



Web-Based Research in Psychology

A Review

Ulf-Dietrich Reips

Psychological Methods and Assessment / Experimental Psychology and Internet Science, Department of Psychology,
University of Konstanz, Germany

Abstract: The present article reviews web-based research in psychology. It captures principles, learnings, and trends in several types of web-based research that show similar developments related to web technology and its major shifts (e.g., appearance of search engines, browser wars, deep web, commercialization, web services, HTML5...) as well as distinct challenges. The types of web-based research discussed are web surveys and questionnaire research, web-based tests, web experiments, Mobile Experience Sampling, and non-reactive web research, including big data. A number of web-based methods are presented and discussed that turned out to become important in research methodology. These are one-item-one-screen design, seriousness check, instruction manipulation and other attention checks, multiple site entry technique, subsampling technique, warm-up technique, and web-based measurement. Pitfalls and best practices are described then, especially regarding dropout and other non-response, recruitment of participants, and interaction between technology and psychological factors. The review concludes with a discussion of important concepts that have developed over 25 years and an outlook on future developments in web-based research.

Keywords: Internet-based research, online experiments, online research, online assessment, web surveys

While Internet-based research began shortly after the invention of the Internet in the 1960s, web-based research in psychology began only in the mid-1990s once the world wide web (short: web) had been invented by Tim Berners-Lee in Geneva in 1992. Web browsers had become available, and subsequently, the http protocol had been amended by the functionality of *forms*. Forms allowed a web browser user to send back responses to what someone had set up on a web page as response options, for example, radio buttons, drop-down menus, check boxes, or text fields (Birbaum, 2004; Reips, 2000).

Web-based research naturally relies on technology and the principles of remote research. As we have seen, a number of technologies have competed and continue to compete for a place among essential requirements. Surviving technologies and concepts are from both server- and client-side (with advancements in quick switching and integration first via AJAX, then in HTML5 technology). Recently, however, Garaizar and Reips (2019) have concluded that the complexity of browser technologies has increased so much that web-based research is facing difficulties that will continue to cumulate. Anwyl-Irvine and colleagues (2021) compared the precision and accuracy of some online experiment platforms, web browsers, and devices. Their data confirm what became clear as an intrinsic difference at the onset of web-based research: Variance

in any measurement is higher than under laboratory-based conditions, especially compared to calibrated equipment of a limited type. Variance on the web comes unsystematically and systematically, as in earlier research, Anwyl-Irvine and colleagues found differences between online experiment platforms, web browsers (and operating systems), and devices. While during the first decades of web-based research, browsers (and main technologies) were updated only every few years or months, browser vendors now update weekly or even daily (Garaizar & Reips, 2019; see e.g., for Firefox at https://en.wikipedia.org/wiki/Firefox_version_history). From a researcher's point of view, both the stimulus display technology and the measurement device change frequently during data collection. Furthermore, browsers are not being optimized to meet researchers' needs. The main reason behind some of the difficulties is the browser vendors' motivation to optimize browsers in terms of the user experience to serve their commercial outlook. Browser vendors primarily want to sell, not help science find the truth.

With the advent of new technologies, the story of web-based research in psychology continues in a similar vein, yet to new frontiers. Similar to Musch and Reips (2000, replicated by Krantz & Reips, 2017), who once surveyed the first web experimenters, Ratcliffe and colleagues (2021) recently web-surveyed those who currently pioneer

remote research with Extended Reality (XR) technology – such as virtual and augmented reality. They found a similar hope by XR researchers to the early web researchers for “benefits of remote research for increasing the amount, diversity and segmentation of participants compared with in-lab studies” (Wolfe, 2017, p. 9) and also in parallel to web-based research. In reality that this hope rarely materializes, probably as a result of many technical and procedural limitations. While XR researchers see opportunities to further develop web-based XR, the history of web-based research shows that research requiring any addition to the core interface – the browser – may be doomed to fail or end up in a small niche of limited showcases that never make it to routine research methods.

Types of Web-Based Research

Essentially, all traditional methods in behavioral research have been transferred to digital research in one way or the other. However, naturally, some methods transfer more easily than others, which may change over time with the change of technology. Major shifts (now sometimes trendily called “disruptions”) in web technology always caused new methods to flourish. Examples are (1) the implementation of forms in HTML in the mid-90s suddenly allowed people to easily send data back to the servers from websites, inspiring web pioneers in psychology to create the first web surveys, experiments, and tests (like André Hahn, who owned the domain *psychologie.de* and is believed to have developed the first web-based psychological test,¹ on perceived self-efficacy); (2) big data methods are evolving from the appearance of search engines and large hubs on the web, for example, picture sharing sites or public databases, with their data being made available (e.g., a method is to use SECEs – search engine count estimates – in research, Janetzko, 2008); (3) web methods that rely on quick interaction between client and server (e.g., AJAX, social interaction online, video transmission) have become possible and increasingly better accessible with higher bandwidth and ubiquity of mobile connections and flat rates; (4) digital experience sampling methodology evolving from the invention and subsequent proliferation of smartphones, smartwatches, and other wearables. Some other technologies have been on the verge of a breakthrough for a long time and may make it in the future or not, for example, VR and augmented reality in, for example, Google glasses.

From a taxonomy valid for the first decade of web-based research Reips and Lengler (2005) empirically determined

from data for reactive studies on the web that within psychology web-based research mostly was conducted in cognitive and social psychology, followed by perception and a few studies each in personality, clinical and Internet science. A later review and survey by Krantz and Reips (2017) revealed a largely unchanged picture. Reips and Lengler also explicitly introduced the concept of *web services*, which is software that runs on a server on the Internet: “Users access it via a web browser and can use it only while they are connected to the Internet. Because the functionality of web browsers is less dependent on the operating system . . . all who access a web service are likely to see and experience almost the same interface (but see, e.g., Dillman & Bowker, 2001, for browser-related problems in Internet-based research). Web services spare the user from upgrading and updating since this is done by the web service administrators at the server. Nothing is installed on the user’s computer, saving space and time” (p. 287). Curiously, in a way with web services (now often called “apps”) and the “cloud” we have returned to the server-terminal model common in the 70s and 80s that was interrupted by a brief phase of only temporarily connected personal computers.

Importantly, when discussing characteristics of web-based research, it is important to not mistakenly attribute the advantages of computerized assessment to the web method only. Many useful functionalities such as item branching, precise timing, filtering, automatic checks of plausibility during data entry, and so on, were already introduced to experimenting during the computer revolution in the late 1960s and early 1970s (Reips & Krantz, 2010). Characteristics of computerized research are valid in web-based research also, but its true advancement comes from its structure and reach as a world-wide network.

In the following, I will present characteristics and developments of the methods of web surveys and questionnaire research (including web-based tests), web experiments, Mobile Experience Sampling, and non-reactive web research.

Web Surveys and Questionnaire Research

This type of web-based research is the most frequently used, yet it also may be the most error-prone. A general advantage of surveying as a methodology is its ease, and thus we see much ad hoc use with everything associated with quick-and-dirty approaches. Related, many errors frequently seen in web questionnaires came and come from a lack of understanding of the technology and that a direct transfer from the paper-based or computer-based formats to an Internet-based format is impossible.

¹ Still available at <https://userpage.fu-berlin.de/ahahn/frageb.htm> – the web archive also contains it at http://web.archive.org/web/*/userpage.fu-berlin.de/~ahahn/ The self-scoring script, a CGI, is not working anymore, at the time it would calculate and show the test takers their own score, their percentile and the distribution of all results.

Some of the typical pitfalls are technological. For example, Birnbaum (2004) observed a case of erroneous coding: in that web questionnaire, if a person were from India or if he or she did not respond, the coded value was the same, 99. Reips (2002a) highlights that the way HTML forms work may lead to overwriting of data due to the same names associated with form elements, for example, assigning the variable name “sex” to both an item reporting one’s own sex and an item asking about frequency of sexual activity.

In a recent meta-analysis, Daikeler and colleagues (2020) confirmed earlier findings that response rates in web surveys are lower than in other survey modes, roughly by 12%. The expected length of web questionnaires is, of course, a major factor in a respondent’s decision-making for participation and response. Galesic and Bosnjak (2009) experimentally varied both expected (10, 20, and 30 min) and actual questionnaire length. They found, “as expected, the longer the stated length, the fewer respondents started and completed the questionnaire. In addition, answers to questions positioned later in the questionnaire were faster, shorter, and more uniform than answers to questions positioned near the beginning” (p. 349).

Reips (2010) lists several typical errors that frequently happen when constructing online questionnaires:

- preselected answer options in drop-down menus, resulting either in submission of the default answer option as a chosen answer when the item is really skipped or provoking an anchoring effect,
- overlapping answer categories,
- no limitations set or announced to size of text to be entered in text fields,
- lack of options that indicate reluctance to answer (e.g., “don’t want to answer” or “no answer”), especially for sensitive items,
- all items on one run-on web page (see OIOS Technique section below),
- incorrect writing (e.g., errors in instructions or questions).

An excellent resource for many topics around web survey methodology is Callegaro and colleagues (2015). It summarizes many important learnings from the scientific literature on moving surveying to the web and supersedes many earlier attempts in scientific depth and applicability.

Despite the many benefits of web-based research, researchers and others have expressed concerns about online data collection in surveying. For the important case, in which generalization to a country population is needed, Dillman and colleagues (2010), in their editorial “Advice in surveying the general public over the Internet” for the *International Journal of Internet Science*, make the prediction “The first decade of this century in web surveying is likely

to be recalled as a time of much uncertainty on whether random samples of the general public could be surveyed effectively over the Internet. A significant proportion of households in most countries are not connected to the Internet, and within households, some members lack the Internet skills or frequency of use needed for responding to survey requests. In addition, households without Internet access differ sharply from those with it. Non-Internet households are older, have less education, and have lower incomes. Their inclusion in surveys is critical for producing acceptable population estimates, especially for public policy purposes. Web survey response rates in general public surveys are often dismal.” (p. 1). Despite the skeptical state of web surveying the general public, the authors made five recommendations that are likely to improve the situation, namely (1) use of a mixed-mode approach, (2) delivering a token cash incentive with the initial mail request, (3) using a mail follow-up to improve response rates and obtain a better representation of the general public, (4) refraining from offering people a choice of whether to respond by web or mail in initial contacts, and (5) using an experimental approach, in order to generate estimates for the meaning and sizes of various effects on response rates and non-response. Loomis and Paterson (2018) empirically investigated the challenges identified by Dillman and colleagues and found limited differences between survey modes when aggregating all results (in their case from 11 studies). Only the non-deliverable rate seemed consistently higher in online surveying than in mail surveying. Non-response rates, item non-response, and content of the results showed no or only small differences for the aggregated data. The authors interpreted differences in response rates according to mode as “random or idiosyncratic in nature, and perhaps more a matter of study population or topic than of mode” (p. 145).

A specific form of web surveying, if it considers verbal items, is *web-based psychological testing* (e.g., of personality or other individual differences constructs). A researcher who has pioneered this branch of web-based research is Tom Buchanan (e.g., Buchanan, 2000, 2001; Buchanan & Smith, 1999; Buchanan et al., 2005). His and others’ studies generally find equivalency to offline testing, with notable exceptions that indicate that any web-based test should be scrutinized with the standard reliability and validity checks during test development.

Web Experiments

Web-based experimenting has evolved since 1994 and was first presented at the SCiP Conference in Chicago 1996, with Krantz presenting the first within-subjects web experiment (Krantz et al., 1997) and Reips (1996) presenting the first between-subjects web experiment and the first

experimental laboratory on the world wide web, the Web Experimental Psychology Lab (Reips, 2001). On track of the question whether lab and web and type of participant recruitment are comparable, Germine and colleagues (2012) addressed the question of data quality across a range of cognitive and perceptual tasks. Their findings on key performance metrics demonstrate that collecting data from uncompensated, anonymous, unsupervised, self-selected participants on the web need not reduce data quality, even for demanding cognitive and perceptual experiments.

Reips points out the ultimate reason for using the web to conduct experiments:

The *fundamental asymmetry of accessibility* (Reips, 2002b, 2006): What is programmed to be accessible from any Internet-connected computer in the world will surely also be accessible in a university laboratory, but what is programmed to work in a local computer lab may not necessarily be accessible anywhere else. A laboratory experiment cannot simply be turned into a web experiment, because it may be programmed in a stand-alone programming language and lack Internet-based research methodology, but any web experiment can also be used by connecting the laboratory computer to the Internet. Consequently, it is a good strategy to design a study web-based, if possible. (2007b, pp. 375–376).

