Psychological Methods
2016, Vol. 21, No. 4, 475–492

1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000081

A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data From the Internet for Use in Psychological Research

Richard N. Landers
Old Dominion University

Robert C. Brusso
ICF International, Fairfax, Virginia, and Old Dominion University

Katelyn J. Cavanaugh and Andrew B. Collmus
Old Dominion University

The term *big data* encompasses a wide range of approaches of collecting and analyzing data in ways that were not possible before the era of modern personal computing. One approach to big data of great potential to psychologists is web scraping, which involves the automated collection of information from webpages. Although web scraping can create massive big datasets with tens of thousands of variables, it can also be used to create modestly sized, more manageable datasets with tens of variables but hundreds of thousands of cases, well within the skillset of most psychologists to analyze, in a matter of hours. In this article, we demystify web scraping methods as currently used to examine research questions of interest to psychologists. First, we introduce an approach called theory-driven web scraping in which the choice to use web-based big data must follow substantive theory. Second, we introduce *data source theories*, a term used to describe the assumptions a researcher must make about a prospective big data source in order to meaningfully scrape data from it. Critically, researchers must derive specific hypotheses to be tested based upon their data source theory, and if these hypotheses are not empirically supported, plans to use that data source should be changed or eliminated. Third, we provide a case study and sample code in Python demonstrating how web scraping can be conducted to collect big data along with links to a web tutorial designed for psychologists. Fourth, we describe a 4-step process to be followed in web scraping projects. Fifth and finally, we discuss legal, practical and ethical concerns faced when conducting web scraping projects.

Keywords: big data, web scraping, data source theory, Python, tutorial

The term *big data* is an umbrella term used to describe several distinct families of technical approaches for data collection, data analysis, data storage, and data visualization (Landers, Fink, & Collmus, in press). When the term is considered holistically, the differences between these component approaches are muddled, making big data appear more inaccessible and unfamiliar to non-computer scientists than its component parts actually are. Even within any particular category, the distinctions between approaches are sometimes fuzzy to those without expertise in each contained technique. For psychologists, one of the most accessible individual techniques to come out of the big data movement is a data extraction approach called web scraping (or web content mining), which involves the automated identification and coding of information found on webpages (Cooley, Mobasher & Srivastava, 1997). Even when this big data technique is used to create a million-case dataset, many of the traditional analytic techniques used in psychology can be applied, including frequentist and Bayesian approaches. In comparison to other data extraction techniques commonly used in psychology, web scraping combines several advantages and disadvantages of both observational and archival research with its own unique capabilities.

Increasingly, psychological questions are being addressed by web scraping, although not generally by psychologists. As the Internet continues to develop, researchers are recognizing the vast quantity of behavioral data being created on the Internet; however, these raw data have in the past been difficult to convert into a format easily importable into a statistical analysis program like R or SPSS without direct human interpretation. Recent software innovations have made this conversion more accessible, reducing what would have previously been a several-thousand-hour undergraduate research assistant coding task into a couple of hours of automatic computer processing after a few hours of computer programming. Research on psychological topics taking advantage of web scraping so far include the exploration of interests (Liu, 2008), personality (Youyou, Kosinski, & Stillwell, 2015), and learning (Baker & Yacef, 2009).

Similar to both observational and archival data collection approaches, web scraping potentially brings many advantages. First, the data extracted from web scraping are behavioral, which is a

This article was published Online First May 23, 2016.

Richard N. Landers, Department of Psychology, Old Dominion University; Robert C. Brusso, ICF International, Fairfax, Virginia, and Department of Psychology, Old Dominion University; Katelyn J. Cavanaugh and Andrew B. Collmus, Department of Psychology, Old Dominion University.

Correspondence concerning this article should be addressed to Richard N. Landers, Department of Psychology, Old Dominion University, 250 Mills Godwin Building, Norfolk, VA 23529. E-mail: rnlanders@odu.edu

significant advantage given the overwhelming popularity of survey methodology in psychology despite criticism of the common method variance problems sometimes associated with this approach (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Second, methodologists often lament the small sample sizes typical of psychological research, and web scraping enables the collection of datasets with potentially millions of cases and minimal effort. Third, data extraction in web scraping is essentially invisible to those being observed, eliminating risk of researcher contamination. In contrast, observational researchers often must either go to great lengths and expense to disguise themselves or simply accept the sacrifice in validity resulting from researcher intrusion. Fourth, the amount of time to conduct a research study is dramatically reduced; the time from hypothesis drafting to data-in-hand is reduced to mere hours, which increases the speed with which high-quality research can be conducted (Landers & Behrend, 2015). Fifth, there are often great barriers to conducting large-scale psychological research in both newly industrialized and least developed countries. Given the popularity of the Internet throughout most of the world and the availability of free and open source software to conduct web scraping, the promotion of web scraping may help eschew the tendency for psychological research to focus upon participants that are Western, educated, industrialized, rich, and democratic (WEIRD; Henrich, Heine, & Norenzayan, 2010).

In this article, we first describe the origin of web scraping to provide historical context for this technique. Second, we describe the advantages, disadvantages, and cautions associated with web scraping alongside observation and the collection of archival data to situate this approach with similar data collection methods. Third, we introduce a new concept, which we call the "theory of the data source" as a vital step to establishing the meaningfulness of scraped data. Fourth, we provide a case study that we conducted to explore the strengths and weaknesses of web scraping as a data extraction technique. Fifth, we explore our case study and existing literature on web scraping to develop a set of practical guidelines for social scientists wishing to explore psychological questions using datasets scraped from the Internet.

Web Scraping as a Technology

Although web scraping is a novel approach to data extraction in psychology, the technology and processes that enable it have been explored extensively in the field of data science, where researchers apply this information to multiple content domains. As described by Marres and Weltevrede (2013), in journalism, web scraping has been used to assess the significance of international news stories by counting the number of times those stories were mentioned by social media users. In marketing, it has been applied to gather customer insight data via online profiles and message board participation. In policy research, it has been used to measure endorsement or opposition of political endeavors as public perceptions change, in real time. This sort of processing is possible due to research in the area of information extraction, which describes how unstructured or semistructured content can be processed to meet specific information end-goals. For example, consider a researcher who needs to convert multiple webpages of content into a usable dataset. For researchers who lack web scraping expertise, the two available options involve a cost-time tradeoff. Undergraduate research assistants could spend many hours reviewing source material and engaging in data entry. Alternatively, a programmer could be hired to develop an algorithm that would enter this automatically. In situations where many pages of material must be processed, the time required for data entry increases linearly with each additional page of source material whereas the time and monetary costs for programming are one-time expenses.

Automated data extraction from the web is possible because the computer language underlying the display of modern webpages, called Hypertext Markup Language (HTML), is hierarchically structured around the meaning of the text or other content it contains, commonly called the semantic web (Antoniou & Van Harmelen, 2008). In practice, this can be seen in the raw code used to create HTML documents in the form of nested virtual objects. It is within these objects that the content of webpages is defined and the recommendations for its display are provided. Although there is no central authority dictating how content is delivered on the Internet, HTML5 is currently recommended by the World Wide Web Consortium (W3C; 2014), which is generally accepted as the most reliable source of standards and protocols for interpreting web-based content. Embracing such standards helps to ensure a cohesive and consistent experience across websites. Further, the ability to view and interact with content in similar ways across web browsers (e.g., Microsoft Edge, Google Chrome, Apple Safari, Mozilla Firefox) exists, for the most part, because the organizations producing these browsers have agreed to implement the W3C standards in determining how they display webpages. However, the display of content and the meaning of content were not always united in HTML. In earlier versions of HTML, the virtual objects contained within were intended to describe how content was displayed rather than its semantic meaning.

An example of early HTML can be seen in Figure 1. Virtual objects are defined with opening and closing tag pairs (e.g., <h1>and</h1>), and text between the tags are content to be displayed on a web page. Several tags are used hierarchically. For example, several <b> tag pairs, which display text as bolded, are nested within an <i> tag pair, which displays italicized text. Thus, in this example, "Jeff A. Jones" would be both italicized and bolded, whereas the following "and" would only be italicized.

```
<h1>Computing Confidence Intervals for Standard Regression Coefficients </h1> <i> by <b>

Jeff A. Jones</b> and <b> Niels G. Waller </b> </i> <br> Psychological Methods <br>

December 2013 <br> DOI: 10.1037/a0033269
```

Figure 1. Example of early HTML (semistructured).

After a few years, the W3C determined that this model of delivering web content was nonoptimal, because the tags used did not imply meaning. Text could be italicized for any number of unknown reasons: because it was important, because it was a special name, because the web designer liked the look of italics, and so on. In response, eXtensible Markup Language (XML) was developed to more naturally convey meaning of content (Antoniou & Van Harmelen, 2008). To illustrate the shift, the information presented in Figure 1 can also be found in Figure 2, but converted to an XML-compliant format. By comparing the figures, it is evident that XML is more human-readable and can be algorithmically parsed more easily due to the use of semantic tags. Although authors are simply marked as "bold" in Figure 1, they are defined as "authors" in Figure 2. Modern HTML (i.e., HTML5) generally meets the XML standard, but remnants of the old nonsemantic web remain in use even today.

Because it contains better-behaved data, XML-compliant code is more amenable to automated parsing and extraction during web scraping. Thus, XML-compliant HTML is preferable for web scraping when available (e.g., when choosing between two different data sources that provide the same content). This difference does not imply that information cannot be extracted from non-XML compliant HTML. However, manual pattern recognition of semistructured or unstructured content is required when developing web scrapers of this type of content, which increases time and resource costs.

On the modern web, information is structured according to two stakeholders: website providers who design the structure of webpages, and content creators who provide the content that is displayed within that structure. For example, a community discussion board website typically allows content from content creators to be displayed in various ways. To the content creator (i.e., an end-user of the web), the underlying programming structure is generally not visually apparent (Cafarella, Halevy, & Madhavan, 2011). A particular target web page appearance can be programmed with an essentially infinite number of approaches, so the underlying structure is only revealed when a person chooses to observe the source code that created it. The visual appearance of a web page is irrelevant to a web scraper; only the underlying structure determines what content is amenable to web scraping. Building a web scraper thus requires a thorough investigation of the underlying structure of the web page intended to be scraped.

Given the structure of HTML, two distinct types of content can be scraped to assess psychological constructs: user-generated data and metadata. User-generated data is any information that the user created and added to the website, either personally or via a proxy (e.g., by linking a social media account to that website). Metadata is defined as data that describes other data (Cafarella et al., 2011; Marres & Weltevrede, 2013; Moorthy et al., 2015). Returning to the example of the community message board, user-generated data would include user provided demographic information (e.g., gender, age, and race) and message board post content. Metadata might include the geographic location of the user based on Internet Protocol (IP) address, the date and time of posts, the length of their posts, and other contextual information (Loukides, 2010).

