unique results is also a good measure of a source's value. If a source provides a large number of accurate results, but all of them are also found by the other two sources that particular source has not added anything. The number of accurate results that are *also* unique to just one source are the most useful results. Figure 1 shows the raw number of results for each source. While Google provided the most overall search results (n=1054) it also returned the most results that were determined to be inaccurate (n=495). Yellowpages.com found fewer total results (n=882), but its results were more likely to be unique and accurate (n=405). In general, however, each source found at least 100 results that were both unique to that source and accurate. Figure 2 shows the results proportionally.

Figure 1. Search result effectiveness by directory – number of results



[15] Kim, A. E., Loomis, B., Rhodes, B., Eggers, M. E., Liedtke, C., & Porter, L. (2015). Identifying e-cigarette vape stores: description of an online search methodology. *Tobacco Control*, tobaccocontrol–2015–052270. http://doi.org/10.1136/tobaccocontrol-2015-052270

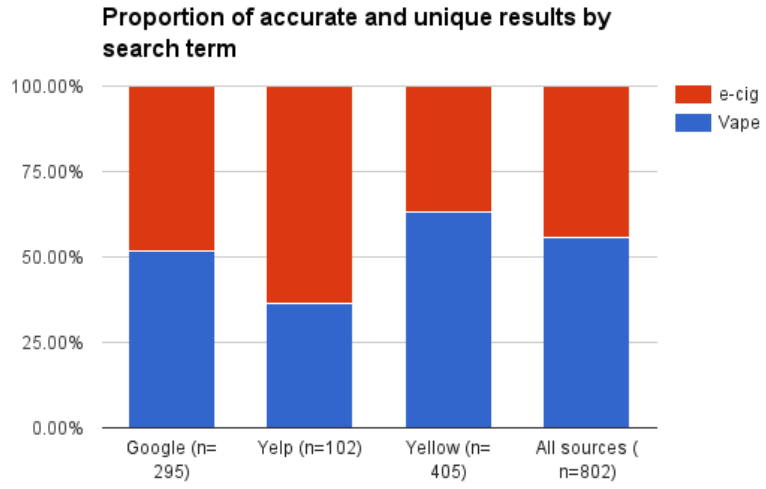
Figure 2. Search result effectiveness by directory – proportion of results



Term analysis

We also examined the relative efficacy of the search terms we used with each source directory. We used six total search terms, four terms were related to vape ("vape," "vapor," "vapin," and "vaper") and two terms related to e-cigarettes ("e-cig" and "e-cigarette"). Figure 3 shows the proportion of the unique and accurate results for each source by search term used. Most of Yelp's most valuable results were found with the "e-cig" related terms. Yellowpages.com on the other hand found most of its best results when searching using the "vape" terms. For Google the proportion was split about evenly. This shows how different terms may work better or worse depending on the directory being searched.

Figure 3. Proportion of accurate and unique search results by search term for each directory



Discussion and Conclusion

This methodology for producing a list of establishments presents several advantages. The first is the relatively low cost and the speed with which the list can be developed. We were able to produce the list in about two days for relatively low costs. This allows researchers the ability generate an up to date list as needed. Because the online

source information is constantly being updated, researchers could also potentially use this methodology to track the growth or decline of a particular type of establishment over time. In addition because this methodology allows researchers to identify very specific types of establishments, it also could save money over other methods that would require the researcher to screen a large number of establishments to identify those of interest.

This methodology, however, is not ideal for generating some kinds of establishment frames. Since many of the directories used in this research focus on retail establishments, this methodology is not necessarily very useful in identifying non-retail establishments (e.g., manufacturers, accounting firms, distributers). That being said, if any online directories exist of a specific type of establishment, this methodology could likely be modified to pull useful information from these directories. In addition, for many types of establishments a "gold standard" list already exists. In these situations this methodology may not be necessary, but it could be worth exploring as a method for potentially expanding the existing list for relatively low cost.

While this methodology is promising for developing specific kinds of establishment sample frames, there are several things to consider before implementing these methods. The first is the terms of use of the websites where a researcher is planning to scrape. Some website terms of use may explicitly prohibit automated web scraping from their site. It is important researchers first look closely at any website terms of use before scraping data from a site. Even if web scraping is not explicitly forbidden by a website's terms of use it may not be perceived favorably by the website owner and may result in banning of the user's IP address. While scraping data on the scale described here is not likely to go noticed by the website administrators of large, high-traffic sites, it is still important to consider the impact on the site when implementing this methodology.

Another consideration is the use of application program interfaces (APIs). APIs are essentially a set of protocols, established by the website itself, that allow a user to automatically access data from a specific website. Through an API a user can request specific information from a website, and receive that information back in a structured format. For websites that provide an API this may be a preferable method (over web scraping) for gathering information. APIs, however, do have some disadvantages compared to web scraping. APIs are not available for many websites (e.g., Yelp and Google have them, but Yellowpages.com at this time does not). APIs also can be limited in the kind or amount of information that can be requested from the site. For example some APIs may only allow the user to access a limited number of search results or provide the geographic coordinates for an establishment rather than a street address.

This methodology for developing establishment frames can and should be further refined going forward. More research is needed in testing this methodology to develop frames for other types of establishments, and comparing these frames to "gold standards" that may exist. In addition, researchers who use this methodology should take advantage of the speed and relatively low cost of this methodology by looking to refine their processes iteratively by exploring multiple web sources and search terms to help identify those that are most effective and efficient.

References

Keating MD, Furberg RD. A methodological framework for crowdsourcing in research. Federal Committee on Statistical Methodology Research Conference. Washington DC 2013.

Keating MD, Rhodes BB, Richards A. Crowdsourcing: a flexible method for innovation, data collection, and analysis in social science research. In: Hill CA, Dean E, Murphy J, eds. Social media, sociality, and survey research. New York: Wiley, 2013:179–202.

Kim AE, Lee YO, Shafer P, et al. Adult smokers' receptivity to a television advert for electronic nicotine delivery systems. Tob Control 2015;24:132–5.

Kim, A. E., Loomis, B., Rhodes, B., Eggers, M. E., Liedtke, C., & Porter, L. (2015). Identifying e-cigarette vape stores: description of an online search methodology. *Tobacco Control*, tobaccocontrol–2015–052270. http://doi.org/10.1136/tobaccocontrol-2015-052270

Lee YO, Kim AE. ‘Vape shops’ and ‘E-Cigarette lounges’ open across the USA to promote ENDS. Tob Control 2015;24:410–12.

Loomis BR, Hebert C, Dench D, et al. Highlights from the 2013 Florida Retail Advertising of Tobacco Survey: Report submitted to the Florida Department of Health. 2013.

Paynter J, Edwards R. The impact of tobacco promotion at the point of sale: a systematic review. Nicotine Tob Res 2009;11:25–35.

Rhodes, B. B., & Keating, M. D. (2013, May). *Are we asking the right questions? An exploration into crowdsourcing survey questions.* Poster presented at American Association for Public Opinion Research 68th Annual Conference, Boston, MA.

Rodriguez D, Carlos HA, Adachi-Mejia AM, et al. Predictors of tobacco outlet density nationwide: a geographic analysis. Tob Control 2013;22:349–55.

Rose SW, Barker DC, D’Angelo H, et al. The availability of electronic cigarettes in U.S. retail outlets, 2012: results of two national studies. Tob Control 2014;23(Suppl 3): iii10–16.

Sussman S, Garcia R, Cruz TB, et al. Consumers’ perceptions of vape shops in Southern California: an analysis of online Yelp reviews. Tob Induc Dis 2014;12:22.