Conducting research on the internet – a new era

Claire Hewson provides updates on practicalities and possibilities

C an psychological research studies conducted via the internet provide valid and reliable data? This is the question I posed in The Psychologist more than a decade ago (Hewson, 2003). Here I consider this question again, drawing upon the wealth of new examples and relevant research, and conclude that the answer is now a resounding ‘yes’.

Many of the more optimistic early predictions about internet-mediated research (IMR, primary research conducted via the internet) have now been borne out. Psychologists have been amongst those most willing to embrace this new mode of data collection. With an emphasis on what IMR has to offer psychological primary research, I here consider recent developments, the ‘state-of-the-art’, and future predictions. In what is now arguably a ‘new era’ of IMR, in which ‘Web 2.0’ and social media activities play a central role, I hope to show that IMR is now an established methodological approach in social and behavioural research, with many demonstrated successes. In order to assist the researcher wishing to implement an IMR study, I also include an updated set of resources, and good practice guidelines.

A decade of IMR

IMR pioneers started publishing the results of their early investigative work from around the mid-1990s. Commentaries on this new mode of data collection started appearing around the early 2000s, in the form of books devoted to the topic (e.g. Mann & Stewart, 2000), and review articles (e.g. Nosek et al., 2002). Issues of concern at that time included sample bias, possible influences of the online communication medium, and reliability of procedures. Some key advantages also made IMR very attractive, including time and cost-efficiency, and facilitated access to a large participant pool. Over the last decade or so IMR has flourished, enjoying expansive multidisciplinary reach. There now exists a wealth of examples and good practice guidelines for today’s IMR researchers to draw upon. Many currently active studies can be found via online study clearing houses: ‘Useful links’ (opposite) lists several such sites, and other dedicated IMR events and resources.

Surveys remain the most widely used method in IMR. Early examples often used e-mail, but developments in the availability of supporting software make web-based approaches most popular now (advantages including enhanced control and functionality). Experimental approaches are also well-established, following the success of early examples. IMR interviews have had a more ambivalent reception, early reports indicating mixed levels of success (see below). Document analysis approaches have proved largely successful, though they are probably the least reported method to date.

Particularly noteworthy over the last decade or so has been the expansion and diversification of observational approaches in IMR. The emergence of ‘Web 2.0’ (the term first introduced by DiNucci, 1999) is intricately linked to these developments. Web 2.0 refers to the notion of the world wide web as a fluid, organic space where people interact and collaborate to create continually updated content; this can be contrasted with earlier conceptions of the web as a relatively static space for the publication and dissemination of information (though Tim Berners-Lee has contested this conception; see: tinyurl.com/dd97ym).

Social networking sites such as Facebook are prime examples of Web 2.0 technologies, as are ongoing mass online collaborations such as Wikipedia and OpenStreetMap. So are blogs, online discussion forums, and (the microblog) Twitter.

The recent fascination with Web 2.0 technologies, and the communities they support, arises from the recognition that the mass of data traces they leave behind of people’s online behaviours and interactions creates a vast pool of

References


Cane, F.D., Ho, T., McFadden, D. et al. (2012). Using the internet to administer more realistic vignette
potential research data. The sheer volume of online potential data sources can seem mind-boggling. Indeed, recent discussions of ‘big data’ reflect this. The term originates from within computer science, referring to data sets that become so large as to be awkward to work with. The potential for online sources to create very large data sets that are of interest to social scientists means that new methods for storage, analysis and presentation of results are needed in order to work with these data sets. Google Analytics (an internet analytics service) is a commercial service (aimed primarily at businesses) which allows the tracking and analysis of very large data sets generated by webpage visitor activity. Possible sources of big data sets with potential relevance to psychological research are considered below.

The changing face of the internet
In the early days of IMR, concerns were raised that samples accessed online were likely to consist largely of white, educated, middle-class, technologically proficient, professional males. If this point was perhaps overemphasised at the time, there were also good reasons to think internet users would be biased towards this profile. Such concerns are today attenuated, due to both shifting patterns of internet usage and new evidence on the profiles of users. Confirming my predictions back in 2003, internet use has grown rapidly, from around 500 million users then to approaching 3 billion now (see tinyurl.com/oeeb9m ). Claims that this population would also diversify are supported (e.g. Hewson & Laurent, 2008), with the Oxford Internet Surveys and the World Internet Project indicating that the individual user profile has both expanded and diversified. Of course, there are still some biases: for example, although the gender divide seems to have disappeared, users are still more likely to be younger, wealthier and more highly educated (Dutton & Blank, 2011).