Theory-Driven Web Scraping in Psychology

To maximize the predictive power of web scraping, many data scientists use data mining or machine learning, which are inductive and computational approaches to identify variable interrelationships in datasets, to determine what content to scrape, and to extract meaning from this content. We propose an alternative approach which we have termed "theory-driven web scraping" that follows the hypothetico-deductive modeling approach common in contemporary psychology. Specifically, we promote an approach in which researchers consider web scraping with research questions and hypotheses already determined and only then develop a web scraping approach to address those questions if web scraping is indeed the ideal approach for a given research question. In doing so, we hope to direct psychological researchers away from the many of the pitfalls associated with hypothesizing after results are known (Kerr, 1998) while still gaining the benefits associated with web scraping.

The constructs most relevant to psychologists that can be obtained from web scraping are persistent individual differences

```
<article>
        <title>Computing Confidence Intervals for Standard Regression Coefficients</title>
        <author>Jeff A. Jones</author>
        <author>Niels G. Waller</author>
        <journal>Psychological Methods</journal>
        <year>2013</year>
        <doi>DOI: 10.1037/a0033269</doi>
</article>
```

Figure 2. Information from Figure 1 presented in an XML-compliant format (structured).

(e.g., user demographic information required in user profiles), behavioral data (e.g., comments posted on a message board, number of Likes), and behavioral metadata (e.g., amount of time spent on a message board). These data are readily available on most social network sites (e.g., Facebook, LinkedIn, Twitter), message boards, and blogs. For psychological research, web scraping allows researchers to (a) analyze web-based data that previously required a typically unattainable number of research assistants to code and (b) draw inferences and assess relationships between these variables and psychological constructs (Marres & Weltevrede, 2013). However, considerations involving operationalization of constructs, content validity, and construct validity are still critical (see Binning & Barrett, 1989; Cronbach & Meehl, 1955; Shadish, Cook, & Campbell, 2002) and are often overlooked or willfully ignored by many big data practitioners (Lazer, Kennedy, King, & Vespignani, 2014). Although demographic data can be easily parsed and extracted from large quantities of web data, operationalizing metadata (e.g., number of message board responses, number of friends, participation records) and user generated content (e.g., text content posted to a message board, endorsement of statements generated by other users, self-reported individual differences) must be done with consideration of the nomological net of target constructs.

Specifically, most data that can be captured via web scraping are behavioral in nature. Because behavior is by definition the product of a person-situation interaction (Lewin, 1935), the situational context in which those data were created influences their interpretation. For example, when scraping a social network site, a researcher may consider operationalizing the personality trait "religiosity" by counting the number of endorsements of religious content made by each user. However, because religiosity is an individual difference (i.e., from a theoretical perspective, a cause) and religious endorsements are a behavior (i.e., from a theoretical perspective, an effect), other individual differences or situational constraints may have influenced the emergence of those behaviors. Specifically, the behavior of endorsing religious content could be indicative of other individual differences (in addition to religiosity), the availability of religious content for the user to endorse, or both. Depending upon research goals, this distinction could, but does not necessarily, influence the interpretation of study hypothesis tests. In such cases, researchers should seek multiple sources of information indicating the same construct; for example, the latent factor capturing shared variance between endorsement of religious content, the use of words commonly associated with religion, and the specification of a religion on a profile page are likely to be more closely related to religiosity than any one of those behavioral indicators in isolation.

Given the behavioral nature of web scraped data, web scraping's strengths and weaknesses as a data extraction technique are most like those of observational and archival research. Similar to observational research, web scraping has the potential to capture phenomena as they occur in real-time in their natural environment (Marres & Weltevrede, 2013), a key strength of observation (Shadish et al., 2002). Researchers can assess user reactions to phenomena or controversies that are currently occurring in the real world or observe the relationships between available demographic information and online behavior as they occur (Marres, 2015). However, in contrast to observational research, web scraping enables much larger quantities of data to be collected with minimal risk that researchers will influence behavior as it is measured (Copes & Miller, 2015). Similar to archival research, web scraping can extract an abundance of historical web content for the purpose of longitudinal analysis. For example, researchers could scrape the Internet Archive (http://archive.org/web/), which has preserved and catalogued much of the public web for decades (Thelwall & Vaughan, 2004), in order to track how behaviors have changed. Even within a single web page, dates and times are often posted, making the collection of many longitudinal datasets quite simple. A strength of web scraping is that it mitigates a prominent weakness of both observational and archival research: both observational and archival research require time and research assistants to observe or review numerous data points, whereas web scraping extracts this same information quickly and automatically.

Web scraping also shares some weaknesses with both approaches. First, as with observational research, the inferences drawn from web scraping results may be limited by the specific sample that is observed, potentially limiting external validity. Boyd and Crawford (2012) describe how user behaviors on the social network site Twitter are sometimes used by researchers to draw generalizations about all people; however, there are likely important differences between Twitter users and non-Twitter users on many variables. Internet users as a population differ on some individual difference variables from non-Internet users (Wallace, 2001). These differences are threats to external validity when interpreting results from a web scraping dataset if those variables are correlated with constructs of interest (Landers & Behrend, 2015). Additionally, unlike observational research conducted in a single locale, web scraping has the potential to develop datasets drawn beyond WEIRD societies, a common limitation of published psychological research (Henrich et al., 2010). However, it should not be assumed that datasets created from web scraping are automatically generalizable to the worldwide population; many groups are likely to remain out of reach. Web scraping simply brings researchers closer to this goal by making the assumptions made by researchers explicit and testable.

Second, as with traditional observational research, researchers do not always have access to all available data, which may create internal validity threats. Websites or their users may limit the availability of data, allowing researchers to only access portions of content (Boyd & Crawford, 2012). The extent that this limiting is correlated with variables of interest also introduces range restriction (Landers & Behrend, 2015); for example, if researchers are studying privacy and if people with a greater need for privacy are more likely to use strict privacy settings on their profiles, statistics calculated from such profile data will be biased. Other sources of sampling bias may also be introduced by specific website policies; for example, a website asking its users to provide personal identifying information may introduce a sampling bias as a result of some portion of individuals choosing not to join in response to such a policy.

Third, as with both archival and observational research designs, inferences about causality cannot be made with correlational designs. Although all three approaches (i.e., observation, archival, and web scraping) allow for the investigation of trends over time, valid inferences regarding causation are greatly threatened by confounding unmeasured variables (Shadish et al., 2002). However, regarding specific experimental design paradigms, web scraping can be used in both within-subject and between-subjects

research designs, including correlational, experimental, and quasi-experimental approaches that allow researchers to target internal, external, and ecological validity goals. For example, in a correlational design involving a comparison of the number of replies to a message board topic to gender captured from user profiles, the elements necessary to draw causal conclusions are missing, so causal inferences between gender and reply behavior are impossible despite the increased ecological validity due to the use of authentic behavioral data. Leveraging web scraping as part of a design where causal inferences are more appropriate is possible with randomized or quasi-randomized experiments involving scraped data. For example, a philanthropic organization might identify 20 message boards where they believe potential donors likely participate, randomly assign half of those boards to a special promotional effort, and then scrape the content of the board to determine if their organization was discussed to a greater degree in the experimental versus the control condition. Although still quasi-experimental, more trust could be placed in causal inferences regarding the effectiveness of the intervention at increasing organizational interest with reasonable internal validity and a degree of ecological validity not possible with a traditional survey-based approach. In both of these examples, external validity remains a concern as with any other convenience sample; an argument should be made explicitly regarding the similarity of the convenient population to the target population, per the recommendations of Landers and Behrend (2015).

Although web scraping can be used as a stand-alone data extraction technique, even greater information can be obtained by integrating it into a multimethod research design (e.g., by conducting a survey as a follow-up and extension of scraped data), consistent with best practices in psychological research, as well as social science research in general (Mathison, 1988). Multimethod research designs allow for the enrichment of results beyond what a single method can offer (Hanson, Creswell, Clark, Petska, & Creswell, 2005). With both small and large datasets, integrating web scraping with other data extraction techniques (e.g., survey or interview) provides additional perspectives to inform the research questions. As an example, Young, Dutta, and Dommety (2009) could have taken advantage of a multimethod web scraping design by scraping Facebook profile information for their research on relationship status, followed by a targeted survey through Facebook's advertising system to reach those same Facebook users.

Importantly, "theory-driven" is not synonymous with "the use of null hypothesis significant testing [NHST]." NHST relies on the assumption that the null hypothesis is meaningful. In most modern applications of NHST in contemporary psychology, this assumption is untrue (Cohen, 1994; Krueger, 2001); the null hypothesis is almost always false. Despite this, researchers in psychology have often been able to create meaningful, new knowledge about psychology. This is in part due to the typically modest sample sizes common to psychology; with small sample sizes, only moderate or large effect sizes are generally found to be "statistically significant." In the context of web scraping, however, the effects of falsely assuming the null to be meaningful become much greater. Because datasets created as a result of web scraping are commonly very large, with tens of thousands, millions, or even billions of cases, almost every relationship found will be "statistically significant" regardless of effect. Given this, an approach relying primarily upon effect size interpretation should be used to determine consistency with the theory to be tested.[1] Prior literature available on the effect to be examined should be used as a basis for conclusions regarding practical significance (Kirk, 1996). Meta-analytic evidence, in particular, should be prioritized to determine what effect sizes might be expected. If NHST must be used, it should only be employed as a preliminary step to effect size interpretation. In situations where this prior evidence could be summarized as a prior probability distribution, Bayesian analyses should be used as well.

To Test Substantive Theory, a Data Source Theory Must Be Created and Tested

In reviewing the data science literature, we were most discomforted by the various assumptions made by researchers drawing conclusions from scraped data. Because such data are preexisting and then interpreted, researchers approach scraping projects with preconceived notions of why the data they are targeting exist and approximately what shape those data take. For example, Boyd and Crawford (2012) describe how researcher expectations for Twitter and the realities of data scraped from Twitter often differ dramatically. According to Boyd and Crawford's investigation of Twitter's development documentation, a researcher using the resources provided by Twitter to download what they believe to be "all public posts" may in fact be downloading only one percent of posted content, and it is often unclear which content are available.

Although this is an extreme example, it highlights how dramatically different a researcher's beliefs about the reasons their data exist may be from the reality of those data. This is not a problem unique to web scraping; any time researchers collect data or choose a dataset relevant to their research questions, such as in the analysis of archival data, assumptions must be made. In the context of web scraping, however, these need not be untested assumptions. Instead, such beliefs about the dataset can be presented as hypotheses and empirically tested. This is enabled by the availability and examination of metadata. For example, if a researcher believes they are looking at relationships within a well-developed, stable web-based community, this notion can be empirically tested by scraping and analyzing metadata that supports (or refutes) this belief. One approach might involve examining patterns of usage over time and the tenure of website members.

To address this potential incongruity, researchers working with scraped data should first build a data source theory that describes the assumptions they have made about those data. Second, a recursive procedure to refine this data source theory through empirical evidence should be embedded in all web scraping projects. When hypothesis-relevant aspects of the data source theory are challenged and metadata are not available or do not support the initial beliefs of the researcher, the dataset in question should be abandoned for the research questions relying upon those beliefs. In this way, we propose that manuscripts describing web scraping projects be explicit about theoretical and empirical support for the assumptions they have made about their data's existence.

[1] To be clear, we recommend this approach for all research, and not just web scraping studies, but this argument is outside the scope of the current article.

A Web Scraping Case Study From Psychology

Throughout the previous two sections, web scraping has been presented as a technique capable of extracting information from unstructured and semistructured data sources, large and small. Similar to how the emergence of survey methodology enabled access to a previously untapped type of data (i.e., introspective and unobservable), web scraping has the potential to shape the data collection methodology of psychological research (Marres & Weltevrede, 2013) as long as the assumptions underlying it are clearly articulated. Because advances in web scraping are led by computer scientists who often embrace a stance of theory-less brute force empiricism, it may be difficult for psychologists to recognize potential applications within psychology without a concrete example involving the creation of a data source theory. To remedy this, we provide the following case study.

To ensure relevance to a broad audience of psychologists, we selected a replicate-and-extend study of gender differences in the use of self-initiated support-seeking coping strategies related to depressive symptoms as an accessible and illustrative case study. We do not intend to demonstrate the full potential of web scraping by incorporating data mining algorithms, machine learning, or follow-up surveys; instead, we seek to provide a relatable entry point to the basic technical requirements of web scraping alongside the basic analytic challenges faced when conducting such a project. From our overall research question, we sought a website containing a large repository of behavioral data that we believed potentially indicative of the target behavior. After considering the benefits and drawbacks associated with various data sources, including issues such as potential sample size, the availability of the variables of interest (i.e., self-reported gender and social coping behaviors), and the difficulty of web scraping (i.e., minimizing time spent programming as much as possible), we decided upon a depression discussion board (Healthboards.com, n.d.), which at the time appeared to contain approximately 120,000 individual posts across 20,000 conversational threads.

Based upon a review of the first hundred pages of discussion board content, we developed our initial data source theory. First, we believed that virtually all people who created new conversations on the board were engaging in support-seeking behavior, which is also consistent with the broader literature on self-help discussion boards (Evans, Donelle & Hume-Loveland, 2012) and the beneficial effects of communicating with other people via the Internet (Shaw & Gant, 2002). Second, we believed from this review that those joining the site were required to self-report gender, which would minimize any potential biases due to missing data or nonreporting. Third, seeing relatively little replication of usernames, we believed that most users of this discussion board joined it in order to seek support from other members. This was a feature we sought explicitly, because a substantial number of long-term board members would violate nonindependence assumptions in our target dataset. Specifically, by targeting a popular, public board, we assumed that most users seeking support for symptoms of depression would do so and then leave, providing a quasi-random sample of support-seekers.

From a review of the research literature, we developed a replication hypothesis and two research questions to extend the literature. First, a sizable body of research has identified gender differences in depressive symptoms and disorders; specifically, women more often report depressive symptoms than men (Nolen-Hoeksema, 2001; Piccinelli & Wilkinson, 2000). Additionally, both coping style and coping effectiveness have been found to vary by gender (Lengua & Stormshak, 2000), and some evidence is available to suggest that support-seeking coping is more common among women (Piko, 2001). Thus, we first hypothesized that women engage in a greater degree of social support-seeking behavior than men. Next, we developed two research questions to explore the extent to which these gender differences also exist in how people respond to such support-seeking behaviors, which are data traditionally difficult to obtain due to the inaccessibility of those providing social support to research participants for follow-up surveys.

> *Replication Hypothesis:* Women engage in more support-seeking coping behaviors than men.
>
> *Research Question 1:* Are there gender differences in the number of responses to these support-seeking coping behaviors?
>
> *Research Question 2:* Is there an interaction between support-seeking gender and respondent gender such that men respond more often to men, women respond more often to women, women respond more often to men, or men respond more often to women than would be expected from the main effects alone?

Method

To maximize the value to readers, this section contains a step-by-step explanation of the reasoning process we followed in pursuing answers to our research questions. It is not expected that method sections of future web scraping studies in psychology will be equally complex or lengthy. However, it should be recognized that the decision-making required influences the quality of the data obtained. Much like observational, archival, and meta-analytic research, what appear to be minor decisions can sometimes have significant implications. Where we identified such implications, we discuss them in greater depth.

We now define five terms used to describe specific features and technical aspects of discussion boards that will be used throughout this case study. First, "post" refers to any single text comment posted to the website. Second, "original post" refers to a post that is not made in response to another post. Third, a "reply" refers to any post created in response to an original post which appears alongside that original post. A "thread" refers to an original post plus all of its replies. Finally, a "user" is a person with an account on the website, who may contribute original posts and/or replies to discussion. In summary, users create new threads by creating an original post, which may or may not be followed by replies made by either additional users or the original user. All discussion board content follows this general format.

To web scrape the target website, each of the authors first learned the open-source Python programming language by following an award-winning 13-hr free Python course developed by Codecademy (http://www.codecademy.com/). Next, we installed the Scrapy package for Python, which is an open-source set of functions for scraping websites. A full tutorial that we have developed for psychologists learning Python and Scrapy can be

found at http://rlanders.net/scrapytutorial.html. After learning Scrapy (for our team, between 5 and 20 hr per person, which varied by prior programming experience), we as a group created the Python/Scrapy code appearing in the Appendix, which took an additional 4 hr. This code took approximately 20 hr to execute with a 2-s delay between each web page processed in order to avoid overburdening the Healthboards.com server. This effort resulted in the collection of 165,527 posts. We next removed duplicate cases, the creation of which is a common side effect of automated web scraping, by deleting cases flagged using SPSS' *Identify Duplicate Cases* function. This resulted in the elimination of 12,487 duplicates, leaving 153,040 posts. We further removed one impossible post, which was blank and associated with a posting date earlier in time than the first nonblank post. This post was likely due to a programming error by the creators of healthboards.com. After dividing the dataset into original posts and replies, we found that our dataset contained 23,325 distinct threads with 129,714 replies, the first of which was made November 9, 2000, and the last of which was made March 22, 2015, which was the day we ran the scraper. A new variable, which we called *epochDays*, was created to capture the total number of days elapsed since November 9, 2000, resulting in a variable with minimum 0 days and a maximum of 5,246 days ($M = 1{,}739.38$, $SD = 1{,}014.94$), representing 14.37 years of data.

Our data source theory was revised several times throughout the extraction and analysis process. Most critically, our assumption that self-reporting of gender was required when registering for the board was challenged. Simple descriptive statistics revealed that gender was unreported for a substantial portion of the dataset. When we crossed this analysis by time, we realized that gender reporting only became required somewhere around December 2004. This was inferred due to ambiguity in the data. Specifically, 65% of users did not specify a gender in September 2004, 41% in October, 23% in November, and 1% in December. Less than 5% of new users each month did not specify a gender until August 2006, at which point 0, 1, or 2 individual users (less than half a percent each month) did not specify a gender in every subsequent month.

To investigate any potential effects of this change, we next regressed reported gender on epochDays with quadratic, cubic, quartic, quintic, sextic, and septic terms added hierarchically. Together, these variables explained 1.1% of the variance in gender, and the final model contained quadratic, cubic, and quintic terms, all of which were statistically significant. Thus, there were some shifts in gender over time, but they were relatively small in effect size magnitude. However, this interpretation does assume that these shifts were in gender representation of the sample and not gender self-reporting, which could affect the interpretation of our research questions. To reduce this uncertainty (i.e., so that the gender measured can be attributed to gender and not the decision to reveal it in a user profile), we next examined what percentage of posts had an associated gender after the first of each month in 2005 to identify at what point less than 1% of posts were missing gender. This was identified to be November 2005, at which point 99.2% of posts had an identified gender. This became the working dataset, which held 66,387 posts.

Next, a dataset was created containing one case per thread (at this point, $N_{threads} = 11{,}484$) from the working dataset. For each case, the gender of the original poster was included, as well as the proportion of female responses (not including replies by the original poster) to each thread using SPSS' AGGREGATE command. Threads where the original post was before the November 2005 cutoff ($N_{threads} = 90$) were removed. Threads without replies were also removed in order to ensure the two research questions were investigable ($N = 1{,}379$). Finally, original posters without gender ($N_{threads} = 1$) and threads where no replies included reported gender ($N_{threads} = 13$) were removed. This produced the final dataset ($N = 10{,}001$).

The dataset was next checked for the severity of independence assumption violations. Specifically, conversation threads were not independent; users could post more than once (either starting or continuing conversations). Thus, if a significant number of users posted across many conversation threads, gender would be confounded with individual posting behavior. Analysis of this revealed that the median number of posts (either starting or continuing a thread) was 2, and the mode was 1. 91.8% ($N = 8{,}580$) of users posted less than 10 times, suggesting this was not a significant threat to internal validity.

Results

To examine the hypothesis, we first noted that among the 10,001 cases in the final dataset, 70.5% (7,046) were female and 29.5% (2,955) were male. The total number of replies to each thread in this dataset appeared to follow a positively skewed Poisson distribution ($M = 4.13$, median = 3.00, mode = 1) as explored in Figure 3. For comparison, we identified a base rate reported by Kessler, McGonagle, Swartz, Blazer, and Nelson (1993) who found 21.3% of women experience depressive symptoms at some point during their lifetime, whereas 12.7% of men do, a ratio of approximately 1.677:1. The ratio of women to men in this dataset is 2.507:1. Given 10,001 threads, it was expected that if the base rate of people posting on this board (those seeking support) were the same as in the population of those with depressive symptoms identified by Kessler et al., 37.4% would be men and 62.6% would be women.

The percentage of female repliers in our dataset was 69.32% ($SD = 35.32\%$) and slightly negatively skewed (skewness = −0.86). Entering this information into a chi-square goodness-of-fit test in comparison to the base rate produced a significant value, although this would be expected from even small differences with this many cases: $\chi^2(1) = 263.10$, $p < .001$. Women seek support more often than would be expected from the base rate of gender differences in depressive symptoms. Specifically, women engaged in 7.9% more support-seeking behaviors than was expected. Given the relatively large effect size, the hypothesis was supported.

To examine Research Question (RQ) 1, a z test was conducted to compare the gender composition of those responding to support-seeking behaviors (70.27% female) to both the base rate of women engaging in on the board (69.32%) and the base rate of women in the world (50.00%; The World Bank, 2014). This difference was statistically significant in both cases ($z_{board} = 2.06$, $p = .039$; $z_{world} = 40.54$, $p < .001$; 95% CI [69.36%, 71.16%]). A more meaningful interpretation of practical significance revealed two dramatically different effects: 0.95% more women responded than would be expected based upon discussion board membership, but

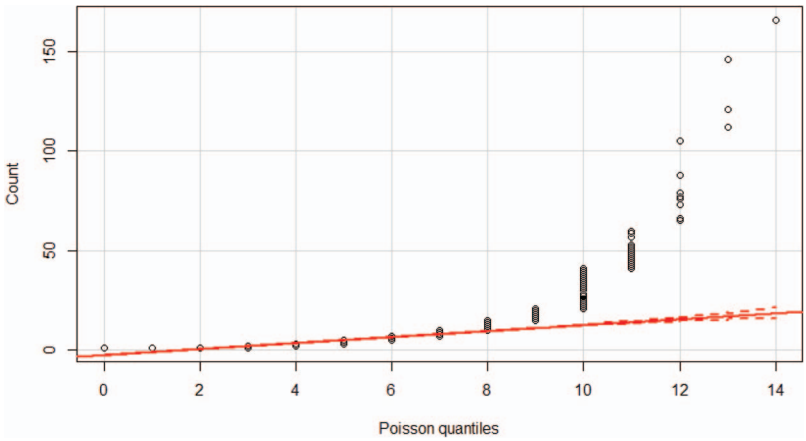


Figure 3. Plot comparing quantiles of total replies versus quantiles of a Poisson distribution. See the online article for the color version of this figure.

20.27% more women responded than would be expected based upon worldwide base rates.

To examine RQ 2, a Mann–Whitney *U* test was conducted to compare the gender balance of people responding to support-seeking behaviors by the original poster's gender due to the nonnormality of the reply gender percentages. This difference was statistically significant ($p < .001$), and a small effect was observed. Specifically, 67% of respondents to male posters were female whereas 70% of respondents to female posters were female, a difference of 3%. The point-biserial correlation between poster gender and respondent gender percentage was .038 ($p < .001$), so we concluded that 0.144% of the variation in respondent gender can be attributed to poster gender.

Discussion

Given these results, we successfully replicated prior literature by finding gender differences in support-seeking coping behaviors, although we did this in the context of an online discussion board, which has not been done previously. We found evidence of gender differences in all tests. The magnitude of this effect in the exploration of RQ 1 varied by comparison group. Specifically, if we assume that because all women in the world have access to this website, the reasonable comparison is 50%, a substantially greater number of women responded than would be expected by chance. However, if we assume that only those initially attracted to the discussion board (i.e., to engage in support-seeking behaviors) are likely to continue using it in order to provide support for others, there is no practical difference. The small effect size found in the exploration of RQ 2 led us to conclude that even if this effect does exist, it is likely too small to be practically useful. Although we cannot make causal conclusions regarding gender due to the study design (i.e., correlational), this is not possible in any study on gender differences. However, without a follow-up survey, we were unable to assess any control variables, which might provide a clearer picture of the drivers of this overall difference. These results thus provide motivation for a variety of follow-up studies.

Lessons Learned and Reflection

The most significant lesson learned from this case study was the value of a data source theory. It was only after running analyses and finding unexpected results that we realized some of our assumptions about the data source had been inaccurate. Thus, we have provided empirical support for the value of creating and testing a data source theory. Any revisions to this data source theory should be clearly documented in manuscripts and publications so that readers can fully evaluate the extent to which such revisions may have capitalized upon chance. Reviewers might even ask for additional tests to verify assumptions that the researchers did not even realize they were making; with a scraped dataset, the effort required to collect additional variables is typically trivial in comparison to that of a large experimental or survey-based study.

We also found it surprising just how little impact the removal of thousands of cases had upon our analyses and effect size estimates. Although we had a well-considered rationale for eliminating cases before November 2005, including those cases in analyses did not change estimates dramatically, or more critically, interpretation. Although the implications of this variation by approach may be important for benchmarking studies, such studies are already inappropriate for most sampling approaches commonly used in psychology (Landers & Behrend, 2015).

Finally, the limited value of relying upon NHST alone was made quite clear. In this case, one of our *z* tests revealed a statistically significant difference with only a 0.95% effect from its test parameter. In larger samples, still smaller differences would be statistically significant. Our findings also provide a reminder of the necessary technical interpretation of confidence intervals (CIs), which is often lost in psychology. In this case, if we assume 70.3% to be the statistical population's true gender percentage, given 95% CI [69.36%, 71.16%], 95% of samples of size $N = 10001$ drawn from this population would find percentages between 69.36% and 71.16%. The confidence interval's lack of overlap with zero or the test parameter implies little about the theoretical population percentage to which researchers typically want to generalize. In this case, the confidence interval provides

no additional information about the gender percentage among depression websites in general, the gender percentage among people experiencing depressive symptoms in general, or any other such population, beyond what is already suggested by the percentage itself. Given this, we reiterate that judgments should instead be based primarily upon effect size interpretation and, if appropriate, a Bayesian evaluation of the posterior probabilities of observed effects.

Guidelines for Web Scraping in Psychology Research Projects

From our case study alongside a careful consideration of extant psychometrics and data science literatures, we have developed a standardized procedure for theory-driven web scraping in psychology. In contrast to this approach, data scientists conceptualize theory as an outcome of data (Cleveland, 2001). In short, psychology researchers typically develop theory deductively from prior research literature and seek to provide support for that theory with data, whereas data scientists collect data and then create or find a theory inductively to explain those data. The particular approach commonly taken among data scientists has been criticized due to its temporal precedence and emphasis on exploratory analysis of extracted data as the principle form of validation (Marres & Weltevrede, 2013). In practice, the large datasets produced by big data techniques are often correlational and taken out of context. Because of the statistical power associated with such large sample sizes, many frequentist analyses, especially those involving NHST, lose value and meaning (Boyd & Crawford, 2012). In response to these criticisms, some data scientists have promoted a priori planning in order to ensure the meaningfulness of these investigations, which echo current practice and recommendations in psychological research. For example, Herman and colleagues (2013) advocate articulating the problem clearly at the beginning and defining a method and analytic plan using knowledge of the subject domain, emphasizing the importance of a driving question with clear and compelling logic driving data extraction and analysis. Loukides (2010) emphasizes that the existence of data is not as important to data science as testing hypotheses using statistics and drawing valid conclusions. Thus, in creating our guidelines, we started with theory-based testable hypotheses and/or research questions to guide each step in the web scraping process as is already common in psychology, but we augmented these standard practices with the perspective of data science. From this process, we developed a set of four steps: identifying the data source, developing a coding system, scraping the web, and cleaning the data collected.

Step 1: Identify Potential Sources of Information

Prior to collecting data, researchers should identify a relevant source of data (i.e., a website or a collection of related websites) and develop a data source theory to explain why those data exist and what type of information they theoretically provide. The purpose of this theory is to summarize researcher understanding of what the data source is, provide context for data contained therein, investigate whether the information contained in the source can be used to test the research questions, and determine how constructs are operationalized in the data source. From the perspective of big data researchers, the sample of a data source provides important framing and context for the relationships found within (Boyd & Crawford, 2012). From the perspective of a psychological researcher, this stage marks the beginning of an empirical investigation of construct validity.

To conduct this investigation, researchers should create specific research questions regarding the data source that would affect interpretation of study hypothesis tests depending upon the answer. Next, the scraper should be programmed so that as many of these research questions can be investigated empirically as possible once data have been collected. Several questions of potential interest appear in Table 1. In addition to these questions, researchers should develop specific hypotheses related to their prospective data source when the answer to a particular research question would influence interpretation of study results.

After the data source has been identified and defined, the data source theory will likely need revision based on unexpected information encountered during initial consideration of the data source research questions, which should be documented clearly. We reviewed discussion board content to ensure the data available reflected support seeking behavior of the content creators. Testing and refining the theory of the data source should be an ongoing process, documented in any resulting manuscripts (Herman et al., 2013), and the theory should be tested more explicitly once data are scraped to ensure that the data collected align with the data source theory.

With an initial data source theory, researchers should next identify technical details of the types of data to be extracted; what are the specific data needed and how are they organized, encoded, and displayed? Raw HTML should be examined to investigate how the website has been created and organized, what kind of data are collected from users, and how those data can be accessed. This can be thought of as mapping the structure of the data (O'Neil & Schutt, 2013), which involves identifying the scope of what to capture and compiling a list of variables and information about each variable. Table 1 provides research questions for researchers to consider for each variable in order to map the structure of the data source.

Verifying that hypotheses regarding the structure align with the structure of the data source has important consequences for all later stages of the project. The data source may contain various types of data (user-generated and metadata; Marres & Weltevrede, 2013), which may be structured or semistructured. Structured data are already categorized into meaningful units by the content provider (Herman et al., 2013) and will simply need to be extracted as-is. Semistructured data contain some organizational markers but have not been formally structured by the content provider, thus requiring additional processing to organize or classify the data meaningfully. As an example, in Figure 1, author names are semistructured whereas in Figure 2, they are structured. The technical details of operationalization should be mapped explicitly before the scraper is programmed so that the scraper is programmed to align with the research question and the available data.

Step 2: Develop a Coding System

Once a theoretical model of the data source has been created, a coding system should be developed, starting with visualization of the final dataset's structure. Programming will be used to create a relational database, in which variables are represented with col-

Table 1
Data Source Theory Research Questions to Be Considered and Empirically Tested If Feasible

Sample research question	Sample hypothesis
Regarding data origin/population characteristics	
Why does this website exist?	The operator of this website is for-profit, earning revenue from advertisements.
Who owns the data available on this website?	Data remain the legal property of the content creators.
Who has access to the website?	All Internet users have unlimited access to this website.
Why would someone want to visit the website?	This website attracts users interested in giving and receiving advice about gardening.
Why would a content creator want to contribute?	Users post on this website because they are seeking creative fulfillment.
What type of data do content creators provide?	Users contribute images but may also comment on those images with text.
Do users pay to participate?	Users must make a $10 yearly payment to post content to this website.
Does the website impose structure on the data provided by content creators or modify the data after it is submitted?	When users submit their user profiles, biographies are limited to 500 text characters and screened for profanity.
Regarding data structure	
How are target constructs represented visually and in code?	Age is shown with every post and is found between and tags.
Is there any inconsistency in how target constructs are represented?	Ages are only available from users that have volunteered that information.
Do data appear on only one type of webpage or are there multiple types?	Lists of interests appear on distinct user profile pages which may have limited access, whereas comments appear on a general form.
How is user content created and captured?	Users can contribute content from any type of computer, including mobile phones.
How much content is available on each page?	The provider allows the display of 20 cases at a time.
Is the target content consistently available on each page?	User profiles cannot be hidden by users.

umns and cases with rows (O'Neil & Schutt, 2013), which is the most common dataset format in Psychology. This dataset must be designed before any coding is done so that the scraper will create precisely this dataset. Essentially, this involves operationalizing the constructs in their original HTML format; for example, gender might be operationalized from semistructured data as "the text content between the and tags within the third tag on each page to be scraped, except when the word 'private' appears within the user profile." Each operationalization must be recorded in one or more datasets, depending upon the data's structure. For example, when examining a discussion board, user profiles might be scraped to create a person-level dataset, whereas their behaviors would create a postlevel dataset nested within the first. In this case, website-provided identifiers must be captured at both levels to later tie the datasets together. Possible levels of analysis vary by the purpose of web scraping, but include visits, people, and posts, among others.