As a consequence, the many advantages of web-based experimenting, for example,

- (1) web experiments are more cost-effective in administration, time, space, and work in comparison with laboratory research,
- (2) ease of access for participants, also from different cultures and for people with rare characteristics (for accessibility see Vereenooghe, 2021),
- (3) web experiments are generally truly voluntary,
- (4) detectability of confounding with motivational aspects of experiment participation,
- (5) replicability and re-usability, as the materials are publicly accessible,

quickly led to the new method's proliferation (Birnbaum, 2004; Krantz & Reips, 2017; Musch & Reips, 2000; Reips, 2000, 2007b; Wolfe, 2017). As a consequence of these characteristics of web experiments, they frequently have been shown to collect higher quality data than laboratory experiments (Birnbaum, 2001; Buchanan & Smith, 1999; Reips, 2000). For example, in a field where all past studies from more than two decades had two digit sample sizes, spatial-numerical association, web experiments finally provided detailed results with small confidence intervals (Cipora et al., 2019). Tan and colleagues (2021) were highly successful in objectively measuring singing pitch accuracy

on the web. With moderate-to-high test-retest reliabilities (.65–.80), even across an average 4.5-year period between test and retest (!) they see high potential for large-scale web-based investigations of singing and music ability. In some areas of psychology, more than half of all studies published are now conducted online, for example, in social psychology – however, due to a certain participant recruitment method discussed further below, the widespread use is not without problems (Anderson et al., 2019).

Beyond what to researchers continues to be a new way and sometimes challenging and questionable advancement beyond laboratory experimental research, web experiments have truly revolutionized digital business. In his bestseller Seth Stephens-Davidowitz (2017) writes:

Experiments in the digital world have a huge advantage relative to experiments in the offline world. As convincing as offline randomized experiments can be, they are also resource-intensive. . . . Offline experiments can cost thousands or hundreds of thousands of dollars and take months or years to conduct. In the digital world, randomized experiments can be cheap and fast. You do not need to recruit and pay participants. Instead, you can write a line of code to randomly assign them to a group. You do not need users to fill out surveys. Instead, you can measure mouse movements and clicks. You do not need to hand-code and analyze the responses. You can build a program to automatically do that for you. You do not have to contact anybody. You do not even have to tell users that they are part of an experiment. This is the fourth power of big data: it makes randomized experiments, which can find truly causal effects, much, much easier to conduct – anytime, more or less anywhere, as long as you're online. In the era of big data, all the world's a lab. This insight quickly spread through Google and then the rest of Silicon Valley, where randomized controlled experiments have been renamed 'A/B testing.' In 2011, Google engineers ran seven thousand A/B tests. And this number is only rising. Facebook now runs a thousand A/B tests per day, which means that a small number of engineers at Facebook start more randomized, controlled experiments in a given day than the entire pharmaceutical industry starts in a year (pp. 210–211).

While early and quickly, a number of *web survey generators* appeared on the market, with "Internet-Rogator" by Heidingsfelder (1997) being one of the first, only a few *web experiment generators* are available today. For a recent listing of software for experiments (laboratory or web) and links to further resources, see the helpful Google page by Weichselgartner (2021). Figure 1 shows the newest version of

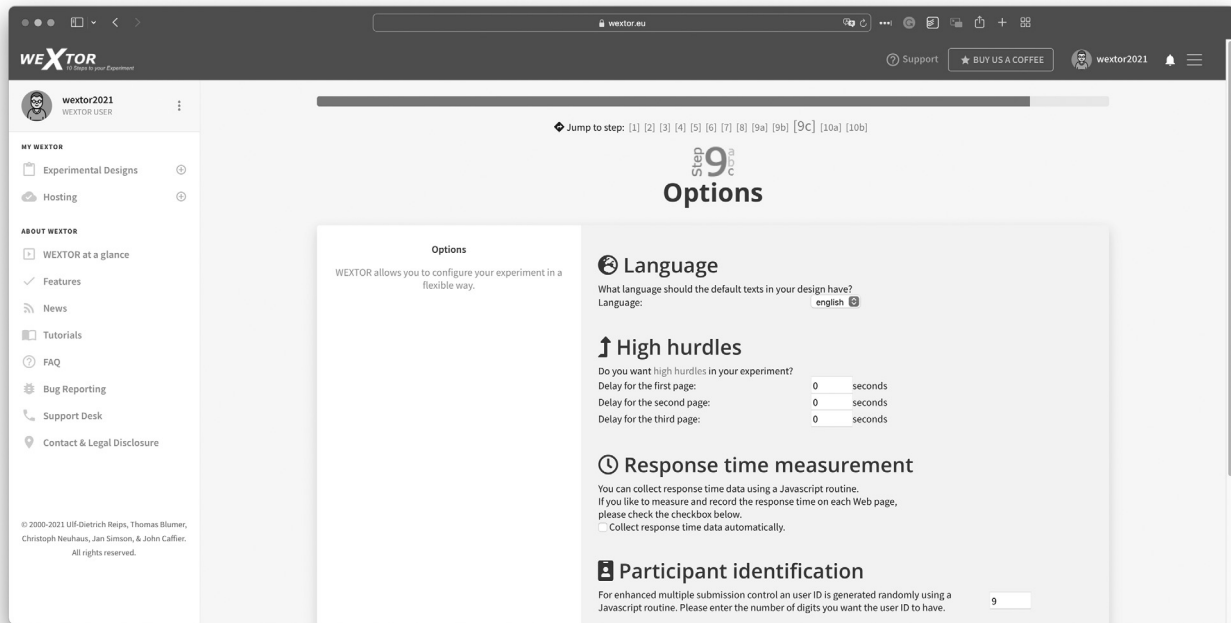


Figure 1. WEXTOR 2021, a web experiment generator available from <https://wextor.eu>. The figure shows the “good methods by design” philosophy implemented by WEXTOR’s authors, that is, methods and best practices for web-based research were implemented in a way that the experimenter is nudged toward using them.

WEXTOR (<https://wextor.eu>, also available on the iScience server at <https://iscience.eu>), one of the longest-standing web experiment generators.

Mobile Experience Sampling

This method is sometimes also described under the terms “ecological momentary assessment” or “ambulatory assessment” (Shiffman et al., 2008). It is a modern form of the diary method and has its strength in the current ubiquitous presence of smartphones, smartwatches, and other wearables in large proportions of populations in many technologically advanced societies. For instance, Stieger and Reips (2019) were able to replicate and refine past research about the dynamics of well-being fluctuations during the day (low in the morning, high in the evening) and over the course of a week (low just before the beginning of the week, highest near the end of the week) (Akay & Martinsson, 2009). The method is more accurate in capturing the frequency and intensity of experiences (Shiffman et al., 2008), but it is relatively burdensome for both participants and researchers, may lead to non-compliance, produces correlational data only, and its reliability is hard to determine. With context-sensitive experience sampling (see e.g., Klein & Reips, 2017), researchers can trigger questions based on times, events, app usage or location or combinations thereof, for example, ask for subjective well-being when someone

leaves university only on afternoons when smartphone sensors registered a practical sports class was attended. Moreover, by using the experience sampling method, different research questions can be analyzed regarding the use of mobile devices in research, for example, whether well-being can be derived from the tilt of the smartphone (Kuhlmann et al., 2016).

Akin to upbeat statements regarding the potential of computers in psychology in the 1960s and 1970s and extension of hopes for the impact of the web in psychological research in the 1990s, the proponents of mobile research declared, “Smartphones could revolutionize all fields of psychology and other behavioral sciences if we grasp their potential and develop the right research skills, psych apps, data analysis tools, and human subjects protections.” (Miller, 2012, p. 221). Critically, however, the reliance on consumer-grade devices in research comes with their limited reliability and fast-changing variance. For example, smartwatches do not agree much in their recordings of workout parameters, and their measures depend on walking speed (Fokkema et al., 2017). Furthermore, our study on smartphone sensors showed that measurements and their reliability vary by type and brand of smartphone and operating system (Kuhlmann et al., 2021). Smartphone sensing in the field will thus systematically suffer from worse measurements than possible in controlled studies in the laboratory.

While mobile experience sampling research is often not exactly web-based, the free open-source platform “Samplify” (Shevchenko et al., 2021; <https://samplify.uni.kn>) for experience sampling enables researchers to access the complete interface via a web browser and manage their present studies. It allows researchers to easily schedule, customize and send notifications linking to online surveys or experiments created in any web-based service or software (e.g., Google Forms, lab.js, SurveyMonkey, Qualtrics, WEXTOR.eu). A flexible schedule builder enables a customized notification schedule, which can be randomized for each participant. The Samplify research mobile application preserves participants’ anonymity and is freely available at the Google and Apple App Stores. Shevchenko and colleagues (2021) demonstrated via two empirical studies the app’s functionality and usability.

“Non-reactive web-based methods” refer to the use and analysis of existing databases and text or media collections on the Internet (e.g., forums, picture collections, server log files, scanned document collections, newsgroup contributions). Such data can also include geolocation, that is, information about the place that may allow analysts to identify routes and timelines. “The Internet provides an ocean of opportunities for non-reactive data collection. The sheer size of Internet corpora multiplies the specific strengths of this class of methods: Non-manipulable events can be studied in natura, facilitating the examination of rare behavioral patterns.” (Reips, 2006, p. 74). While lab-based research creates social expectations that might motivate participants to answer and perform in unusual ways, data from archival or non-reactive research will not contain biases that come from reacting to the research situation or personnel (hence “non-reactive”). Such non-reactive research is easy to do on the web.

For some archival data, there are even specific interfaces. For example, upon its initiative to scan as many of the world’s books as possible and make them available to the public, Google also created a specific search engine to search this corpus, *Google Books Ngram Viewer* (short: Google Ngram, <https://books.google.com/ngrams/>). With it, relative frequencies of words within the corpus can be analyzed per year, and thus, it is possible to create timelines that show word use over time since the year 1800. Michel and colleagues (2011) describe how this tool allows for research options unprecedented in the history of science, in data mining cultural trends as reflected in books. Younes and Reips (2019) provide guidelines for such research, for example, the use of synonyms, word inflections, and control words to assess a word’s base rate in a given year and language corpus.

Research That Could Not Be Done Without the Web

Of course, there is a lot of web-based research that cannot be done without the web because it is research that concerns the web. I will not review this research here. Instead, I will focus in this section on research that was impossible or only very difficult to do before the advancement of the web.

Sensitive or Illegal Topics

Research with people who have rare and unusual characteristics of interest used to be impossible to do or very costly and burdensome. Similarly, for research asking sensitive questions about illegal or taboo behaviors (e.g., drug dealing, Coomber, 1997; or ecstasy and cannabis consumption, Rodgers et al., 2001, 2003) or for information that the responders may be reluctant to disclose the web with its veil of anonymity has become a promising route.

With just two web-based surveys conducted via the hub of people concerned with having the rare condition *sexomnia* and their family and friends, *sleepsex.org*,² Mangan and Reips (2007) reached more than five times as many participants from the target population than in all previously published studies combined.

Large Crowdsourcing Studies

In most cases, studies that rely on thousands of participants or even more can be conducted much more efficiently and less burdensome on the web. Deshpande and colleagues (2016), for example, took a collection of thousands of so far unanalyzed hand-marked forms, with which hundreds of researchers and research assistants before the advent of modern media and the Internet had collected color names in almost 200 middle American cultures that used different yet related languages, and crowdsourced helpers on the web to categorize these entries. The results of this ongoing project promise unparalleled insights into the essentials of perception and language. Similar projects can be found on *Citizen science* websites, for example, *zooniverse.org*, where people from the general public can help researchers via the web by categorizing images from outer space or remote zones on Earth or read and classify diaries and letters by soldiers who fought in the American Civil War. Honing (2021) discusses citizen science studies in music cognition.

Large collections of entries or traces from human behavior on the Internet have become an accessible source for research. Examples include the definition of points of interest via data mining in uploaded pictures (Barras, 2009),

² Site now dysfunctional, see web archive at, for example, <https://web.archive.org/web/20080827050316/http://www.sleepsex.org/> for a snapshot of the site.

prediction of influenza outbreaks from searches (Ginsberg et al., 2009), and our own work on attributions of personality characteristics to first names accessed via Twitter mining (Reips & Garaizar, 2011). Upon the big success of its search engine that became available on the web in 1997, Google has created freely available interfaces to their search data. These include Google Trends, Google Insights, Google Correlations (now disconnected, like the more specific services Google Flu Trends and Google Eurovision). These services can be used in psychological research, sometimes with profound results achieved within hours, but of course, there are also limitations, ethical issues, and scientific principles that are sometimes at odds with characteristics of big data services provided by companies.

Other big data research is, of course, possible without the web, and all of it comes with a number of limitations. For example, Back et al. (2011) showed for a big data study on emotions in digital messaging how a term from an automatically generated message that was frequently sent by the system unknowingly completely messed up the results and the conclusions that were first drawn from it because it contained the word “critical,” which was interpreted as a marker for anger. I expect us to see many more cases with such artifacts in the future, along with the proliferation of large-scale studies.

In principle, for many studies conducted in psychology, no large participant crowds are needed. