Searching for Information Online

Using Big Data to Identify the Concerns of Potential Army Recruits

Salar Jahedi, Jennie W. Wenger, Douglas Yeung

Key findings

- Google search queries can be used to better understand how interest in military careers has evolved over time and geographic location.
- It is possible to use these tools to identify the chief Army-related concerns that potential recruits have, including the qualifications for, procedures for, or benefits of enlisting.
- It is possible to predict with reasonable accuracy what individuals were searching for months before or after searching for Army-related terms
- Including Google Trends terms in a model of factors influencing the number of Army accessions increases the predictive power of the model.

SUMMARY ■ In this report, we assess some empirical applications of web search data and discuss the prospective value such data can offer to Army recruiting efforts. We discuss three different tools—Google Trends, Google AdWords, and Google Correlate—that can be used to access and analyze readily available, anonymous data from Internet searches related to the Army and to Army service. We find that Google search queries can be used to better understand how interest in military careers has evolved over time and geographic location, and even identify the foremost Army-related concerns that potential recruits experience. Moreover, it is possible to predict with reasonable accuracy what non-Army related terms people are searching for in the months before or after an Army query. Finally, our results suggest that search terms can serve as a measure of propensity and can be used to predict the overall proportion of highly qualified Army accessions. We close with a brief discussion of the implications that can be drawn and fruitful areas for future research.

INTRODUCTION

Despite the high trust that the American public places in the Army, many young people today appear to have a lower propensity to serve and a lower likelihood of contacting a recruiter than in past years. At the same time, the Internet has become an increasingly vital source of information, especially for young people. Both of these trends have implications for recruiting. In this report, we focus on exploring the potential of new analytic techniques for using Internet data to improve recruiting. In particular, these new methods can benefit the Army and help to understand youth attitudes, concerns, and interests; and ways to predict or quantify trends in youth propensity for the military (i.e., interest in military service) and recruitment decisionmaking. Traditionally, the Army has obtained information about attitudes and propensity from youth surveys. Surveys are advantageous for several reasons, including the specificity of data collected, but recruiting survey samples are relatively small, potentially unrepresentative, and cover only a limited number of topics. More importantly, there is a substantial delay between the time the data are collected and when they are available for analysis. This suggests that other sources of information might complement surveys, especially if these sources could track information in something closer to real time, or could reveal a wider range of information, including that which the Army currently does not collect. "Big data"—which refers to the extremely large data sets made possible through the Internet, mobile devices, and other sources—have the potential to provide such information.

In this report, we focus on three specific research questions related to how the Army might use big data:

- How have Army-related searches changed over time and across locations?
- What sorts of questions and concerns are prevalent in Army-related searches?
- How is the number of relevant searches related to the number of people who enlist?

The remainder of this report is divided into four sections:

- In the next two sections, we define "big data" and provide a brief overview of the types of analyses that are possible with big data, focusing on analyses that may be especially relevant to Army recruiting.
- In the third section, we discuss three tools that the Army might consider using, and demonstrate how each can be used to answer aspects of the questions above.
- The final section details insights, lessons learned, and suggestions for future research on the use of Internet search data in Army analyses.

WHAT IS BIG DATA?

In recent years, big data have emerged as an important addition to traditional data sources derived from behavioral research (which generally involves data collection via either surveys or accessing administrative records, followed by analysis). The term "big data" is evolving and refers to extremely large data sets derived from the Internet, mobile devices, sensors, and other sources, as well as the wealth of available information that, if analyzed appropriately, can reveal valuable insights (Manyika et al., 2011).

In most cases, working with big data requires specialized analytical tools that make it computationally feasible to process and analyze large amounts of complex information. Analyzing such data may require access to and careful handling of large volumes of relatively unstructured data, including personally sensitive or identifiable information. Other necessary capabilities include front-end tools that allow analysts to interact with data (e.g., visualization), and back-end tools (i.e., hardware and software infrastructure) that can process the large volumes of data. For this reason, big data analysis often involves knowledge discovery rather than hypothesis testing—i.e., it is often used more for finding correlations and making predictions than for inferring causal relationships.

"Big data" refers to the extremely large data sets made possible through the Internet, mobile devices, and other sources.

These large data sets can come from many different sources. Everyday gadgets, such as smartphones, are constantly gathering information about people's behaviors, as are wearable sensors that measure health and fitness, or smart home technologies that automate lighting, security, and energy use (i.e., the "Internet of things," Schindler et al., 2013). Most actions on the Internet are automatically recorded by software. For instance, such Internet search engines as Google or Yahoo! often gather anonymized data regarding the topics that people search for, as well as the date and location of these inquiries. In some cases, people use social media, such as Facebook or Twitter, to make public more-detailed information about their thoughts, as well as pictures, videos, and sometimes, their location.

Because social media posts contain information that people actively choose to share, they can provide a rich source of insight toward understanding attitudes and opinions. Naturally, the data are limited to the extent that people choose to present themselves. A somewhat larger drawback is that this information is not readily available; it generally requires contracting with a commercial provider that may have exclusive access. However, these providers typically can help organize the data into workable data sets. Given that social media data reside predominantly on commercial platforms, access to these data is subject to the profit motives and specific terms that social media companies and third-party providers dictate.

In contrast, Internet search data offer several advantages over other types of big data, particularly for the Army. Assuming that people search online for information they do not possess but find useful, Internet search data are likely to provide valuable insights into youth preferences and interests that may be otherwise unavailable. Thus, while Internet search data may not be wholly representative of the population, several factors suggest that people in the Army's target population are likely to search for information online. First, younger people are more likely to use the Internet in general: 98 percent of 18- to 29-year-olds use the Internet for any reason, the highest proportion of any age group (Pew Research Center, 2013). Those who do not go online are more likely to be older (Zickuhr, 2013). Furthermore, Yeung and Gifford (2011) found that potential military recruits ask a wide range of questions in online forums, suggesting that these and other information needs may be further reflected in usage of online search engines.

Google makes aggregated and anonymized search data available to the public; this offers several advantages. The aggregated data contains no personal information at all, potentially allaying many privacy concerns. Also, Google has internalized much of the computational costs of handling the data; this information is not only free, but it is easy to access and surprisingly easy to use. This is a huge benefit in the sense that it does not require significant resources to access valuable information. At the same time, only a partial glimpse of the data is available and a lot of rich analysis is simply not possible. Nevertheless, the Google platform is particularly useful for capturing trends in how people are searching for and accessing military career information and resources. Trillions of web searches are conducted worldwide every year, and Google is by far the leader in this marketplace: 83 percent of U.S. search users most often use Google (Purcell et al., 2012). Google provides anonymous, aggregated volume data at the weekly level on the queries that are searched. Thus, using Google's publicly accessible analytic tools to explore web searches could help to identify aggregated statistics on youth perspectives and information searches related to Army recruiting.

Internet search engines such as Google or Yahoo! often gather anonymized data regarding the topics that people search for, as well as the date and location of these inquiries.

WHAT CAN INTERNET SEARCH DATA TELL US ABOUT ATTITUDES AND TRENDS?

Over the last decade, researchers have begun to use Internet search data to learn about human behavior. The content of searches can shed light on a wide variety of people's concerns, ranging from dieting to divorce. Furthermore, researchers have found that the volume of searches conducted for queries is

predictive of important outcomes, from health to employment status. For example, Ettredge and colleagues (2005) analyzed search terms that might be used by people who are seeking employment ("job search," "jobs," "monster.com," "resume," "employment," and "job listings") to study whether the volume of searches for those terms could predict forthcoming federal monthly unemployment reports. The authors found that the total number of searches during each week in 2001–2003 was indeed correlated with the unemployment figures published by the Bureau of Labor Statistics.[1]

Internet search data have also been used to analyze trends in health care. Cooper and colleagues (2005) showed that searches for cancer-related terms were positively related to the American Cancer Society's estimates of cancer incidence and mortality. A succession of papers followed, seeking to examine whether Internet searches could predict influenza outcomes. By far the most influential paper was by Ginsberg and colleagues (2009), who demonstrated a novel method to pick the search terms that will be most predictive of flu outcomes. They took the 50 million most-commonly searched terms on Google and tested the correlation of each to the percentage of flu-related physician visits over the past five years. They then created an index of search terms using the 45 highest search queries as their predictor, finding it to be highly accurate in predicting in real-time the number of flu-related doctor visits, as reported by the Centers for Disease Control and Prevention (CDC).

This example illustrates the occasionally serendipitous nature of big data—Google's services were *not* explicitly intended for disease tracking, but researchers discovered that user search data could be repurposed in this manner. The CDC has collaborated with Google to incorporate this information into its flu estimates.[2] But this example also serves to illustrate the challenges of data repurposing and building complex algorithms that often remain opaque to the users. Analyses of Google Flu Trends' estimates, based on existing search data, have suggested that its initial predictive ability was significantly overstated (e.g., Olson et al., 2013; Lazer et al., 2014). However, as Lazer and colleagues pointed out, Google's flu data in *combination* with the CDC's data allowed for predictions that were more accurate than either set of data in isolation. Such challenges are important to note, suggesting not that big data approaches to behavioral research and policy analysis should be discarded, but that their implications need to be carefully considered.

The widely reported development of Google Flu Trends activated an interest in conducting research using search data more generally. A number of these efforts focused on understanding economic behavior.[3] The recent work of Seth Stephens-Davidowitz, reported in a series of New York Times articles, demonstrates ways to combine Google search data with more traditional data to shed light on phenomena that are difficult to measure with traditional survey tools (Stephens-Davidowitz, 2013). For example, he argued that child abuse was significantly underreported during the Great Recession, noting that areas with the largest cuts in social services spending reported lower numbers of child abuse cases, but in those same areas, searches for phrases such as "my dad hit me" or "child abuse signs" increased.

Other examples further suggest how Internet search data have the potential to reveal concerns that may be difficult to elicit on traditional surveys. Using another Google tool, which provides search volume data on exact words searched and suggestions for similar searches that are popular, Stephens-Davidowitz (2014b) found that pregnant women across the world searched for such concerns as whether they could drink alcohol or cold water. He also demonstrated that parents are more likely to search "is my son gifted" than conduct a similar search for daughters (Stephens-Davidowitz, 2014a). In fact, for many intelligence-related terms (e.g., "a genius," "intelligent," "stupid," "behind"), search activity focuses more heavily on sons; the opposite is true of searches related to beauty.

In many respects, the use of Internet search volume data in research is still in its infancy. Researchers are constantly seeking to develop new methods by which to take advantage of this rich data source, and many new methods are likely to be available in the near future. These methods may help us to better understand the attitudes and concerns that are implicitly expressed by search behavior, and, specifically, how the Army can understand the recruiting concerns of the youth population.

What Can Internet Search Data Tell Us About Attitudes Toward Military Service?

Data from search queries consist of a record of the terms people enter into search engines and the frequency with which various words are searched. Thus, data from search queries implicitly communicate the topics users are interested in. Although the resulting search information is available only in an aggregated form, a key advantage of these data is that no new collection is required; therefore, the analysis can be carried out quickly and with limited resources.

Here, we explore the extent to which information from Army-related Internet searches may be able to capture trends

These trends would give interesting real-time information to the Army that it could use to effectively manage recruiting.

that reveal youth attitudes and concerns, such as the perceived benefits and drawbacks of military service. This method could help the Army understand the extent to which the youth population is concerned about potentially negative aspects of service and how these perceptions have changed over time or in response to specific events as they occur. Correlation analyses could also reveal the concerns of specific sub-groups of youth—for example, young people who are interested in military service but who lack citizenship may be hesitant to contact recruiters; tracking searches in regard to aspects of military service *and* citizenship could provide insights into the size of this market, as well as the concerns and interests of this sub-group. Analyzing Internet search data also has the potential to provide more-detailed information about youth interests and concerns. In particular, it is possible that general searches specific to Army benefits, such as education benefits, training, or loan-repayment, provide useful information about key economic trends and patterns that can ultimately be used to help with recruiting.

Beyond revealing a population's specific questions and concerns, Internet search data may also provide a measure of the trends over time concerning interest in military careers, as well as differences across regions of the country. Such search data could provide a measure of propensity to join the military and how propensity has changed over time. These trends would give interesting real-time information to the Army that it could use to effectively manage recruiting, especially as the civilian labor market continues to improve and desirable recruits have a wider range of career options available. To provide a simple example, such an analysis could focus on terms such as "join the Army," and could examine how the overall number of such searches has varied over time and across states or regions. The Army might also use search data to understand the extent to which specific advertising campaigns or information are reaching target audiences. Finally, in the same way that searches on flu-related terms have added explanatory power to models built on other types of data, the Army may be able to use search data, along with existing data sources, to improve the explanatory power of recruiting models.

In the following sections, we use Internet search data to explore both general interest in Army careers and the ways in which this interest has varied over time and geographic areas. We also consider patterns in search queries that may be related to key recruiting markets or benefits. In this discussion, we focus on three different publicly available tools provided by Google. These tools report on similar underlying data on search terms, but differ slightly in their methodologies and outputs:

- Google Trends returns frequencies across time and location for broad searches
- Google AdWords returns search volumes for exact searches
- Google Correlate helps to find search terms that are similar to search terms of interest, or that are related to user-uploaded data (e.g., seasonal or geographic patterns).

We discuss each tool in turn, and use our three research questions to demonstrate the capacity of these tools.

INTERNET SEARCH DATA: THREE PUBLICLY AVAILABLE ANALYTIC TOOLS AND ARMY-RELEVANT EXAMPLES

Internet search data can be used to answer many questions that are relevant to Army recruiting. Here, we focus our attention on Google searches, because of the widespread use of the platform and the availability of analytic tools. We describe each tool in turn, providing relevant examples for the Army related to the three questions introduced earlier:

- How have Army-related searches changed over time and across locations?
- What sorts of questions and concerns are prevalent in Army-related searches?
- How is the number of relevant searches related to the number of people who enlist?

The tools are all similar in the sense that they deliver data on search frequencies across time and location, but they differ

enough in their attributes that the types of analysis that can be done varies across the tools. Table 1 offers a brief description of each tool and some of its chief features. The tools will be able to answer the questions posed above from different perspectives.

Google Trends

Google Trends is a public database that returns the frequency at which a given query is searched, relative to all Google searches within a given region, the default being set to the United States. For any term(s) searched, Google Trends will plot the popularity of that term across time (at the weekly level, going back to 2004) or across geography (at the country, state, metro, or city-level aggregated over a given time period). With some caveats, this tool is well-suited to explore our first research question: "How have Army-related searches changed over time and across locations?"

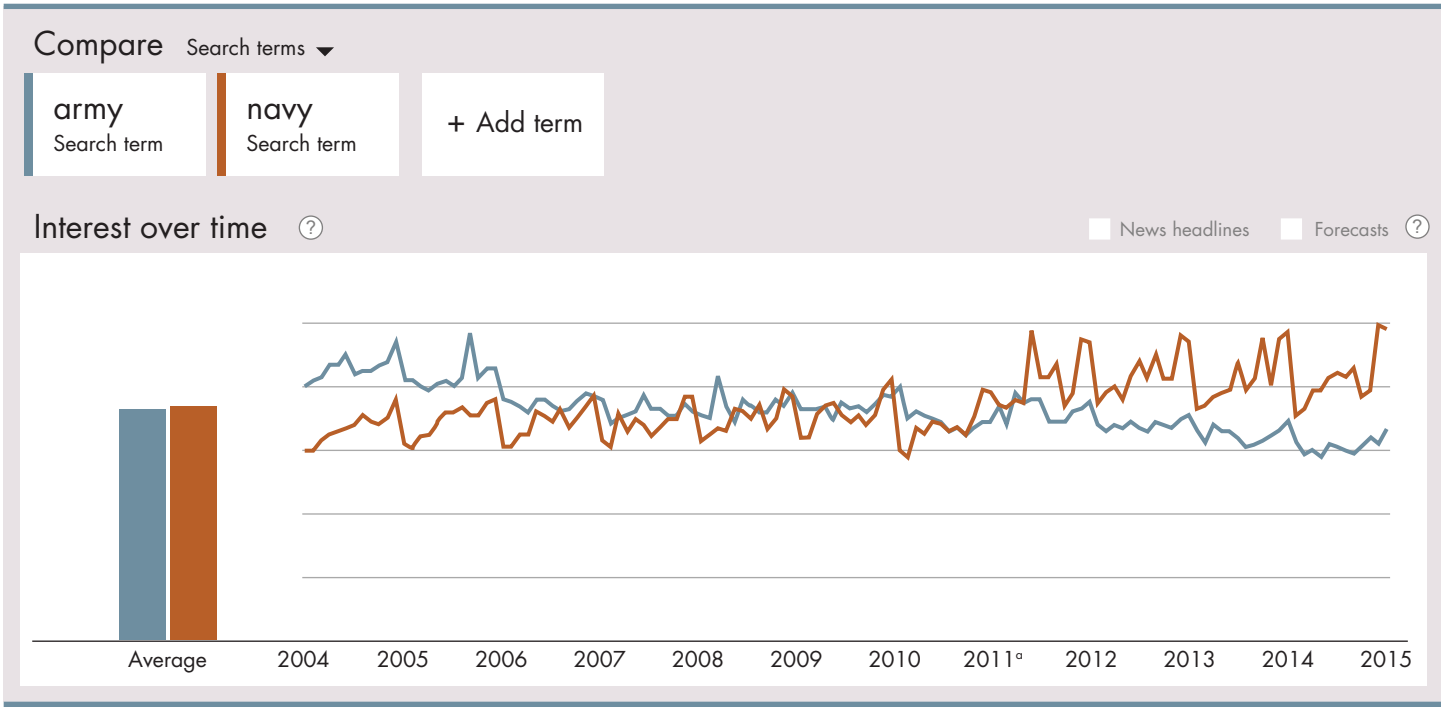
It is important to understand how Google Trends rates and aggregates search queries. Google Trends assigns the period with the highest search volume a value of 100 and scales the value assigned to all other dates to be a percentage of the peak value. The data come from broad matches, meaning that searches for the term "apple" include searches for the terms "red apple" as well as "apple iPhone." The interpretation of the data can be tricky, given that the math behind the normalization is unobservable, but relative comparisons are fairly straightforward by plotting two separate terms on a single figure. Figure 1 plots the output from two such queries: (1) "army" and (2) "navy." In 2004, the term "army" was more popular than the term "navy." Over time, however, the term "army" slowly decreased in popularity while the term "navy" increased in popularity.[4] By 2015, it is clear that "navy" is a relatively more popular search term than "army."

It may be tempting to conclude that this says something about the popularity of the branches of the Armed Forces—in other words, that people search for the Navy more frequently than they search for the Army. However, recall that the search results are provided for broad terms, meaning that a search for "army" includes terms like "Salvation Army" and a search for "navy" includes terms such as "Old Navy." Removing these terms from the results, as shown in Figure 2, "army" (minus "salvation") has a larger search volume than "navy" (minus "old").