Frequently, there are multiple approaches to scraping a particular website, so researchers should explore which is the ideal method for the purpose of the investigation (Herman et al., 2013). This includes both technical and theoretical concerns. From a technical perspective, webpages may be represented in multiple ways, some of which may be more easily scraped than others. For example, when a web page is displayed in "print view," which is an option offered on many websites, data may be more structured, and thus more accurately scrapable, than from the default view. From a theoretical perspective, much information contained on webpages is unnecessary to extract when theory is guiding decision-making (Marres & Weltevrede, 2013) and should not be scraped. Specifically, data scientists advocate for simplicity in data extraction; extracting only what you need in the most straightforward process possible (O'Neil & Schutt, 2013) is often faster and more accurate (Herman et al., 2013).

Throughout this process, the theory of the data source should be empirically tested to the maximum extent possible. This provides evidence of data veracity, which refers to the accuracy of subjective judgments of how data exist within their context (Herman et al., 2013). The quality of data collected directly affects conclusions drawn later (Loukides, 2010), so veracity is key for subsequent conclusions to be valid. Each time the data source theory is revised, researchers must question if study research questions and hypotheses are still testable. If not, the data source must be abandoned.

Some data scientists at this point incorporate machine learning as a technique to extract meaning. However, this approach is not necessary to answer many psychological research questions and could instead obscure the meaning associated with theory-driven research. In the context of web scraping, machine learning is a process used to iteratively examine patterns and relationships within the raw data algorithmically, typically with little human intervention (Harrington, 2012). This can reveal unexpected relationships but generally does not consider the context of the data (Conway & White, 2012). Without the meaning provided by a theoretical grounding, the risk of erroneous conclusions driven by chance relationships is greatly increased.

Step 3: Code a Scraper and Crawler and Scrape the Web With Them

Once the coding system is finalized, web scraping software must be developed; fortunately, two of the programming languages commonly used for this purpose, R and Python, are well within the skill sets of many psychologists and are already used in psychological research.

Python is a high-level programming language known for fast, simple programming and short, easy-to-read code (Wentworth, Elkner, Downey, & Meyers, 2012). R is an interactive, interpreted, object-oriented language that was designed by statisticians (R Development Core Team, 2015) and is increasingly used for psychological research (Culpepper & Aguinis, 2011). For example, some psychologists have included R code in published articles to encourage replicability (e.g., De Corte, Lievens, & Sackett, 2007).

From a technical perspective, two programs are used for web scraping: a crawler and a scraper. First, a crawler systematically downloads data from the Internet, based on coded parameters, such as a set of addresses, number of pages, or type of files. As each page is crawled, a scraper extracts meaningful information from the raw crawled data, stripped of useless or unrelated pieces, combined with data from other locations, and formatted. Together, these programs automatically extract semistructured data or structured data in a nondesired format from one or more websites and convert that data into a structured dataset according to the coding commands (Marres & Weltevrede, 2013).

Because website data exist for a purpose other than ease of web scraping, creativity and problem solving may be needed to reformat those data appropriately to test research questions. Although the process, based on research questions, is set up before scraping, the approach and/or data may need to be adapted to ensure the data collected are as expected. While creating scraper code, the drafted code should be executed on just one or a few pages to ensure the data collected matches the planned dataset structure from the previous step and code should be revised until correct. Experimentation is often required to arrive at solutions when problems arise in the collection and organization of previously unstructured data (O'Neil & Schutt, 2013). Collecting data this way is recognized as an iterative and lengthy process in data science, as novel problems with complex causes are encountered, and researchers may need to test and rule out multiple solutions to solve them (Herman et al., 2013). Even with dedicated planning effort, researchers should expect to encounter data that do not match expectations, and be prepared to think critically and creatively to solve these problems.

The resources required to complete a web scraping project necessitate planning and consideration; the technical and time resources required are minimal, but sufficient coding expertise in R or Python is most critical. Many of the technical resources needed for web scraping are already widely used or freely available; these include computer hardware, high speed Internet, web scraping software, database management programs (e.g., Microsoft Excel), and statistical analysis packages (e.g., SPSS; Cleveland, 2001). Once the scraper and crawler are coded, running them takes no more effort or time on the part of the researcher, but depending on the size of data source and amount of data collected, hours or days may be needed to collect the data. Coding expertise is necessary, and coding time will vary by skill level. In terms of time investment for psychologists, learning a language such as Python or R is approximately as complicated as learning a new complex statistical procedure, such as structural equation modeling; specific training is needed.

Responsible web scraping practices from the broader data science literature should be incorporated in all projects. A website provider is a person or an organization that has made data available on a website. Scraping that data offers the provider no benefits but collecting datasets by web scraping does introduce processing and cost burdens. Following data science best practices ensures that an unnecessary burden is not placed on the website providing data and that researchers are not blocked by the website (Larson, 2008), which can occur midproject. Within the code of the scraper, researchers should include language that ensures parsimonious data extraction at a reasonable rate. The scraper should only send as many website fetch requests as are required to address hypotheses; collecting extraneous information will take longer for the researcher to collect and will increase the website's bandwidth. For example, if the testing of project hypotheses does not require downloading images, images should not be downloaded. Delays should be incorporated between website fetch requests; the standard guideline is a 2-s delay (Larson, 2008). A steady stream of requests could be interpreted as a denial-of-service attack (i.e., a type of malicious hacking) and could crash the server hosting the website. Outside the code of the scraper, researchers should make sure to minimize the number of times a website is scraped by saving and reusing data whenever possible. Multiple people working on one project should not scrape data separately but should instead work together so that a website is scraped only once. Additionally, scraped datasets should be submitted as supplementary resources of publically available data so future researchers wishing to replicate analyses or explore this dataset to answer other research questions do not need to rescrape the data on this website, as long as participant privacy is not infringed in doing so.

Step 4: Clean the Data and Revise the Data Source Theory

Once the initial dataset has been created, a data cleaning process is necessary to ensure that the structure and form of the data collected match the structure and form intended, consistent with the data source theory. For example, if the data source theory indicates that only American adults (i.e., aged 18 and over) use the website and that age is represented as an integer (e.g., 21, 22, 23), all scraped values should be integers greater than 17. Unlike the smaller datasets that psychologists typically analyze, which usually have fewer than a thousand cases, scrolling through such datasets to scan for impossible data is not useful; with such a large number of cases collected through web scraping, there will likely be far too much data to look through all data points in order to conduct any meaningful review. Instead, cleaning a scraped dataset involves calculating descriptive statistics (e.g., minimum and maximum values, mean, median, skew and kurtosis estimates), producing data visualizations, and creating frequency tables in an effort to find impossible values for every variable collected. Unexpected findings could indicate problems in any of the previous three steps, and, as with any research study, necessitate an investigation into those data points. Finally, assumptions for each statistical test that has been planned will need to be checked.

Each step in the data extraction process should be carefully documented, because future researchers may attempt to replicate investigations of research questions using the original dataset or additional datasets. Data science scholars encourage planning analyses, investigating time dependence, and reporting all relevant method, code, and analyses to facilitate reproducibility (O'Neil & Schutt, 2013). Planning analyses prior to data extraction minimizes idiosyncratic decisions about removing cases. If time dependence is hypothesized to influence the research questions of interest, this

can be easily estimated by statistical approaches using timestamps and examining sections of the dataset, similar to the subset approach championed by meta-analytic scholars (Schmidt & Hunter, 2001). After initial data extraction, the theory of the data source may again need revision. As described in the case study, we discovered only after web scraping had concluded that gender was not included in the dataset for a substantial number of cases. Because gender was a central variable for our hypothesis and research questions, it was important to investigate this variable to ensure it existed for the reasons we thought it did, and we revised our data source theory accordingly based upon that investigation.

Legal, Practical, and Ethical Concerns When Web Scraping

Although the technical requirements of web scraping are relatively straightforward, researchers must also consider the legal, practical, and ethical implications of this method. Legally, researchers must primarily concern themselves with intellectual property and cybersecurity laws in their local jurisdictions. These laws differ across the globe, and thus researchers should review local and national policy on these issues before beginning a web scraping project (Band & Gerafi, 2013). To date, most legal action regarding digital rights has been in the United States, so the U.S. legal system will be the focus here. In contrast, practical and ethical concerns are more universal. Each will be discussed in turn.

Legal Concerns

The U.S. Legal System consists of officially codified legislation called statute law and noncodified standards derived from judicial decisions called case law (Ponzetto & Fernandez, 2008) that interact to determine the legality of a particular web scraping project. When a trial goes to court, the court's findings (i.e., their interpretation of existing statutes and case law) set legal precedent for future court cases (Field, 1986), establishing case law. Because technological advances can occur much more rapidly than the typical speed of legislatures, the language in legislation designed to govern technology and computers must be general or risk becoming obsolete. As a result, case law generally dictates the legality of online activities, and due to its dynamic nature, the case law surrounding web scraping, data aggregation, and digital property can seem inconsistent and difficult to interpret, even to lawyers specializing in this area. Thus, the historical context of a given set of laws, including their legal interpretations, is key to understanding and predicting the contemporary legal environment, especially so for laws that govern technology and digital property. For the sake of brevity, hereafter the term law will refer to the practical combination of statute and case law.

Federal copyright law, originally established by the Copyright Act of 1790 to protect original works of authorship from being copied, used, or distributed in a manner unapproved of by the author (Joyce, Leaffer, Jaszi, & Ochoa, 1998), is commonly applied to both websites and the databases they reference. In general, copyright law requires consent from the owner to repurpose or republish such works. However, there are many instances in which express consent is unnecessary for the use of materials, which is referred to as fair use (Madison, 2004). Fair use laws protect critics, commentators, reporters, teachers, researchers, authors, and artists, who wish to use, reuse, or repurpose copyrighted materials under certain circumstances. The evaluation of fairness includes four specific factors: the purpose of the use, the nature of the copyrighted work, the portion of the copyrighted work that was used, and any effect the use has on the copyrighted work's market value (Madison, 2004).

Specific legislation throughout the 1980s and 1990s provides the most relevant statutory groundwork for the legality of web scraping (Jarret & Bailie, 2011). In 1986, the U.S. Congress enacted the Computer Fraud and Abuse Act (CFAA) to combat nefarious activities, such as unauthorized access to government computers and national security information, general computer fraud and extortion, and property damage to computers or data, although this act has since been criticized as having too broad of language that is rooted in outdated concepts of computing (Jensen, 2014). To preserve online commerce by striking a balance between the rights of content creators and the rights of content consumers and sharers, the U.S. Congress passed the Digital Millennium Copyright Act of 1998 (Nimmer, 2000), which has two key provisions: Title I, which makes it a criminal and civil offense to circumvent measures that are put in place to protect digital copyrighted work, and Title II, which protects Internet service providers from being held liable for maintaining or distributing copyrighted material (Benchell, 2002). Under Title I, it may be legal to obtain unencrypted copyrighted materials from the web but illegal to remove encryption from any data not owned by the researcher, even if the data were purchased or if the encryption algorithm was legally obtained. Additionally, the computing era sprouted a reconceptualization of trespass to chattels, a common law concept that had mostly disappeared from American jurisprudence in the 1800s. Chattel refers to mobile personal property and is distinct from land property, which has its own set of laws. Trespass to chattels occurs when a third party intentionally and unwantedly meddles with, possesses, or modifies an individual's physical property, causing some tangible form of monetary or physical damage (Quilter, 2002). Trespass to chattels came to new relevance in 1996, when a California appellate court ruled that electronic signals are tangible physical property, finding two teenagers guilty of trespass when using a computer to access a telephone system. Subsequent cases have loosened the tangible damage requirement in trespass to chattels findings, and charges may be brought when even potential damage could occur (Quilter, 2002).

In the 2000s and 2010s, several prominent cases shaped digital property and copyright law specifically in relation to web scraping. According to Quilter (2002), a key case regarding web scraping was *eBay v. Bidder's Edge* (2000), in which eBay, Inc. sought an injunction against aggregator Bidder's Edge for utilizing a scraper to access several Internet auction sites, including eBay, and then compiling that information for its users on its own website. The court ruled against Bidder's Edge, finding that their crawler intentionally accessed eBay's system in a manner that caused potential harm by slowing down the connection for eBay's other users, thus constituting trespass to chattels (O'Reilly, 2007). In a similar case that same year, *Ticketmaster Corp., et al. v. Tickets.com, Inc.* (2000), Ticketmaster brought suit against Tickets.com for crawling and aggregating data from Ticketmaster's website. In this instance, it was found that ticket prices were factual, and therefore unable to be copyrighted and considered fair use (O'Reilly, 2007). The next year, Princeton Professor Edward Felten received a letter from the Recording Industry Association of America ordering him to destroy his research materials and to avoid

publicly speaking about his work which related to weaknesses in the Secure Digital Music Initiative copyright encryption system, in response to the challenge issued by that organization to demonstrate the strength of their product (Benchell, 2002). Felten responded by filing a legal complaint against the Recording Industry Association of America, which subsequently dropped the suit (Benchell, 2002). In 2013, security researcher and self-described hacker Andrew Auernheimer was sentenced to 41 months in prison for violating the Computer Fraud and Abuse Act (CFAA). He and his colleague, Daniel Spitler, found a security flaw in AT&T Inc.'s public web page during the 2010 release of the Apple iPad. Specifically, they were able to demonstrate that anyone could obtain the private email addresses of individuals who had purchased iPads. To demonstrate this, they designed a web crawler to iteratively gather email addresses from AT&T's public website (Doherty, 2013). Although the pair were security researchers and did not publically release private email addresses, the court found Auernheimer guilty of identity fraud and conspiracy to access a computer without authorization (Zetter, 2013). Auernheimer sought an appeal, hoping to highlight his case in an effort to change the laws governing Internet activity (Zetter, 2013), and his conviction was overturned in April 2014. Despite this, the appeals court did not clarify or modify the specific legality of his actions under the CFAA. Rather, the reversal was based on a technicality regarding the discrepancy between the physical location of AT&T's servers and New Jersey, the state in which the original trial took place (Bilton, 2014).

Given this ambiguity and inconsistency, we recommend researchers only scrape publicly available, unencrypted data sources to avoid legal risk. If a login or password is required to gain access, or if headers are included by the website provider to discourage automated software, web scraping could be considered illegitimate access to data by the website provider. In such cases, written permission must be acquired. The tension that arises when applying statutory law originally drafted long before computers existed to novel technological concepts and possibilities will undoubtedly continue, resulting in further high-profile cases mired in controversy. This tension shapes legal context not just for legal professionals, but for academics, practitioners, and society in general.

Practical and Ethical Concerns

The proliferation of nonhuman visitors to websites, including web crawlers, has led many sites and organizations to deploy various prevention mechanisms which can affect the practicality and ethicality of a web scraping project. Some websites include specific code in their webpages to request web crawlers not process them. However, web crawlers can easily be programmed to ignore these requests (Von Ahn, Blum, Hopper, & Langford, 2003). When this happens, some websites attempt to identify nonhuman visitors and block them from accessing the website or database (Von Ahn et al., 2003). In turn, some crawlers employ countermeasures to work around these restrictions (Chellapilla, Larson, Simard, & Czerwinski, 2005). Researchers should generally not engage in such practices, and scrapers should identify themselves as automated software, clearly indicating their origin. For example, the web scraper described in this paper included a few lines of code identifying itself as an "Old Dominion University Research Project." The presence of access-prevention measures is a clear signal that a website provider does not wish to share its data with scrapers, and researchers should adhere to this request in most circumstances. Ethically, the provider does not want to share data, and researchers should not take it by force. Practically, some countermeasures are specifically designed to damage the quality of data scraped. For example, a honeypot, which is a decoy used to lure a crawler away from desired content, can trap the crawler in an inescapable loop collecting useless data (Kreibich & Crowcroft, 2004; Spitzner, 2002).

Broad ethical guidelines for web scraping can be applied directly from the ethical standards codified by the American Psychological Association (APA). The Data Science Association (DSA), which is a major data science professional organization, also provides ethical guidelines, but these primarily describe the relationship between data scientists and the individual or organization that retains their service and are decidedly less comprehensive than those of the APA. The most recent APA (2010) directive, *Ethical Principles of Psychologists and Code of Ethics*, describes five overarching principles of conduct and 10 categories of ethical standards. Within the 10 ethical standards are subcategories, which are intended to encompass the wide range of psychologists' professional endeavors. The APA document covers a variety of topics such as research and reporting, record keeping, confidentiality, education and training, interacting with other professions, client relationships, and obligations that psychologist have to each other, the public, and the scientific community (APA, 2010). These ethical guidelines apply to web scraping just as they apply to any other data extraction or collection method.

The interpretation of these guidelines in the context of scraping does create unique ethical questions. To conduct the study described here, we sought ethics approval from our institution, which deemed the study to be exempt from review by our institutional review board. In our application for exemption, we stated that the study involved the observation of public behavior on the Internet, which thus does not require consent. Although this is consistent with U.S. law, there is increasing recognition that a number of Internet users, especially those on social network sites like Facebook, believe that the content they submit to those sites will be kept private (O'Brien & Torres, 2012). Although recent privacy controversies (e.g., regarding Kramer, Guillory & Hancock, 2014) have increased awareness that any data shared over the Internet has been in effect shared publicly, many people remain unaware of this. It is currently unclear if psychologists have a responsibility to obtain consent in such cases. Additionally, many websites that could be targeted for web scraping exist behind easily overcome barriers. For example, a discussion board website might require registration in order to view content, but registration is free and without restriction. In such cases, the data are publicly available, but the registration requirement might lead members to believe that their sharing exists in a "safe space" (Reid, 2011). Given this, it may be unethical to scrape such information, despite its public availability, without explicit consent. The specifics of ethics in this space remain an open question, and we encourage exploration of these issues.

More broadly, the theory-based web scraping method presented in this article also addresses some of the ethical challenges associated with the mainstream analytic approach of data science, which can be represented as empiricism absent of guiding theory. For example, the DSA (2015) describes the data science method as observation of data and relationships between data, deduction of meaning from data and data relationships, formation of hypotheses and finally experimental

or observational testing of the hypotheses. Existing theory surrounding the phenomenon being examined is not generally considered; instead, data science theory is concerned almost entirely with the nature of data in the abstract. When new theory is developed, it is more likely to concern the algorithms used rather than the target of those algorithms. This purely empirical approach may be appropriate for artificial intelligence research and techniques based upon it (Poole, 1989) but may be dangerous in the context of social science. In data-driven approaches to analyzing social science data, with potentially thousands of variables and no theoretical basis for their inclusion, one would expect a sample to contain a large number of relationships that exist only due to sampling error. A theory-driven approach reduces the temptation to examine these relationships and interpret them post hoc.

Conclusion

Although unanswered questions remain as with any new data collection technique, we contend that web scraping offers a great deal of potential for psychology by increasing access to behavioral data without significant researcher intrusion, dramatically increasing sample sizes, decreasing the amount of time spent during the data collection phase, increasing access to researchers in newly industrialized and undeveloped nations who cannot afford traditional large-scale research projects, and improving the interdisciplinary application of our vast research literature on psychometrics. Web scraping, at its core, is about finding meaning in patterns of human behavior, the fundamental goal of all psychological research. It is primarily the context that has changed. Regarding their own field of study, Savage and Burrows (2009) write that sociology is "losing whatever jurisdiction we once had over the study of the 'social' as the generation, mobilization and analysis of social data become ubiquitous." The same risk faces psychology. If our field does not more thoroughly explore Internet behavior—which may be viewed as just as valid an expression of human thought and emotion as "real life" behavior—the damage to our field's impact may be lasting.

In conclusion, theory-driven web scraping offers exciting possibilities for psychology, but there are many caveats. Most importantly, researchers should be cautious not to regress into brute force empiricism but instead establish a clear theory of their data source and target research questions a priori. It is worthwhile for researchers to carefully consider what theory can be tested, which results can be replicated meaningfully, and where theory can be reasonably expanded. Constructs should be operationalized carefully and explicitly in order to ensure their validity. The data source theory developed must be empirically tested to the maximum degree possible to establish data source validity. If such tests reveal that the data source is not as hypothesized and if those differences have interpretive implications, that data source should be abandoned, lest we be tempted into post hoc rationalization and hypothesizing. The greatest benefits from web scraping will be realized when combining the behavioral data to which it provides access with other data sources already commonly in use. Psychologists can have the most confidence in their theories when those theories are supported regardless of the approach taken to data collection and analysis. Perhaps most critically, we must avoid the data science version of Maslow's (1966) hammer: if all you have is web scraping, everything begins to look like data.

References

American Psychological Association (APA). (2010). *Ethical principles of psychologists and code of conduct*. Washington, DC: American Psychological Association.

Antoniou, G., & Van Harmelen, F. (2008). *A semantic web primer* (2nd ed.). Cambridge, MA: MIT press.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Eductional Data Mining, 1,* 3–16.

Band, J., & Gerafi, J. (2013). *The fair use/fair dealing handbook*. Retrieved from http://www.policybandwidth.com/publications

Benchell, N. A. (2002). The Digital Millennium Copyright Act: A review of the law and the court's interpretation. *Journal of Computer & Information Law, 21,* 1–18.

Bilton, N. (2014, April 11). Appeals court overturns conviction of AT&T hacker known as "Weev". *The New York Times*. Retrieved from bits.blogs.nytimes.com/2014/04/11/court-overturns-conviction-of-att-ipad-hacker

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478–494. http://dx.doi.org/10.1037/0021-9010.74.3.478

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society, 15,* 662–679. http://dx.doi.org/10.1080/1369118X.2012.678878

Cafarella, M. J., Halevy, A., & Madhavan, J. (2011). Structured data on the web. *Communications of the ACM, 54,* 72–79. http://dx.doi.org/10.1145/1897816.1897839

Chellapilla, K., Larson, K., Simard, P. Y., & Czerwinski, M. (2005). *Computers beat humans at single character recognition in reading based Human Interaction Proofs (HIPs)*. Paper presented at the CEAS, Stanford University, CA.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review, 69,* 21–26.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Conway, D., & White, J. M. (2012). *Machine learning for hackers*. Sebastopol, CA: O'Reilly Media.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Wide Web. In *Proceedings of the Ninth IEEE International Conference on Tools With Artificial Intelligence,* 558–567.

Copes, H., & Miller, J. M. (2015). *The Routeledge handbook of qualitative criminology*. New York, NY: Routeledge.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. http://dx.doi.org/10.1037/h0040957

Culpepper, S. A., & Aguinis, H. (2011). R is for revolution: A cutting-edge, free, open source statistical package. *Organizational Research Methods, 14,* 735–740.

Data Science Association (DSA). (2015). *Data science code of professional conduct* [PDF document]. Retrieved from http://www.datascienceassn.org/code-of-conduct.html

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92,* 1380–1393.

Doherty, M. (2013). *Legal risk in computer security research in Canada*. Unpublished manuscript, Department of Computer Science, Dalhousie University, Nova Scotia, Canada.

eBay v. Bidder's Edge, 100 F. Supp. 2d 1058 (N. D. Cal. 2000).

Evans, M., Donelle, L., & Hume-Loveland, L. (2012). Social support and online postpartum depression discussion groups: A content analysis. *Patient Education and Counseling, 87,* 405–410. http://dx.doi.org/10.1016/j.pec.2011.09.011

Field, M. A. (1986). Sources of law: The scope of federal common law. *Harvard Law Review, 99,* 881. http://dx.doi.org/10.2307/1341238

Hanson, W. E., Creswell, J. W., Clark, V. L. P., Petska, K. S., & Creswell, J. D. (2005). Mixed methods research designs in counseling psychology. *Journal of Counseling Psychology, 52,* 224–235.

Harrington, P. (2012). *Machine learning in action*. Shelter Island, N. Y.: Manning Publications Co.

Healthboards.com. (n.d.). *Depression message board & forum discussions—HealthBoards*. Retrieved from http://www.healthboards.com/boards/depression/

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33,* 61–83. http://dx.doi.org/10.1017/S0140525X0999152X

Herman, M., Rivera, S., Mills, S., Sullivan, J., Guerra, P., Cosmas, A., . . . Kim, M. (2013). *The field guide to data science*. Retrieved from http://www.boozallen.com/insights/2013/11/data-science-field-guide

Jarret, H. M., & Bailie, M. W. (2011). *Prosecuting computer crimes OLE Litigation Series*. Washington, DC: Office of Legal Education.

Jensen, S. (2014). Abusing the Computer Fraud and Abuse Act: Why broad interpretations of the CFAA fail. *Hamline Law Review, 36,* 81–138.

Joyce, C., Leaffer, M., Jaszi, P., & Ochoa, T. (1998). *Copyright law* (6th ed.). Newark, NJ: LexisNexis.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217. http://dx.doi.org/10.1207/s15327957pspr0203_4

Kessler, R. C., McGonagle, K. A., Swartz, M., Blazer, D. G., & Nelson, C. B. (1993). Sex and depression in the National Comorbidity Survey I: Lifetime prevalence, chronicity and recurrence. *Journal of Affective Disorders, 29,* 85–96.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759. http://dx.doi.org/10.1177/0013164496056005002

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academies of Science of the United States of America, 111,* 8788–8790. http://dx.doi.org/10.1073/pnas.1320040111

Kreibich, C., & Crowcroft, J. (2004). Honeycomb: Creating intrusion detection signatures using honeypots. *Computer Communication Review, 34,* 51–56. http://dx.doi.org/10.1145/972374.972384

Krueger, J. (2001). Null hypothesis significance testing. On the survival of a flawed method. *American Psychologist, 56,* 16–26. http://dx.doi.org/10.1037/0003-066X.56.1.16

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizations, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8,* 142–164. http://dx.doi.org/10.1017/iop.2015.13

Landers, R. N., Fink, A., & Collmus, A. B. (in press). Using big data to enhance staffing: Vast untapped resources or tempting honeypot? In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed.). New York, NY: Routledge.

Larson, W. (2008, August). *An introduction to compassionate screen scraping*. Retrieved from http://lethain.com/

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: Traps in big data analysis. *Science, 343,* 1203–1205. http://dx.doi.org/10.1126/science.1248506

Lengua, L. J., & Stormshak, E. A. (2000). Gender, gender roles, and personality: Gender differences in the prediction of coping and psychological symptoms. *Sex Roles, 43,* 787–820. http://dx.doi.org/10.1023/A:1011096604861

Lewin, K. (1935). *A dynamic theory of personality*. New York, NY: McGraw-Hill.

Liu, H. (2008). Social network profiles as taste performances. *Journal of Computer-Mediated Communication, 13,* 252–275. http://dx.doi.org/10.1111/j.1083-6101.2007.00395.x

Loukides, M. (2010). What is data science? The future belongs to the companies and people that turn data into products. *O'Reilly Radar Report*. Retrieved from http://www.verigazeteciligi.com/wp-content/uploads/2014/12/What_is_Data_Science.pdf

Madison, M. J. (2004). A pattern-oriented approach to fair use. *William and Mary Law Review, 45,* 1525–1690.

Marres, N. (2015). Why map issues? On controversy analysis as a digital method. *Science, Technology & Human Values, 40,* 655–686. http://dx.doi.org/10.1177/0162243915574602

Marres, N., & Weltevrede, E. (2013). Scraping the social? Issues in live social research. *Journal of Cultural Economics, 6,* 313–335. http://dx.doi.org/10.1080/17530350.2013.772070

Maslow, A. H. (1966). *The psychology of science*. New York, NY: Harper & Row.

Mathison, S. (1988). Why triangulate? *Educational Researcher, 17,* 13–17. http://dx.doi.org/10.3102/0013189X017002013

Moorthy, J., Lahiri, R., Biswas, N., Sanyal, D., Ranjan, J., Nanath, K., & Ghosh, P. (2015). Big data: Prospects and challenges. *The Journal for Decision Makers, 40,* 74–96.

Nimmer, D. (2000). A riff on fair use in the Digital Millennium Copyright Act. *University of Pennsylvania Law Review, 148,* 673–742. http://dx.doi.org/10.2307/3312825

Nolen-Hoeksema, S. (2001). Gender differences in depression. *Current Directions in Psychological Science, 10,* 173–176. http://dx.doi.org/10.1111/1467-8721.00142

O'Brien, D., & Torres, A. M. (2012). Social networking and online privacy: Facebook users' perceptions. *Irish Journal of Management, 31,* 63–97.

O'Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. Beijing, China: O'Reilly Media.

O'Reilly, S. (2007). Nominative fair use and Internet aggregators: Copyright and trademark challenges posed by bots, web crawlers and screen-scraping technologies. *Loyola Consumer Law Review, 19,* 273–288.

Piccinelli, M., & Wilkinson, G. (2000). Gender differences in depression: Critical review. *The British Journal of Psychiatry, 177,* 486–492. http://dx.doi.org/10.1192/bjp.177.6.486

Piko, B. (2001). Gender differences and similarities in adolescents' ways of coping. *The Psychological Record, 51,* 223–235.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88,* 879–903. http://dx.doi.org/10.1037/0021-9010.88.5.879

Ponzetto, G. A. M., & Fernandez, P. A. (2008). Case law versus statute law: An evolutionary comparison. *The Journal of Legal Studies, 37,* 379–430. http://dx.doi.org/10.1086/533421

Poole, D. (1989). Explanation and prediction: An architecture for default and abductive reasoning. *Computational Intelligence, 5,* 97–110. http://dx.doi.org/10.1111/j.1467-8640.1989.tb00319.x

Quilter, L. (2002). The continuing expansion of cyberspace trespass to chattels. *Berkeley Technology Law Journal, 17,* 421–443.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reid, J. (2011). "We don't Twitter, we Facebook": An alternative pedagogical space that enables critical practices in relation to writing. *English Teaching, 10,* 58–80.

Savage, M., & Burrows, R. (2009). Some further reflections on the coming crisis of empirical sociology. *Sociology, 43,* 762–772. http://dx.doi.org/10.1177/0038038509105420

Schmidt, F. L., & Hunter, J. E. (2001). Meta-analysis. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work & organizational psychology: Vol. 1: Personnel psychology* (pp. 51–70). Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shaw, L. H., & Gant, L. M. (2002). In defense of the Internet: The relationship between Internet communication and depression, loneliness, self-esteem, and perceived social support. *Cyberpsychology & Behavior, 5,* 157–171. http://dx.doi.org/10.1089/109493102753770552

Spitzner, L. (2002). *Honeypots: Tracking hackers* (Vol. 1). Boston, MA: Addison Wesley Reading.

Thelwall, M., & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research, 26,* 162–176. http://dx.doi.org/10.1016/j.lisr.2003.12.009

The World Bank. (2014). *Population, female (% of total)*. Retrieved from http://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS

Ticketmaster Corp., et al. v. Tickets.com, Inc., CV 99–7654 HLH (BQRx) (2000 U.S. Dist.).

Von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. (2003). CAPTCHA: Using hard AI problems for security. *EUROCRYPT, 2003,* 297–311.

Wallace, P. (2001). *The psychology of the Internet*. Cambridge, United Kingdom: Cambridge University Press.

Wentworth, P., Elkner, J., Downey, A. B., & Meyers, C. (2012). *How to think like a computer scientist: Learning with Python 3*. The Open Book Project: http://openbookproject.net/thinkcs/python/english3e/

World Wide Web Consortium (W3C). (2014, October 28). *HTML5*. Retrieved from http://www.w3.org/TR/html/

Young, S., Dutta, D., & Dommety, G. (2009). Extrapolating psychological insights from Facebook profiles: A study of religion and relationship status. *Cyberpsychology & Behavior, 12,* 347–350. http://dx.doi.org/10.1089/cpb.2008.0165

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 1036–1040. http://dx.doi.org/10.1073/pnas.1418680112

Zetter, K. (2013, March 18). AT&T hacker "Weev" sentenced to 3.5 years in prison. *Wired*. Retrieved from http://www.wired.com/2013/03/att-hacker-gets-3-years/

Appendix

Python Code Used in This Study

The following code was used for the case study contained within this article. It requires the installation of the open source Python programming language, as well as the open source package for Python called Scrapy. It can be easily modified for similar scraping projects. This code takes approximately 20 hr to execute and will contact the healthboards.com website at least 120,000 times, so it is recommended that readers do not execute it exactly as written. A full tutorial introducing psychologists to Python and Scrapy can be found at http://rlanders.net/scrapytutorial.html.

In /depression/items.py:

```
import scrapy

class DepressionItem(scrapy.Item):
    threadNumber = scrapy.Field()        # count from the URL
    threadTitle = scrapy.Field()         # title from either the URL or the post itself
    postNumber = scrapy.Field()          # top right of each post
    postDatetime = scrapy.Field()        # in blue bar at top of each post
    postUsername = scrapy.Field()        # larger font at top left of each post
    postCount = scrapy.Field()           # information below username
    postJoinDate = scrapy.Field()        # information below username
    postGender = scrapy.Field()          # information below username
    postContent = scrapy.Field()         # all text, in original HTML
    postID = scrapy.Field()              # post ID number
```

In /depression/settings.py:

```
BOT_NAME = 'depression'
SPIDER_MODULES = ['depression.spiders']
NEWSPIDER_MODULE = 'depression.spiders'
USER_AGENT = 'Old Dominion University Research Project'
DOWNLOAD_DELAY = 2                  # enforces a mean 2-second delay between requests
```

(*Appendix continues*)

In /depression/spiders/depression_spider.py:

```
import scrapy
import re
from depression.items import DepressionItem
from scrapy.contrib.spiders import CrawlSpider, Rule
from scrapy.contrib.linkextractors import LinkExtractor

class DepressionSpider(CrawlSpider):
    name = "depression"
    allowed_domains = ["healthboards.com"]
    start_urls = (
        'http://www.healthboards.com/boards/depression/',
    )

    rules = (
        Rule(LinkExtractor(
            allow=('www\.healthboards\.com/boards/depression/\d*-.*.html', ),
            deny=('www\.healthboards\.com/boards/depression/\d*-.*-print.html'), ),
             callback='parse_crawled', follow=True),
        Rule(LinkExtractor(
            allow=('www\.healthboards\.com/boards/depression/index\d*\.html', ))),
    )

    def parse_crawled(self, response):
        pagecases = scrapy.Selector(response).xpath('//div')
        items = []

        for case in pagecases:
            if ",".join(re.findall('^<div id="edit(\d*)"',case.extract())) != '':
                item = DepressionItem()
                item["postID"] = ",".join(
                    re.findall('^<div id="edit(\d*)"',case.extract()))
                item["threadNumber"] = ",".join(
                    re.findall('depression/(\d*)-',response.url))
                item["threadTitle"] = case.xpath(
                    'table/tr/td[@class="alt1"]/div/strong/text()').extract()
                postNumWrapper = case.xpath('table/tr/td/a/@name')
                item["postNumber"] = ",".join(
                    re.findall("^\d*$",postNumWrapper[1].extract()))
                postDateTime = case.xpath('table/tr/td')
                item["postDatetime"] = ",".join(
                    re.findall(
                       "\d\d-\d\d-\d\d\d\d, \d\d:\d\d .M",postDateTime[0].extract()))
                item["postUsername"] = case.xpath(
                    'table/tr/td[@class="alt2"]/div/a/text()').extract()
                postCountList = case.xpath(
                    'table/tr/td[@class="alt2"]/div[@class="smallfont"]/div')
                for postCount in postCountList:
                    if ",".join(re.findall('Posts',postCount.extract())) != '':
                        item["postCount"] = ",".join(
                            re.findall("Posts: (\d*)",postCount.extract()))
                    if ",".join(re.findall('Join Date',postCount.extract())) != '':
                        item["postJoinDate"] = ",".join(
                            re.findall("Join Date: (.*)</div>",postCount.extract()))
                item["postGender"] = case.xpath(
                    'table/tr/td[@class="alt2"]/font/text()').extract()
```

(*Appendix continues*)

```
            contentString = ('table/tr/td[@class="alt1"]/div[@id="post_message_' +
                             item["postID"] + '"]/text()')
            item["postContent"] = "\n".join(
                case.xpath(contentString).extract()).strip()

            items.append(item)

    return items
```

Received May 30, 2015
Revision received January 4, 2016
Accepted January 27, 2016 ■

UNITED STATES POSTAL SERVICE® Statement of Ownership, Management, and Circulation (All Periodicals Publications Except Requester Publications)

1. Publication Title	2. Publication Number	3. Filing Date
Psychological Methods	0 1 6 - 1 0 9	October 2016
4. Issue Frequency	5. Number of Issues Published Annually	6. Annual Subscription Price
Quarterly	4	Mbr $64 Indiv $145 Inst $648

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4®): 750 First Street, NE, Washington, DC 20002-4242 — Contact Person: Barbara Spruill, Telephone (Include area code): 202 336-5578

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer): 750 First Street, NE, Washington, DC 20002-4242

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank)

Publisher (Name and complete mailing address): American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242

Editor (Name and complete mailing address): Lisa Harlow, Department of Psychology, University of Rhode Island, 10 Chafee Road, Ste. 8, Kingston, RI 02881

Managing Editor (Name and complete mailing address): Susan J.A. Harris, 750 First Street, NE, Washington, DC 20002-4242

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

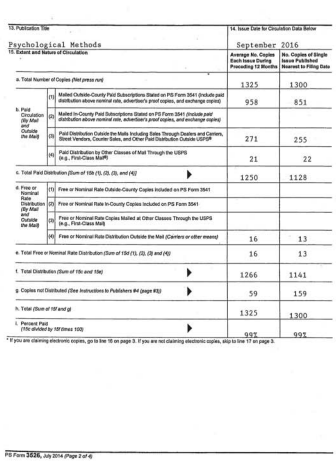
Full Name	Complete Mailing Address
American Psychological Association	750 First Street, NE, Washington, DC 20002-4242

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → ☒ None

Full Name	Complete Mailing Address
None	

12. Tax Status (For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

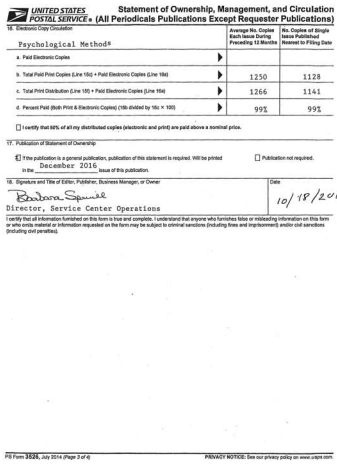
PS Form 3526, July 2014 (Page 1 of 4 (see instructions page 4)) PSN: 7530-01-000-9931 PRIVACY NOTICE: See our privacy policy on www.usps.com.



13. Publication Title: Psychological Methods	14. Issue Date for Circulation Data Below: September 2016	
15. Extent and Nature of Circulation	Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net press run)	1325	1300
b. Paid Circulation (By Mail and Outside the Mail) (1) Mailed Outside-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)	958	851
(2) Mailed In-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)		
(3) Paid Distribution Outside the Mails Including Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Paid Distribution Outside USPS®	271	255
(4) Paid Distribution by Other Classes of Mail Through the USPS (e.g., First-Class Mail®)	21	22
c. Total Paid Distribution (Sum of 15b (1), (2), (3), and (4))	1250	1128
d. Free or Nominal Rate Distribution (By Mail and Outside the Mail) (1) Free or Nominal Rate Outside-County Copies included on PS Form 3541		
(2) Free or Nominal Rate In-County Copies included on PS Form 3541		
(3) Free or Nominal Rate Copies Mailed at Other Classes Through the USPS (e.g., First-Class Mail)		
(4) Free or Nominal Rate Distribution Outside the Mail (Carriers or other means)	16	13
e. Total Free or Nominal Rate Distribution (Sum of 15d (1), (2), (3) and (4))	16	13
f. Total Distribution (Sum of 15c and 15e)	1266	1141
g. Copies not Distributed (See Instructions to Publishers #4 (page #3))	59	159
h. Total (Sum of 15f and g)	1325	1300
i. Percent Paid (15c divided by 15f times 100)	99%	99%

* If you are claiming electronic copies, go to line 16 on page 3. If you are not claiming electronic copies, skip to line 17 on page 3.

PS Form 3526, July 2014 (Page 2 of 4)



UNITED STATES POSTAL SERVICE® Statement of Ownership, Management, and Circulation (All Periodicals Publications Except Requester Publications)

16. Electronic Copy Circulation: Psychological Method®	Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Paid Electronic Copies		
b. Total Paid Print Copies (Line 15c) + Paid Electronic Copies (Line 16a)	1250	1128
c. Total Print Distribution (Line 15f) + Paid Electronic Copies (Line 16a)	1266	1141
d. Percent Paid (Both Print & Electronic Copies) (16b divided by 16c × 100)	99%	99%

☐ I certify that 50% of all my distributed copies (electronic and print) are paid above a nominal price.

17. Publication of Statement of Ownership
☒ If the publication is a general publication, publication of this statement is required. Will be printed in the December 2016 issue of this publication. ☐ Publication not required.

18. Signature and Title of Editor, Publisher, Business Manager, or Owner: Barbara Spruill, Director, Service Center Operations — Date: 10/18/2016

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).

PS Form 3526, July 2014 (Page 3 of 4) PRIVACY NOTICE: See our privacy policy on www.usps.com.