Considering all the evidence, IMR researchers today have an extremely large, diverse population of potential participants to draw upon.

Reliability and representativeness
Reconsidering the question of whether IMR methods can generate valid, reliable data, there is now very strong evidence that this is the case, across a broad range of domains and methods.

Early ‘validation studies’ offered promising findings, suggesting that IMR data was of at least comparable quality to data gathered offline. Such research also showed IMR samples to be, if anything, in many ways more diverse than traditional offline samples (particularly in psychological research, where traditional approaches often use undergraduate student samples: see Arnett, 2008). Many more recent studies have reached similar conclusions. For example, Hewson and Charlton (2005) administered the Multidimensional Health Locus of Control (MHLC) Scale (Walston & Walston, 1981) in both web-based and pen-and-paper modes; the internet data was found to be at least as good as the offline data, considering scale reliabilities and factor structures. Other studies have generated similar support for IMR questionnaires (e.g. Brock et al., 2012) and experiments (e.g. Limman et al., 2006). Only a few studies have reported a lack of equivalence (e.g. Barbeite & Weiss, 2004), and it is often unclear in these cases whether the online or offline data is superior. Whereas previously it has been suggested that experiments involving precise timings are problematic in IMR (e.g. Hewson et al., 2003), several studies have now demonstrated that millisecond accuracy can be achieved using web-based methods (e.g. Keller et al., 2009). Experiments requiring the download of large audio and video files were also once considered problematic, but examples have now shown this can be successfully achieved (e.g. Caro et al., 2012).

Regarding the issue of sample representativeness in IMR (the most common approach being to recruit participants online), a number of studies have now set out to directly address this issue. Studies comparing online volunteer samples with traditional similar offline recruitment have reported internet samples to be more diverse (e.g. Gosling et al., 2004). While we would not expect online volunteer samples to be as representative as those randomly selected offline and this has indeed been demonstrated, e.g. Malhotra & Krosnick, 2007), some studies have nevertheless found such online samples to generate
methods

comparable data, for example when looking at relationships between variables, or psychometric test properties (e.g. Miller et al., 2010). Also, some studies have reported successfully obtaining probability samples online that display good levels of generalisability (e.g. Yeager et al., 2011). A noteworthy recent development is the availability of large online participant panels, some of which have been recruited using offline probability methods (see ‘Useful links’). However, these are typically expensive to access.

In summary, the evidence to date on the quality of internet samples is extremely promising, and a range of sampling options are currently available (see Hewson et al., in press, for a more detailed overview). IMR methods may help to address the heavy reliance on students in some disciplines (e.g. psychology), and even, to some extent, on populations biased towards the ‘WEIRD’ (Western, Educated, Industrialised, Rich, Democratic: see Henrich et al., 2010). Recently, obtaining representative probability samples online has become more viable. As always, individual research goals will dictate the most suitable sampling procedures to employ, whatever the mode of administration.

Favoured methods

Surveys and experiments are now well-established, commonly used IMR methods, as noted above. Interview and focus group research in IMR has not enjoyed quite the same success. Perhaps early mixed reports contributed to the less than enthusiastic uptake of these methods. For example, some authors noted problems in establishing good levels of rapport with participants, and obtaining high-quality, rich, reflective data (e.g. Bowker & Tufﬁn, 2004). Asynchronous approaches seem to have been more widely used than synchronous approaches (the greater technical skill levels demanded by the latter approach being a possible barrier: O’Connor et al., 2008). Nevertheless, some beneﬁts of online interview methods have been evidenced, such as the facilitation of access to hard-to-reach groups (e.g. Barratt, 2012). Several authors do report managing to obtain high-quality data using these methods, including synchronous approaches (e.g. O’Connor & Madge, 2003).

Certain elements of good practice seem important in distinguishing between the more and less successful examples, such as adopting clear strategies to establish good rapport with participants prior to conducting an interview or focus group. One noteworthy point is that early speculations that enhanced levels of anonymity in IMR might have beneﬁts in reducing biases from perceptions of biosocial features, or perhaps allow the manipulation of such features (e.g. Hewson, 2003), have not been embraced by interview researchers. Rather the approach has been to make every effort to recreate as closely as possible the intimate and personal nature of the offline face-to-face interview setting (e.g. O’Connor & Madge, 2003). When carefully implemented this seems to have worked well. While some researchers have tried incorporating audio and video in online interviews (e.g. Hanna, 2012), this approach awaits further technological developments to achieve good reliability.

Observation methods in IMR, including unobtrusive data-mining techniques, are currently receiving great interest. Early examples have illustrated the value of unobtrusive observational approaches for gathering linguistic data in ways that are hard to achieve online (e.g. Bordia, 1996). Recent examples include Brady and Guerin’s (2010) unobtrusive, qualitative analysis of postings to a discussion board on a parenting support website. This approach can be particularly useful in ethnographic research, where it can be combined with other methods, such as interviews and surveys.

While there are now many examples of the use of linguistic traces and sometimes real-time ‘live’ discussions (e.g. Brotsky & Giles, 2007), the web now offers a multitude of possibilities for also gaining access to non-linguistic traces and data sources, including multimedia material and online user activity. It is useful to distinguish between the use of ‘contentful’ material, and data on ‘structures, patterns and processes’. Recent examples which use ‘contentful’ data include an analysis of the content of multimedia social networking websites used by alcohol marketing agencies (McCreanor et al., 2013), and a content analysis of 417 obesity-related YouTube videos (Yoo & Kim, 2012). Techniques using data on structures and processes include online social network analysis (SNA), which focuses on analysing patterns and connections, such as friendship links, status likes, and so on (see Hogan, 2008).


Another example of process data is webpage navigations; such data traces have the potential to generate potentially extremely large ‘big data’ sets. In psychological research, the role and use of such data sets remains largely to be explored. Tonkin et al. (2012) present a relevant example; they analysed 600,000 tweets on the August 2012 riots in the UK, looking for evidence that, amongst other things, Twitter served as a central organisational tool to promote illegal group action (which the authors report finding not to be the case). This demonstrates the type of approach that might generate findings of interest to psychological research. See also the current project being run at the Oxford Internet Institute (OII) investigating the role of big data in social science research (tinyurl.com/b9fwsmq).

While observational methods in IMR can be characterised as involving online behaviours and interactions (both intra- and inter-individual), document analysis approaches make use of relatively static, published products (e.g. scientific articles, newspaper articles, webpages, repositories of art work). Though less widely used than other methods, examples relevant to psychological topics have emerged. For example, Horvath et al. (2012) sampled data from 462 webpages to look at the nature of webpage defacement by hackers.

In some contexts (e.g. blogs analysis), document analysis and observation approaches in IMR can become blurred. One such example which could be viewed as straddling this boundary is the ‘We Feel Fine’ project by Jonathan Harris and Sep Kamvar (http://wefelfine.org). This fascinating project involves collecting large volumes of data on human feelings from blog posts (in English) worldwide, on a daily basis and updated by the minute. Visiting the homepage provides access to an ongoing live feed of human emotional expressions harvested from internet blogs around the world, offering summary data presentation formats, and access to specific individual posts in multimedia formats (images and text).

Practicalities and ethics
Tools and software developments over the last decade or so have dramatically facilitated web-based survey implementation (and potentially data quality, e.g. by incorporating answer validity checks, preventing multiple submissions, etc.). Resources for assisting in constructing online experiments also exist, but generally demand more advanced computing skills. Interviews

**Tools and texts**

**Software tools**
- SurveyMonkey: surveymonkey.com
- Qualtrics, online survey package offering a sophisticated suite of facilities: qualtrics.com
- Google forms, free solution for creating basic online surveys: docs.google.com/forms
- Websm.org: comprehensive database of online survey software solutions
- WEBEXP, developed at the School of Informatics, University of Edinburgh, requires advanced computing skills, including running a web server: www.webexp.info
- Web-Harvest, open source web data extraction tool: www.web-harvest.sourceforge.net
- WEXTOR, online experiment generator providing web interface and server hosting: http://wextor.org/wextor
- LogAnalyzer, a tool for analysing server log files, such as those generated by webpages: www.sclog.eu
- iScienecemaps, a tool for researching Twitter content: http://maps.iscience.deusto.es

**Textbooks**
- M.P. Couper, Designing effective web surveys (Cambridge University Press, 2008)
- S.D Osling & J.A Johnson (Eds.), Advanced methods for conducting online behavioral research (American Psychological Association, 2010)
- C. Hewson, D. Laurent & C. Vogel, Internet research methods: A practical guide for the behavioural and social sciences, 2nd edn (Sage, in press)
- C. Hine, Virtual methods: Issues in social research on the internet (Berg, 2005)
- J. Salmons, Online interviews in real time (Sage, 2009)
are relatively straightforward to set up and run, particularly now more people are familiar with online discussion and chat technologies (which now also have more user-friendly interfaces). Observational approaches involving data-mining and scraping techniques can be rather complex, but tools to assist have started emerging. See ‘Tools and texts’ (previous page) for a selected list of current useful resources.

Of course, ethical considerations must play a central role in informing design decisions. One issue that remains unresolved in IMR is the blurred distinction between public and private spaces online, and the implications for approaches such as undisclosed data mining which waive informed consent. The reader is referred to the recent BPS guidelines on ethics in internet research (BPS, 2013) for further discussion, and some updates and additions to the good practice guidelines I suggested in 2003 are included (see box).

To the future

Ongoing technological developments, as well as cultural developments in internet usage, are bound to continue to change the landscape of IMR. Some noteworthy developments are the expansion in the use of mobile devices, which creates a situation of users being ever-connected. Smartphone apps facilitate this integration, enhancing the range of online activities that can be conducted ‘on the move’. Apart from the wealth of traces of online activities from people all over the world that smartphones create, opening opportunities for unobtrusive approaches, intriguing potential applications in unobtrusive online methods emerge. Thus, surveys and interviews may start to probe participants while on the move, in different specified locations and contexts, using multimedia and geolocation sources as data rather than relying on subsequent self-reports. ‘Walking interviews’ (e.g. Jones et al., 2008) might be an approach particularly amenable to such strategies. Interactive surveys also may be feasible, gathering richer in situ data. There are now even apps that can monitor physiological states, such as heart rate. The notion of a virtual survey interviewer has been suggested (Vehovar & Manfreda, 2008). All these possible developments have the potential to reassess the way we think about traditional methods, and the delineations between them, perhaps leading to increasingly blurred boundaries between methods in an IMR context.

Another intriguing development attracting current interest is the so-called ‘Internet of Things’; that is, the increasing internet-connectedness of everyday objects, such as televisions, fridges, cars, running shoes, tablets (of the medicinal, as well as the personal computer, variety), etc. The scope for ongoing data-collection possibilities tracking people in their everyday lives is of course enhanced by such developments. Indeed, the logging of contextual information on people’s everyday activities via traces from mobile devices is already pervasive; for example, consider the range of mobile apps that now request permission to collect and store geolocation information, often used to improve the user experience (e.g. photography apps that time, date and location stamp individual snapshots). Thus, individual profiles can now often be linked to physical locations and activities. Historically, behaviour sampling and other such methods have been used to try to examine the everyday lives of individuals; now individuals are voluntarily logging this information themselves online. In future it is easy to see how this data could be linked with other psychologically relevant data, such as personality, health, and so on.

Possibilities for mixed methods research (which traditionally is often cost- and time-intensive) are also likely to be expanded. Thus, while unobtrusive observational data can be automatically recorded or extracted from existing traces, interactive survey and/or interview methods can be efficiently implemented to supplement such data, potentially with participants from a diverse range of locations. Virtual reality environments (VREs) may also play a more prominent role in IMR methods in the future, being used to conduct interviews or focus groups (e.g. Stewart & Williams, 2005), as well as carry out observations.

The internet is constantly evolving, and will always stay one step ahead of my crystal ball gazing. But when you consider a site collecting longitudinal psychometric test data that has now obtained over 7.5 million responses (http://mypersonality.org), it is clear that the sheer scale of what is possible is mind-boggling. See you in 10 years?

I Claire Hewson is a lecturer in psychology at the Open University claire.hewson@open.ac.uk

Good practice

General principles

- Aim to use software solutions that have functions to help maximise control, reliability and validity, such as multiple submission checking and prevention, and response format and completeness checking.
- Aim to use software solutions that conform to established accessibility standards (e.g. compatibility with screen reader software).
- Remain mindful of legal (as well as ethical) issues, including copyright and data protection, especially when harvesting online data sources.

Sampling

- Remain mindful that volunteer samples may not be suitable where broad generalisability is required, and consider alternative offline and online probability methods.
- Remain mindful of the trade-offs involved in ensuring anonymity and gaining participant characteristics information, assessing decisions in the context of study design, context and goals.
- If posting participation requests to newsgroups and other social spaces, it is typically good practice to contact moderators or gatekeepers first to request permission.

Ethics

- If planning to use data without gaining informed consent, carefully consider issues related to the blurred public–private distinction online, and particularly the potential risks of harm due to leakage of personally identifiable data.
- Remain mindful of the enhanced traceability and searchability of data in online contexts, and convey to participants [in informed consent procedures] any substantial risks to the confidentiality of their data, and possible consequences.
- If considering using deception, or highly sensitive materials, be particularly aware of the extra risks that may arise from lower levels of reliability online in (a) verifying participant characteristics and (b) presenting debrief information.