Table 1. Comparison of Google Tools

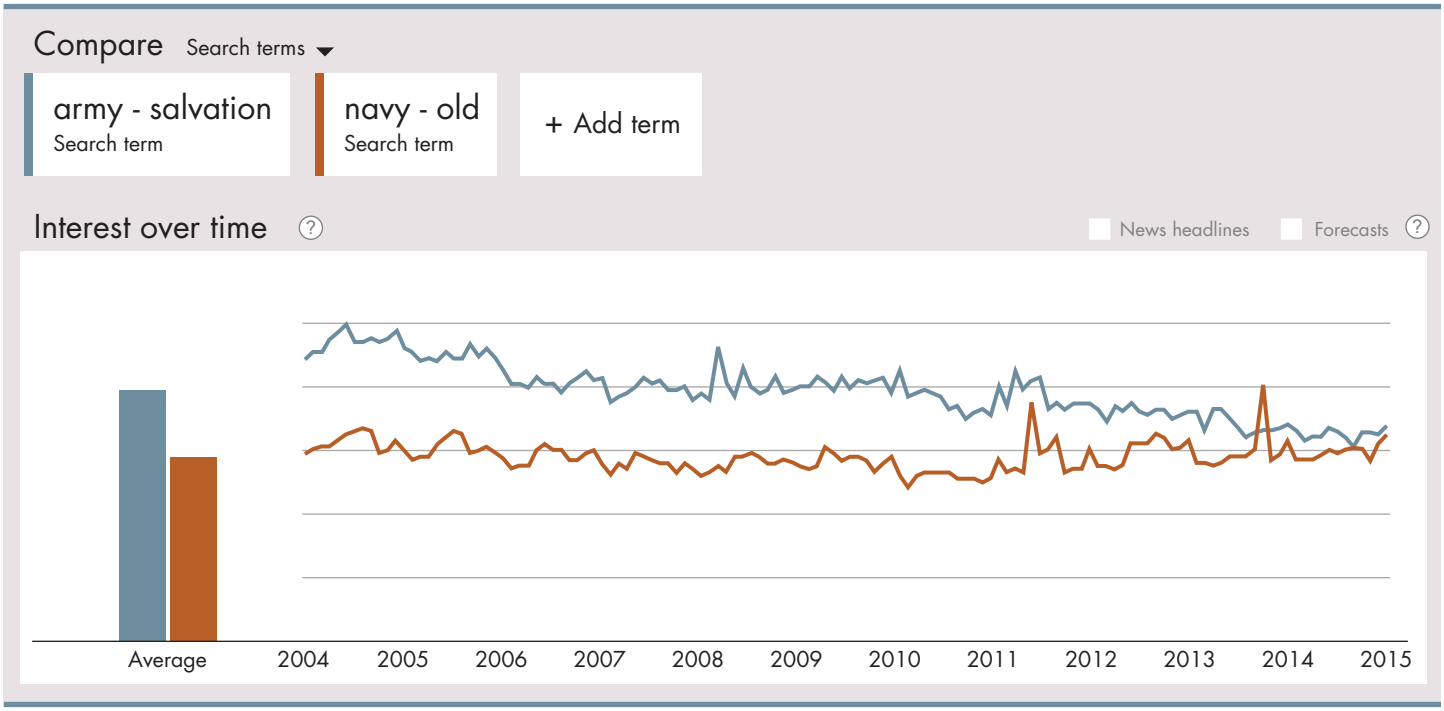
	Google trends	Google AdWords	Google correlate
Usage	Compare search term(s) across time and location. Suggests related popular searches (broad match).	Absolute volume of search term across time and location. Suggests related popular searches (exact match).	Find queries that have a similar pattern across time or state. Data can be of an entered query or a variable that is uploaded from a dataset.
Time period	Ten years	Two years	Ten years
Frequency	Weekly	Monthly	Weekly
Search type	Broad	Exact	Broad
Search units	Relative to peak value	Absolute number	Normalized to zero
Geographic filters	Country / state / metro / city	Up to ten locations, as precise as zip code level	Country / state
Side-by-Side Comparison	Up to five queries (any search term)	n/a	Up to two queries (from 100 most-correlated terms)
Negative keywords	X	X	
Additional filters	Web, Image, YouTube, News, and Shopping	Computer, Mobile, and Tablet	n/a

Figure 1. Trends in "Army" and "Navy" Queries from January 2004–March 2015



SOURCE: RAND Arroyo Center analysis based on Google Trends data (www.google.com/trends).
NOTE: Data from U.S. searches, January 2004 to March 2015.[5]
[a] See footnote 5.

Figure 2. Trends in "Army" and "Navy" Queries Over Time, Excluding Some Irrelevant Data



SOURCE: RAND Arroyo Center analysis based on Google Trends data (www.google.com/trends).

Indeed, one could always go further and exclude "army" terms that include words such as "swiss" (knife), "wives" (TV show), and "surplus" (clothing store), or "navy" terms that include words such as "blue" (color), "pea coat" (outerwear), and "pier" (tourist attraction).

Fortunately, Google Trends allows users to avoid unintentionally including terms that are unwarranted. Rather than entering a specific word, it is possible to choose a predefined search term category that is similar to the word of interest. One of the predefined categories is called "United States Army (Armed Force)," which includes all the search queries that Google has determined are about the Army and excludes "Army-related" queries. (In order to explore the broad attitudes of the Army over time and location, this category will be used later in this report.)

An Application to Army Recruiting

Using state-level search data from Google Trends to identify and plot how the volume of searches in each state has changed over time can provide insight regarding how interest in the Army may have changed across time and place. The value of this analysis, over and beyond that of a survey, is that it can be conducted retroactively on any topic. Furthermore, it is quick, inexpensive, and can often get at attitudes that are not typically disclosed on surveys.

By default, Google Trends provides only the aggregated search volume by location, defined at the state, metro, or city level. A search for the "United States Army (Armed Force)" category from 2004-2014, for example, will output only the average search volume of a location during that entire time period.[6] However, conducting multiple searches over different time periods and then combining the results can produce a time-series data set of search terms over time and across place.

We used Google Trends to conduct 11 separate searches for the predefined category "United States Army (Armed Force)," over single years beginning in 2004 and ending in 2014.[7] The data were appended into a single data set that tracked how the overall popularity of the searched category varied by state and by year.[8]

Figure 3 displays the first and last years of this data set. The difference in the shading of states represents the difference in relative search volume of the category across states in a given year. The darker overall shading in the top figure indicates that there were generally more Army-related searches in 2004 than in 2014. Indeed, the relative search volume of the category in 2014 is almost half the volume in 2004. The relative search volume across states remains generally stable across time. Hawaii and Alaska consistently rank at the top in search volume, as do Alabama, Kentucky, Kansas, District of Columbia, Virginia, North Carolina, South Carolina, and Georgia. The biggest decreases in search volume come from the District of Columbia and Virginia, whereas there are increases in search volume in several states, such as North and South Dakota, as well as Vermont and Delaware. These results suggest that "United States Army (Armed Force)" searches may be carried out by personnel currently on active duty; Virginia, North Carolina, Georgia, and Kentucky are among the states with the highest numbers of active-duty personnel, while Hawaii, Alaska, the District of Columbia, and North Dakota are among the states with the highest ratios of active-duty personnel to total state population.[9]

One strength of this type of analysis is that it is straightforward to replicate with different search terms. For instance, it is possible to track how the searches for Army benefits have changed as a result of the Great Recession or how searches for Army risks have changed as a result of the recent wars. In this manner, it is possible to quickly "take the pulse" of the country regarding any topic of interest and for a specific location.

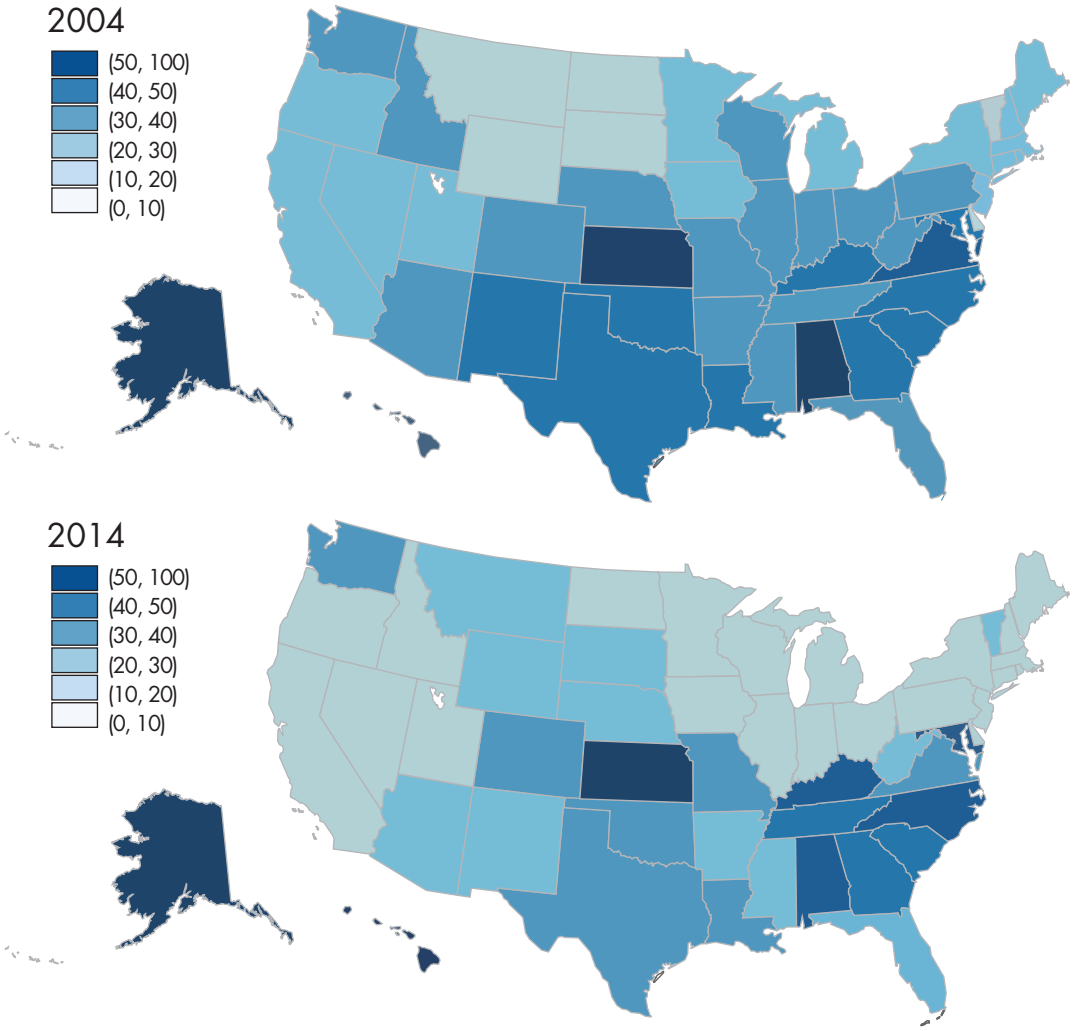
Google AdWords, Keyword Planner

The Keyword Planner tool in Google AdWords is intended to be used by online advertisers to determine which search terms to "bid on" in order to purchase ads targeted to those terms, and thereby to the people who conduct searches for those terms. However, use of this tool is not restricted to advertisers, although it is necessary to sign up for a Google AdWords account prior to accessing any search volume data.

Once a term is entered into the Google AdWords Keyword Planner, the tool outputs both the average monthly search volume (in actual total searches) of that exact term for up to two years, as well as information on similar alternate search terms. This lengthy list of suggested alternate search terms is a key feature of Google AdWords, and something that might be of interest to the Army.[10] This tool is well-suited to explore our second research question: "What sorts of questions and concerns are prevalent in Army-related searches?"

We used this tool to produce a list of questions for which potential recruits might be searching, by including words that are likely to be part of such questions in an initial search. We began with the phrase "can army," which was searched only about ten times per month.[11] This is definitely a small search, especially given that the Army signs up approximately 5,000

Figure 3. Frequency of "United States Army (Armed Force)" Searches by State in 2004 and 2014



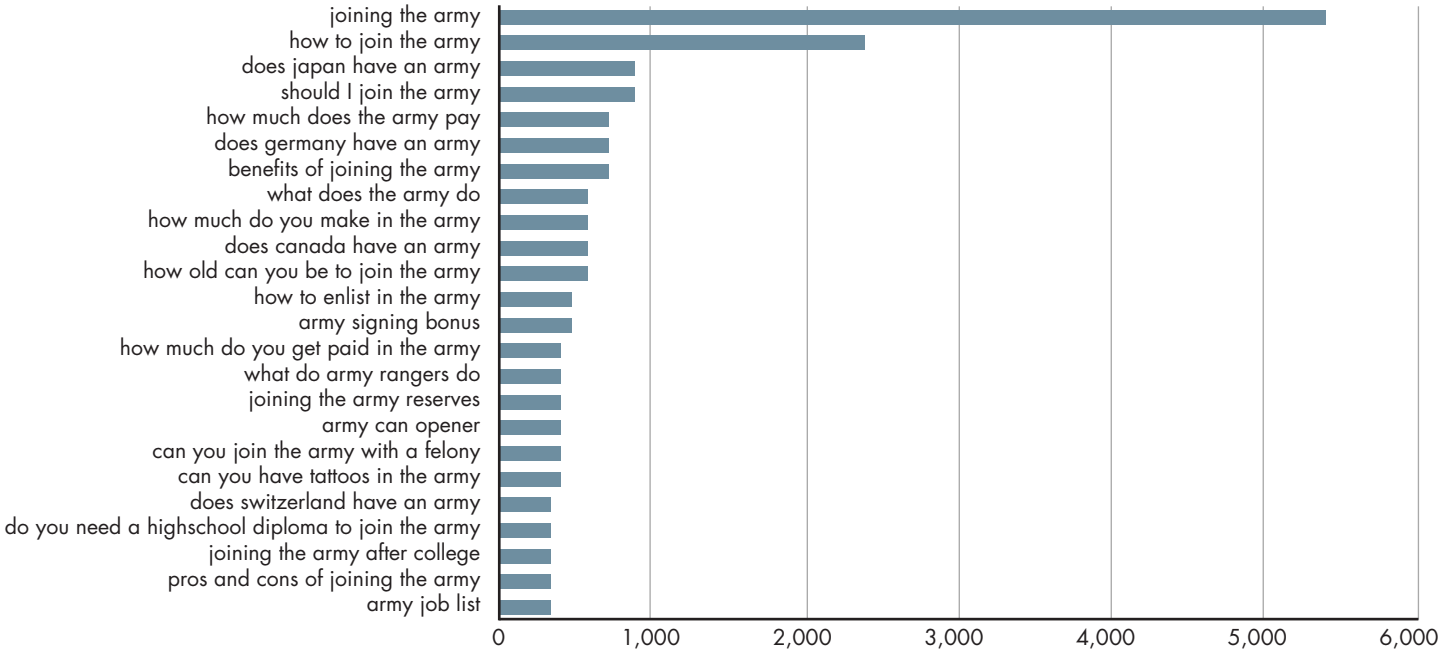
SOURCE: RAND Arroyo Center analysis based on Google Trends data (www.google.com/trends).

recruits per month. The tool, however, revealed many more-common searches. For example, the keyword suggestions show that "how old can you be to join the army" is the most common search term to include the words "can" and "army," and is searched about 480 times per month. Some other popular searches include inquiries about age requirements, felony convictions, tattoos, flat feet, asthma, DUIs, and immigrant status. In some cases, the search patterns for keywords differed across time. This might denote seasonal patterns in the concerns people have about topics. For example, the phrase "how old can you be to join the army" seems to exhibit cyclical variation, whereas the general inquiry of "how can I join the army" seems to have a relatively constant search volume across time. Such findings could be informative about when and for what reasons people are searching about the Army.[12]

Note that the results described above are for a specific type of concern that involves the words "can" and "army." To examine a larger list of concerns, Google AdWords can be used to conduct multiple searches, each using a different question. The result of this technique is a detailed list of frequently searched questions related to the Army. On the list, shown in Figure 4, the most frequently searched relevant queries are "joining the army" and "how to join the army," followed by "should I join the army" and "how much does the army pay."

One characteristic of this list is that some of the queries are clearly not relevant to our area of interest (e.g., "does japan

Figure 4. Average Number of Monthly Searches Related to Joining the Army



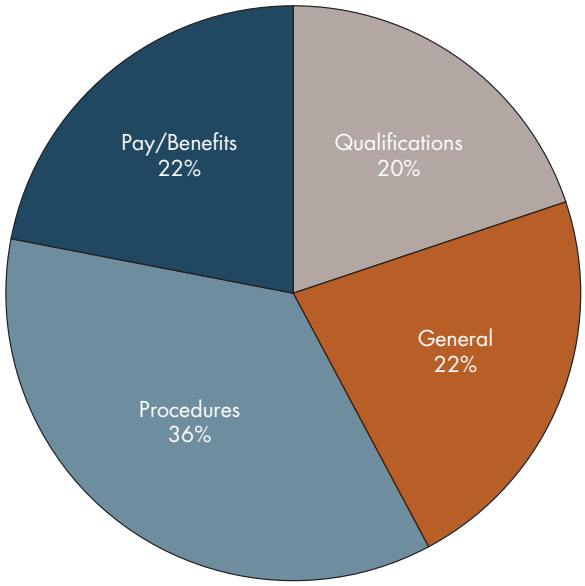
SOURCE: RAND Arroyo Center analysis of Google AdWords data.
NOTE: U.S. searches from March 2013 through February 2015.

have an army" or "army can opener"). As was the case with the first example, caution is necessary to determine the relevance of search results.

To better understand the output from this search, we sorted each query on the list into one of the following categories: (1) questions regarding "Qualifications," such as "can I join the Army if…" followed by restrictions about age, drug use, education, health, and miscellaneous (which include inquiries about tattoos, immigration status, and hair/beard requirements); (2) questions about "Pay and Benefits"; (3) procedural questions, such as "how do I join the Army?"; and (4) "General" inquiries about the Army, mostly related to the job.[13] Figure 5 displays the shares of overall searches in each category, with the largest number of searches in the "Procedures" category—that is, basic questions about how to join the Army.

Of note is that many of the "Qualifications" questions involve specific enlistment requirements. Also, a substantial number of these queries cover topics that young people might be hesitant to discuss with recruiters, such as health issues or past violations of the law. To examine the variation in topics and the relative prevalence of key ideas, we summarized all the queries related to the Qualifications category and weighted each

Figure 5. Prevalence of Army-Related Search Terms by Category



SOURCE: RAND Arroyo Center analysis of Google AdWords data.
NOTE: U.S. searches from March 2013 through February 2015.

word based on the number of times it appeared in searches. Figure 6 presents these data as a simple word cloud.[14]

Figure 6 suggests that many Google searches involve issues that may make people ineligible to enlist and that searches are being used to determine likely eligibility.

These results, however, should be taken with caution. First, it is unlikely that searches are being conducted by the average population; rather, they are likely conducted by those people who already have a propensity to join the Army. The questions asked by such people may differ in substance from questions asked by those without such an inclination. A second concern is that the results are highly dependent upon the language of the search; that is, if the researcher does not phrase the search term in a similar fashion to the searcher, many concerns are likely to be overlooked. For example, a search for "post 9/11 GI bill" can overlook the education-related searches conducted under the query "does army pay for college."

Despite these cautions, Google AdWords results are valuable in that they provide an overview of questions and concerns related to enlistment, which potentially helps paint a picture of people's thoughts related to Army enlistment and qualification. Finally, such results may also suggest areas in which the Army should concentrate online advertising efforts; monitoring the responses to the most-common relevant searches (i.e., "join the Army") is another potential strategy for reaching prospective recruits.[15]

We also formed word clouds from the searches categorized as "General," "Pay/Benefits" and "Procedures." The searches related to "General" are dominated by the word "job." The searches related to "Pay/Benefits" are dominated by words that relate to the amount of pay ("pay," "paid," "much," and "get") while the searches related to "Procedures" are dominated by "requirements," "qualifications," and "get," the last referring to "get into the army" and similar searches.

Google Correlate

As the earlier examples demonstrate, the specific search words chosen for analysis have an enormous influence on the results. While Google AdWords is helpful in suggesting related search terms, it is not a very elegant solution for determining the complete set of relevant search terms. Google Correlate helps to eliminate much of the guesswork surrounding the choice of

Figure 6. Word Cloud Based on Searches Related to Army Enlistment Qualifications

SOURCE: RAND Arroyo Center analysis of Google AdWords data; U.S. searches, March 2013 to February 2015.

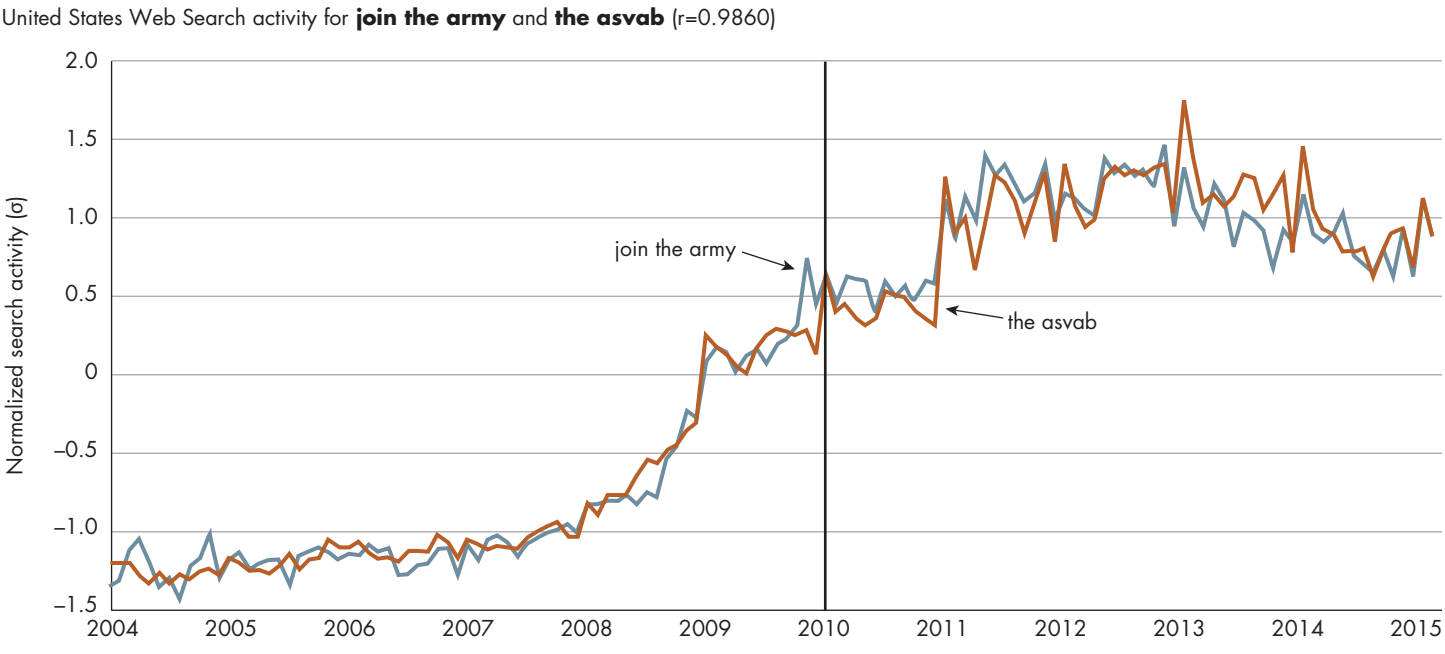
search terms to use. It allows researchers to employ the same methodology as Google Flu Trends, which essentially automated the process of choosing flu-related search terms.

Google Correlate allows users to upload their own data series (a weekly or monthly time series, or data that differs across states), enter a search term, or even use an online pencil to draw a pattern that describes how search volume varies across time. Google Correlate then returns the list of search words that exhibit a pattern of relative search volume that is similar to the data series that was entered. Specifically, the tool provides the top 100 search terms that are most-highly correlated with the term entered along with their correlation coefficients; the same analysis is available for geographic data.

Whereas Google Trends can be used to compare multiple queries across time and location and Google AdWords can be used to suggest searches containing keywords, the strength of Google Correlate is its ability to find queries that match a specific search pattern. Specifically, Correlate searches through its database to identify the queries that have the highest correlation coefficient, *r*, to the entered term. A search for "join the army" will output 100 terms that have a similar search intensity. Figure 7 displays the normalized search volume for the terms "join the army" and "the asvab,"[16] which is almost perfectly correlated ($r = 0.99$). Indeed, the full list of correlated terms sheds light on some of the searches that are likely conducted by the same or similar people. The term "join the army" is correlated highly with terms such as "join the marines" ($r = 0.99$) and "join the military" ($r = 0.99$), which suggest that people who are thinking of joining the Army may also be considering jobs in other branches of the military.

Interestingly, the term "join the army" is also highly correlated to more ambiguous terms such as "the easiest" ($r = 0.99$), "what are some good" ($r = 0.99$), and "get unemployment" ($r = 0.99$). It is impossible to say for certain whether these correlations are non-spurious, but certainly it is possible that they all reference credit constraints.[17] It is categorically important to recognize when using this tool that highly correlated search terms may or may not be searched by the same people. In fact, it is impossible to discern whether the same people who are searching for "join the army" are also searching for "get unemployment" ($r = 0.98$). Nonetheless, these correlations may be useful to personalize recruitment opportunities, as Google Correlate may reveal information about questions and concerns that are prevalent in Army-related searches.

Figure 7. Results From Google Correlate, "Join the Army" vs. "the ASVAB"



SOURCE: Google Correlate screenshot.
NOTE: Results from January 2004 to March 2015.

An Application to Army Recruiting

Perhaps the most interesting way to use Google Correlate is to track the evolution of people's concerns across time. As noted above, Google Correlate does *not* link searches of people across time, but information in aggregate search patterns may be useful nonetheless. For a given search, Google Correlate can be configured to find search terms related to the original word or words searched a set number of weeks earlier or later. To take a non-Army example, a Google Correlate search for the term "weight loss" returns the correlated term "best vacation spots" ($r = 0.89$), but also returns such terms as "why am i not losing weight" ($r = 0.81$), when the correlation is set to find search patterns that occur three weeks in the future. This suggests that a large fraction of people who conduct searches on a given topic will return—consistently and predictably—to inquire about related items.

Therefore, Google Correlate might be used to identify searches that are conducted prior to and just after an individual searches for "join the army." Interpreting the surrounding searches, however, is speculative at best; given the millions of search queries that exist, it is quite possible that two terms exhibit similar patterns of search because of chance. To test this, we used Google Correlate to conduct 13 separate monthly correlations for "join the army," beginning six months prior to the search and ending six months after the search.[18] The data on the 100 most-correlated search terms in each month prior to, during, and after the "join the army" search were collected and appended into a single data set.

The single words that follow the same search volume pattern as "join the army" but occur during the six months prior to this search are relatively more likely to contain terms such as "song" (lyrics and meaning), "email" (how to send email), and "hub" (porn-related). In contrast, the single words that follow the same search volume pattern as "join the army" but occur during the six months after the search are relatively more likely to contain terms such as "jquery,"[19] "document," and "facebook." It is hard to take much meaning away from this list as spurious correlations cannot be ruled out; of course, this is an important limitation of this type of analysis.

Rather than using single words, another method of comparing the query list before and after a "join the army" search is to focus on terms containing more words. If we look at queries that have at least four words in the six-month period prior to the Army search, most searches concerned song lyrics. In the six-month period following the "join the army" search, many searches concerned relationships. These results suggest that many people who are searching for information about the Army are also looking to improve themselves and develop lasting relationships. The Army may benefit from understanding these goals and what they may reveal.

Combining Google Trends Data with Army Accession Data

The information discussed above, particularly trends in searches over time, suggests that data from Google searches may provide a measure of propensity. Such a measure could be valuable in its own right, but here we will show that it could also be used to complement traditional recruiting models. Propensity is acknowledged as a key factor in recruiting, but the information on propensity or on changes in propensity over time is limited. Most research on this topic relies on specialized surveys of the youth population (Woodruff et al., 2006; Orvis et al., 1996); data from these surveys suggest that measured propensity is strongly linked to enlistment.[20] However, the surveys include information on a limited population and occur only occasionally; inevitably, there is a lag between the fielding of the survey and the release of the information. For these reasons, Google search data may provide additional, timelier information for Army recruiting. To explore this issue, we consider our third research question: "How is the number of relevant searches related to the number of people who enlist?"

If Google data might reveal a shift in propensity, it is not immediately obvious how that shift could be measured. Military enlistment involves several steps. Interested youth discuss options with a recruiter, travel to a Military Entrance Processing Station (MEPS) for testing, enlist, generally spend at least some time in the Delayed Enlistment Program, and eventually ship to boot camp. Factors encouraging enlistment include recruiters, enlistment incentives (e.g., bonuses), and advertising. It seems likely that a shift in propensity would show up most quickly among applicants; if propensity to enlist increases, we might expect an increase in the number of young people discussing enlistment with recruiters and traveling to a nearby MEPS for testing. Any increase in accessions might be expected to occur after some delay, if at all. Given the Army's overall enlistment goals, a more likely possibility is that a change in propensity could affect the total number of applicants and therefore could have a positive impact on the overall quality of accessions (because the Army could now choose among more "high-quality" or "highly qualified" candidates to fill its set mission).[21] Again, we might expect that any change to acces-

sions would occur after a lag because enlistment generally takes several months.

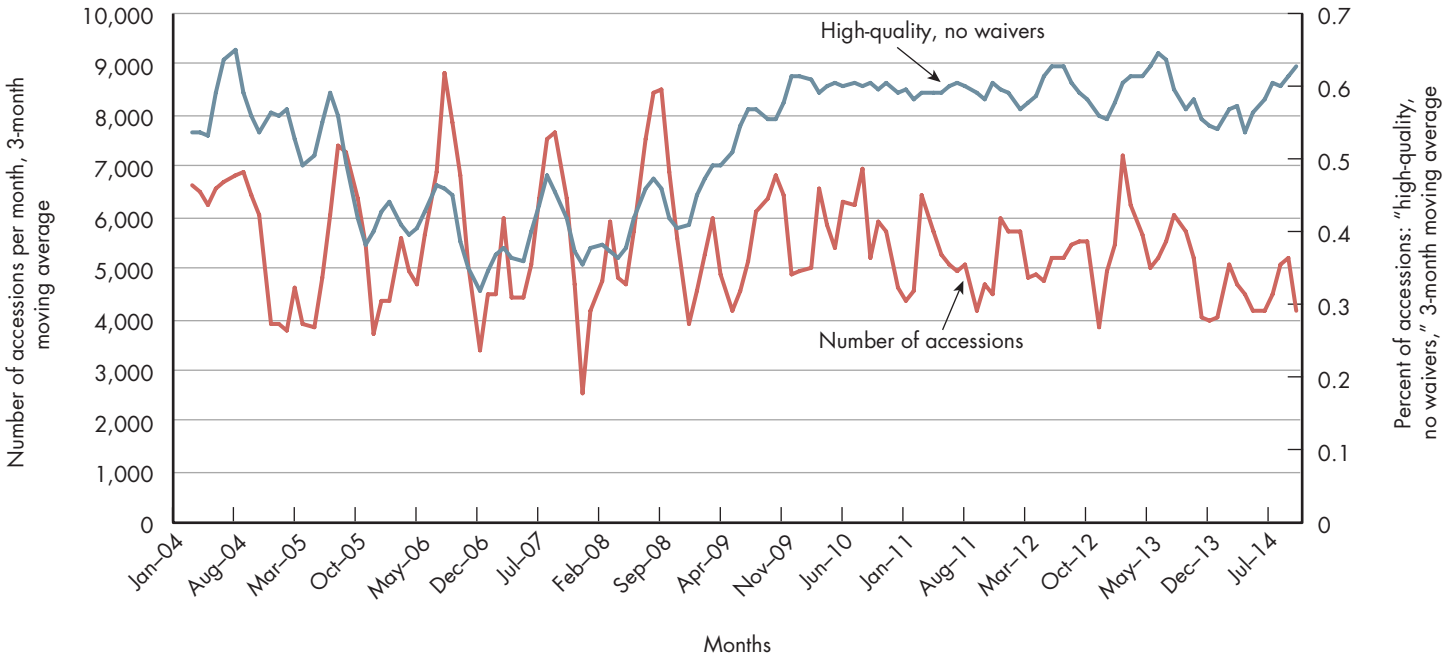
Because of the time constraints for writing this report, we were unable to obtain data on applicants; therefore, we tested our hypothesis on Army accessions data from January 2004 through October 2014 (the most recent accession data available at the writing of this report). During this period, the total number of non-prior-service Army enlisted accessions per fiscal year varied from fewer than 60,000 to nearly 75,000. Our initial file included all accessions in the relevant time frame, as well as information on education, age, Armed Forces Qualifying Test (AFQT) score, home of record, and waivers.[22]

Quality of enlistees is measured in several ways. The traditional measure of "high-quality" recruits is based on education and standardized test scores; those who complete a traditional high school diploma (or any credential considered equivalent) *and* who score in the top half of the AFQT are considered "high-quality."[23] However, the majority of accessions meet this quality threshold; therefore, we also tested several more-stringent measures of high quality.[24] One measure in particular combines the high-quality education and test score requirements with an added requirement that the enlistee have no waivers for alcohol-, drug-, or judicial-related reasons; the proportion of enlistees meeting this measure varied considerably over our time frame.[25] We refer to this measure as "high-quality, no waivers."

For each month, we calculate the total number of accessions, as well as the number meeting our quality measure; from this, we determine the percentage of "high-quality, no waivers" accessions. As shown in Figure 8, there is a considerable variation in the number of accessions per month; to some extent, this is due to the seasonality of recruiting and training. Many new enlisted personnel enter the Army during the summer; in contrast, the number of accessions during the winter months is much lower. However, the total number of accessions per year also trended downward during this period. As accession missions decreased, the proportion of accessions meeting our "high-quality, no waivers" standard increased.

Next, we compare our Google search data with this quality measure. To do so, we choose two different measures: a general term formed from the sum of the number of searches on 31 different Army-related searches suggested by AdWords (which we refer to as "army searches"), and the term "United States Army (Armed Force)," which we refer to as "US Army" for brevity. (Recall that the "US Army" search term is a Google-defined

Figure 8. Changes in Number and Quality of Accessions over Time



SOURCE: RAND Arroyo Center calculations based on Army accession data.

category, including all search terms determined by Google to be relevant to the U.S. Army.) If search data have the potential to provide additional information about the enlistment process, then at a minimum we would expect search intensity (number of relevant searches in a time period) to be correlated with measures of quality. Indeed, we find that search intensity, as measured by "army searches" and "US Army," is correlated with the percentage of enlistees who meet the "high-quality, no waivers" standard. The correlation is positive in the case of "army searches" and negative in the case of "US Army."[26] The difference in signs could suggest that these two search terms capture rather different aspects of Army-related searches; for example, one search term could capture more of the queries about potential disqualifying conditions while the other could capture more queries about jobs or opportunities.

Correlation is a useful measure, but it implies neither causality nor even a predictive relationship between the variables. Granger causality is an additional test to learn more about the likely relationship between search intensity and enlistments (or any two variables that change over time). While still not a definitive test of causality, the Granger test examines the extent to which past values of one time series can be used to predict future values of another. In our case, we are interested in the extent to which searches in some period of time can be used to predict enlistments in some future period. We perform Granger tests of both of our search terms (separately) on "high-quality, no waivers" enlistments. The results suggest that past measures of these search terms contain information that is related to future measures of enlistments.[27] This finding is what we would expect if search intensities measure something about youth propensity.

Next, in an attempt to learn more about the relationship between search intensity measures and enlistment, we include search intensity measures in a very simple regression model measuring the proportion of enlistees meeting our "high-quality, no waivers" definition. Because the number and quality of enlistments tends to vary both across years and across seasons, models also included indicators for calendar year and for quarter of the year. Because our Granger tests suggest that past search information predicts future enlistments, we also include lagged search variables in some specifications.[28] The results suggest that Internet search data has the potential to improve recruit supply models. In Table 2, we report two ordinary least squares regressions. In Model 1 we consider the baseline where no internet search is considered. Model 2 includes the search term for "US Army." We find that a one-standard-deviation increase in "US Army" search intensity is associated with about a 3-percentage-point increase in the proportion of "high-quality, no waivers" recruits and that this is significant at the 10 percent level. The change in the proportion of "high-quality, no waiver" enlistees over the time period was much larger than three percentage points (see Figure 8), but this is a substantive difference. These regression results are consistent with Internet search data providing a measure of youth propensity.

Table 2. Regression Results, High-Quality Enlistments

Variable	Model 1 coefficient/ standard error	Model 2 coefficient/ standard error
Constant	55.75[a] (1.62)	29.80[a] (14.2)
"US Army" searches	~	0.31[b] (0.17)
Year, Quarter Dummy Variables	X	X

NOTE: Dependent variable is percent of "high-quality, no waivers," accessions. Monthly data, January 2004 through October 2014. Models include year, quarter dummy variables. Sample size is 132. Standard errors are in parentheses.
[a] 5-percent level or better.
[b] 10-percent level or better.

We also examine the extent to which adding such search terms could aid in making out-of-sample predictions. The results were promising, although more-detailed data would be required to carry out a more stringent test.

We stress that these results are suggestive at best. The model we present is very simple and incomplete; a more-complete recruit supply model would include additional information about the civilian economy, recruiting resources, and the population. And we note that our measure of "US Army" searches is only marginally significant while our alternate measure of "army searches" did not achieve significance in a similar equation. Also, testing such a model on applicant data seems a more precise method of measuring the influence of search behavior. However, our results suggest that trends in search behavior may have the potential to improve the quality of Army recruiting models.

In summary, and despite the somewhat opaque nature of these measures, our results suggest that Internet search data have the potential to reveal information about the concerns and interests of youth. Our results also suggest that search data may provide an alternate measure of propensity, and thus may be a useful addition to existing recruit supply models.

IMPLICATIONS AND FUTURE WORK

In the past few years, data collection has grown tremendously. This is due, in no small part, to the growth of the Internet and the detailed record-keeping associated with posting information online. Indeed, many businesses, especially in the tech industry, have begun to appreciate the potential of big data to improve business analytics. Tools are being developed regularly to access and analyze data in real time, with the goal of generating insights and improving the predictive power of models that are used to expand products and services. In this section, we discuss how three readily available online tools (used to track Internet search patterns) can be used to improve Army marketing and recruiting efforts.

We summarize our key findings about the usefulness of these tools as follows:

- At the macro level, our findings indicate that Google search queries can be used to better understand how interest in military careers has evolved over time and geographic location
- At a micro level, it is possible to use these tools to identify the chief Army-related concerns that potential recruits experience, whether with regard to the qualifications, procedures, or benefits of enlisting
- At a deeper micro level, our findings suggest that it might even be possible to predict with reasonable accuracy what people were searching for months before or after searching for the terms "the army"
- Finally, we find that including Google Trends terms in a simple model of Army accessions increases the predictive power of the model.

Perhaps the most instructive observations from this study pertain to things that we did not find. For instance, there were very few searches related to the personal harms or other negative aspects that can arise from joining the service. In general, searchers did not inquire about the probability of dying or getting injured, or the likelihood of being deployed in a war zone. In the same vein, people did not search for retirement plans or details about the medical coverage offered by the Army. This suggests that the Internet is a source of information for those who are already likely to enlist; those who are apprehensive about the negative facets of the Army are not searching for how to join.

Indeed, one of the biggest limitations of using Google search data is that our sample of searchers is not random. This selected sample curbs our ability to draw general inferences. Our study necessarily excludes people without Internet access, people who do not conduct Internet searches for the purposes described, or people who do not use Google as their search engine. Our conclusions, therefore, cannot be generalized to the entire population, but are conditional on the population of searchers. Similarly, it should be stressed that the interpretation of the results depends critically on the extent to which the search terms used are appropriate measures to answer the questions posed. Though we tried our best to choose broad queries that are robust to contextual framing, it is certainly possible that a different set of queries will change the results. The development of a systematic method for choosing search queries would certainly be a big contribution to this area. While caution is required before inferring causality, big data may nevertheless help discover correlations that merit further exploration.

The extent to which big data can supplement—or even replace—burdensome, traditional data collection methods is unknown, but initial indications suggest search data may capture aspects of propensity. These data may also be able to answer more-specific questions—for example: What explains changes in propensity and attitudes toward the military? In which types of advertising should the Army consider investing? Information from search data also provides insights into the concerns and areas of interest of one group of potential enlistees.

Further research is required to better understand ways of using Internet search data to inform Army recruiting initiatives, with the eventual goal of drawing causal inferences and making accurate predictions. One promising path involves combining the Google search data with a secondary data source, such as the U.S. Census Bureau, which has information about the percentage of the military-aged population within a geographic region. Insofar as we believe that the majority of searches related to Army enlistment are being conducted by 18- to 35-year-old men, then one can always reweight the search terms appropriately to have a more precise comparison of search intensity across time and location. Indeed, the inclusion of additional data sources can help in other ways as well. Data on the number of deployments to Iraq or unemployment rates during the Great Recession can be informative of the attitudes of people toward the Army. The inclusion of secondary data can be helpful in better determining the factors that underlie the results.

Another promising path is to closely mirror models used by such companies as Amazon and Netflix to offer choices (e.g., to recommend additional movies or books) to people based on the behavior of similar shoppers. One potential way to achieve this is to use accession data in combination with Internet search

> Data on the number of deployments to Iraq or unemployment rates during the Great Recession can be informative of the attitudes of people toward the Army.

data to identify the queries that are most often searched by each demographic group. For example, our data might show that 18- to 20-year-olds who joined the Army from Michigan often search for college loans. Insofar as 18- to 20-year-old civilians in Michigan are generally interested in learning about college loans, then offering this information to them can be helpful. In fact, a randomized control trial can be used to send some people the standard Army recruitment packet while sending others a specialized packet that includes relevant information about the college loan program. By comparing the number of Army applications that result from each packet, it is possible to test whether providing people with information that an analyst has determined is valuable based on knowledge of the individual's demographic leads to a difference in their choices.

Big data is still a relatively new resource. It has already yielded some important insights, and more research is required to realize its potential. In going forward, it would be crucial to move beyond a proof-of-concept and examine thoroughly how behavioral models can utilize big data to make causal inferences and improve prediction models. The potential impact is vast.

TIPS, TRICKS, AND OTHER CONSIDERATIONS

We formed our data set on enlisted Army accessions from the RA Analyst file, which included information on every service member who enlists in the Army. We included all non-prior-service active-duty enlisted accessions. While this analysis could easily be expanded to include officers and/or personnel enlisting in the Reserve Component, we focused on those enlisting for the first time in active duty. We note that search terms are likely to differ between enlisted personnel and officers; search terms may also differ between those interested in serving in the active component versus those interested in serving in the Reserve. These differences constitute additional areas for future research.

In this section, we include some additional information about our accessions data set, as well as a list of "lessons learned" that may be helpful for others exploring big data and Google analytics.

1. **Robustness of Search Results.** The volume of data may depend on the specific phrasing of a search query. For example, "can I join" will return more search results than "can you join." It is important to check variants of chosen search terms and phrases to ensure that the results are relatively stable. Ideally, the search results would be accompanied by standard errors, allowing analysts to test whether trends significantly increase or decrease over time or location.

2. **Language Can Matter.** The search volume for the term "school," for example, does not account for searches in other languages. Analysts interested in what the Spanish-speaking population is searching for may therefore want to add "escuela" as a search term. Such considerations are especially important when performing cross-country comparisons.

3. **Add or Remove Terms from Search Results.** A sizeable volume of searches for "navy," for example, is for the clothing store Old Navy. It is important to frequently check whether the search is returning the desired results. Examining the list of similar queries on Google Trends, or suggested ideas for Keywords on Google AdWords can achieve this goal. As discussed earlier in the report, it is possible to conduct searches that specifically include or exclude certain words.

4. **Identify Who is Searching.** It may be possible to infer who is conducting the search by using such terms as "my girlfriend" or "my husband." For instance, searches such as "is my son" are likely being conducted by parents, while searches for "can pregnant women" are likely conducted by moms-to-be or their partners.

5. **Think Like the Searcher.** Phrasing search queries in similar ways as the target population should help return more valid results. It may be unlikely, for example, that potential recruits search for "post-9/11 GI bill" when considering the education benefits of the military. A more appropriate search might be something like "does army pay for college."

6. **People Who Do Not Search.** Surprisingly few searches are conducted for even the most common terms. It may be worth considering the ways in which those people who might perform a given search may differ from the sample of people who would not perform that search. For example, it could be the case that those most interested in joining the Army ask their friends to answer their questions and also use Google to search for information. It may also be possible that people interested in joining the Army do *not* search for such information. Therefore, the sample of people conducting searches does not necessarily represent the population, or the population of interest.

Notes

[1] At the time, data for frequently searched terms were being published by an agency in a monthly report listing the 500 most searched words every day on the Internet's largest search engines.

[2] See, for example, Centers for Disease Control and Prevention, "Weekly U.S. Influenza Surveillance Report." As of January 4, 2016: http://www.cdc.gov/flu/weekly/

[3] See, for example, Choi and Varian (2012); Askitas and Zimmerman (2009); Baker and Fradkin (2011); Vosen and Schmidt (2011); Goel et al. (2010); and Guzman (2011).

[4] It is important to note that a decreasing trend does *not* necessarily mean that fewer absolute searches were conducted. The output generated by Google Trends is relative to all searched queries; if people are conducting more searches for other items than they are for "army," then the graph will slope downward despite the fact that the number of "army" searches increased.

[5] The trend lines are not exact; rather, they are furnished using a sampling method and can vary by a few percentage points from day to day. Without knowing the standard errors on the estimates, it is impossible to rule out whether "navy" and "army" had similar search volumes. Google instituted a change in 2011 to provide better geolocation data for search queries. Its explanation reads: "an improvement to our geographical assignment was applied retroactively from 1/1/2011." Given the new definition of locations, certain queries can have minor discontinuities at that point in the trend line. If the box marked "News headlines" is checked after a search, the graph will also include lettered points where relevant terms appeared prominently in news headlines, such as the Fort Hood shootings (November 5, 2009). The box marked "forecasts" will provide a simple prediction of future search volume based on past search history.

[6] The state, metro, and city with the highest relative search volume for the U.S. Army category are Alaska, Watertown, N.Y., and Fort Knox, Ky., respectively.

[7] It is not necessary to use a one-year time period between frames; it is possible to go as low as the weekly level. With shorter time periods, however, it is more likely that any changes are due to random fluctuations rather than structural changes.

[8] It should be noted that the data across years are not on the same scale. Google Trends assigns a value of 100 to the data point that is the highest in any given search period. To produce the results in Figure 3, we weighted each period according to its relative popularity.

[9] See, for example, U.S. Department of Defense (2012).

[10] AdWords does not offer the same ease of outputting results by geographic area as Trends, though with some effort it is possible to obtain these data as well.

[11] These are not unique searches. Indeed, a single person who searches the same term many times per month can influence the outcome of the results.

[12] By restricting the data to specific states, cities, or zip codes, it is also possible to use state-level search data from Google AdWords to identify the main concerns for each state. For example, searching "can army" for each state produces a list of top concerns containing those words in each state. The top concern for the East and West coasts appears to be "What age can I join the army," whereas the top concern for the Midwest is "Can you have sex in the army." Interestingly, concerns about the ability of illegal immigrants to enlist ranks at the top in multiple states (Arizona, Delaware, Michigan, New Mexico, and Vermont). Other concerns are about criminal records (West Virginia, Nebraska), tattoos (Florida), asthma (Rhode Island), flat feet (South Dakota), herpes (North Dakota), bad credit (Maine), and whether women can join (Wyoming).

[13] We excluded unrelated searches such as "army can opener;" such searches comprised about 20 percent of the total number of searches. Excluding these alters the total search volume.

[14] We first exclude the words "join" and "army." If we do not do so, these words dominate and the other information is more difficult to discern. We used the tool made available at wordle.net to produce these images; this non-commercial site allows users to reproduce the images without restrictions.

[15] The Army has tools in place to determine the most common searches. Army Marketing Research Group, private conversation, undated.

[16] The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple-choice test given to potential recruits to determine qualification for enlistment in the United States Armed Forces.

[17] Some of the most popular searches containing "what are some good" refer to jobs, while some of the most popular searches containing the term "the easiest" are "what is the easiest credit card to get" and "the easiest way to make money." Certainly, however, the terms may well be referencing the military, such as "what is the easiest branch of the military."

[18] It is also possible to search for terms that were searched at smaller intervals (weeks). With smaller time frames, however, it is more likely that any changes are due to random fluctuations rather than structural changes.

[19] As it turns out, searches involving the term jquery (a Javascript library application), such as "jquery document" and "jquery ready," are quite common in the post sample. For the purposes of analysis, "jquery" was deemed wholly unrelated to the Army-related searches and thus excluded from the text-based word comparison.

[20] Woodruff et al. (2006) surveyed soldiers in two infantry battalions, with results indicating that there were substantial numbers of low-propensity soldiers in combat arms occupations (their enlistment decision was influenced by other factors). Orvis et al. (1996) found that individual propensity predicts enlistment, but that the overall propensity changed over time; this research also forms part of a larger body of work that establishes the relationships between economic factors, recruiting resources, and the number of high-quality recruits.

[21] Of course, this assumes that the increase in propensity results in at least some increase in the number of highly qualified people. This assumption could be tested using applicant data.

[22] About 1.5 percent of our sample had a non-U.S. address; in nearly all cases, the address listed was in a U.S. territory. Google data produced by the tools discussed include information on all Internet searches from within the United States utilizing the Google platform. While it is possible to obtain information on searches carried out in other countries, searches from U.S. territories are not included in Google U.S. data. Therefore, we excluded records on enlistees with addresses in U.S. territories. After these exclusions, roughly 710,000 observations on non-prior-service enlistees from January 2004 through October 2014 remain; these form our sample.

[23] This definition of *high-quality* is used extensively in the military manpower and recruiting literature; see, for example, Orvis et al., 1996.

[24] The other quality measures we tested included higher AFQT requirements, more-stringent education requirements, requirements for no waivers of any type, and age requirements. We found that the quality measures tended to move in a concerted manner, especially before 2009.

[25] Waivers may be granted for a large number of reasons, including those related to weight, height, body fat, other health issues, age, educational attainment, alcohol- or drug-related offenses, and other judicial offenses.

[26] The correlation between the percentage meeting "high-quality, no waivers" and "army searches" is 0.5388 ($p < 0.0001$); the correlation between the percentage meeting "high-quality, no waivers" and "US Army" is -0.302 ($p < 0.0001$). In each case, such results would occur by chance less than one time in 10,000. Both search terms are also correlated with the number of accessions per month.

[27] We utilize chi-squared tests to determine the relevance between past search intensity and enlistment quality. In each case, the chi-squared test rejected the null (of no relevant information) at a level suggesting such a relationship would occur by chance no more than three times in 100. The Granger causality test also suggested that, as we would expect, future values of enlistments are *not* predictive of past search behavior. If we found that future values of enlistment were predictive of past search behavior, this would suggest correlation but a lack of causality. See, for example, Dimpfl and Jank, 2012.

[28] This model represents a very simple "recruit supply" model. A more complete model is beyond the scope of this effort but it would also include measures of the civilian economy, recruiting resources expended on bonuses and advertising, and various population measures. When we include one-, two-, three-, and six-month lagged measures of "US Army" searches, the measures as a group are statistically significant ($F(4,106)=2.66$). This indicates that such a relationship would occur by chance no more than one time in 20.

References

Askitas, Nikos, and Klaus F. Zimmermann, "Google Econometrics and Unemployment Forecasting," *German Council for Social and Economic Data (RatSWD) Research Notes No. 41*, June 1, 2009.

Baker, Scott, and Andrey Fradkin, *What Drives Job Search? Evidence from Google Search Data*, Discussion Paper No. 10-020, Stanford Institute for Economic Policy Research, March 30, 2011.

Centers for Disease Control and Prevention, "Weekly U.S. Influenza Surveillance Report." As of January 4, 2016: http://www.cdc.gov/flu/weekly/

Choi, Hyunyoung, and Hal Varian, "Predicting the Present with Google Trends," *Economic Record*, Vol. 88, Issue Supplement No. s1, June 2012, pp. 2–9.

Cooper, Crystale Purvis, Kenneth P. Mallon, Steven Leadbetter, Lori A. Pollack, and Lucy A. Peipins, "Cancer Internet Search Activity on a Major Search Engine, United States 2001-2003." *Journal of Medical Internet Research*, Vol. 7, No. 3, July 1, 2005, p. e36.

Dimpfl, Thomas, and Stephan Jank, "Can Internet Search Queries Help To Predict Stock Market Volatility?" *December 2012 Finance Meeting EUROFIDAI-AFFI Paper*, Paris: European Financial Management, June 6, 2012.

Ettredge, Michael, John Gerdes, and Gilbert Karuga, "Using Web-Based Search Data to Predict Macroeconomic Statistics." *Communications of the ACM*, Vol. 48, No. 11, November 2005, pp. 87–92.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature*, Vol. 457, No. 7232, February 19, 2009, pp. 1012–1014.

Goel, Sharad, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts, "Predicting Consumer Behavior with Web Search," *Proceedings of the National Academy of Sciences*, Vol. 107, No. 41, October 12, 2010, pp. 17486–17490.

Guzman, Giselle, "Internet Search Behavior as an Economic Forecasting Tool: The Case of Inflation Expectations," *Journal of Economic and Social Measurement*, Vol. 36, No. 3, 2011, pp. 119–167.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani, "Big Data. The Parable of Google Flu: Traps in Big Data Analysis," *Science*, Vol. 343, No. 6176, March 14, 2014, pp. 1203–1205.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H. Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, May 2011.

Olson, Donald R., Kevin J. Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen, "Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales," *PLoS Computational Biology*, Vol. 9, No. 10, October 17, 2013.

Orvis, Bruce R., Narayan Sastry, and Laurie L. McDonald, "Military Recruiting Outlook: Recent Trends in Enlistment Propensity and Conversion of Potential Enlisted Supply," Santa Monica, Calif.: RAND Corporation, MR-677-A/OSD, 1996. As of December 24, 2015: http://www.rand.org/pubs/monograph_reports/MR677.html

Pew Research Center, *How Americans Go Online*, infographic, September 25, 2013. As of March 30, 2015: http://www.pewinternet.org/2013/09/25/how-americans-go-online/

Purcell, Kristen, Joanna Brenner, and Lee Rainie, *Search Engine Use 2012*, Pew Internet and American Life Project, Pew Research Center, March 9, 2012.

Schindler, Helen R., Jonathan Cave, Neil Robinson, Veronika Horvath, Petal Hackett, Salil Gunashekar, Maarten Botterman, Simon Forge, and Hans Graux, "Europe's Policy Options for a Dynamic and Trustworthy Development of the Internet of Things: SMART 2012/0053," Santa Monica, Calif.: RAND Corporation, RR-356-EC, 2013. As of December 24, 2015: http://www.rand.org/pubs/research_reports/RR356.html

Stephens-Davidowitz, Seth, "How Googling Unmasks Child Abuse," *The New York Times*, July 13, 2013. As of January 14, 2016: http://www.nytimes.com/2013/07/14/opinion/sunday/how-googling-unmasks-child-abuse.html

———, "Google, Tell Me. Is My Son a Genius?" *The New York Times*, January 18, 2014a. As of January 14, 2016: http://www.nytimes.com/2014/01/19/opinion/sunday/google-tell-me-is-my-son-a-genius.html

———, "What Do Pregnant Women Want?" *The New York Times*, May 17, 2014b. As of January 14, 2016: http://www.nytimes.com/2014/05/18/opinion/sunday/what-do-pregnant-women-want.html

U.S. Department of Defense, Office of the Deputy Assistant Secretary of Defense, "2012 Demographics: Profile of the Military Community," *Military Community and Family Policy*, 2012.

Vosen, Simeon, and Torsten Schmidt, "Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends," *Journal of Forecasting*, Vol. 30, No. 6, September 2011, pp. 565–578.

Woodruff, Todd, Ryan Kelty, and David R. Segal, "Propensity to Serve and Motivation to Enlist Among American Combat Soldiers," *Armed Forces & Society*, Vol. 32, No. 3, April 2006, pp. 353–366.

Yeung, Douglas, and Brian Gifford, "Potential Recruits Seek Information Online for Military Enlistment Decision Making: A Research Note," *Armed Forces & Society*, Vol. 37, No. 3, July 2011, pp. 534–549.

Zickuhr, Kathryn, *Who's Not Online and Why*, Pew Internet and American Life Project, Pew Research Center, September 25, 2013.

About This Report

The work presented in this report was sponsored by the U.S. Army and was conducted within RAND Arroyo Center's Personnel, Training, and Health Program. RAND Arroyo Center, part of the RAND Corporation, is a federally funded research and development center sponsored by the U.S. Army. The Project Unique Identification Code (PUIC) for the project that produced this report is HQD156928.

We would like to thank our RAND colleagues, James Hosek and Angela O'Mahony, as well as two anonymous reviewers, for their thoughtful reviews. We also appreciate our colleague Kristin Leuschner's assistance. We are indebted to Michael Hansen for providing helpful feedback throughout this project.

About the Authors

Salar Jahedi is an associate economist at the RAND Corporation and a professor at the Pardee RAND Graduate School. His research is in the area of behavioral economics, where he studies the role of information in economic decisionmaking. Prior to joining RAND, he was an assistant professor at the University of Arkansas.

Jennie W. Wenger is a senior economist at the RAND Corporation, where she studies secondary education, alternate credentials, and noncognitive skills, as well as military compensation and skill requirements. Before joining RAND, she was a senior research scientist and project manager at CNA.

Douglas Yeung is a social psychologist at the RAND Corporation. His research has examined communication styles, behaviors, and mental health when using technology (e.g., social media, mobile devices). Before coming to RAND, Yeung was a product analyst at Oracle.

Limited Print and Electronic Distribution Rights

For more information on this publication, visit www.rand.org/t/rr1197.

Library of Congress Cataloguing-in-Publication Data is available for this publication.
ISBN: 978-0-8330-9414-8

www.rand.org

ARROYO CENTER

RR-1197-A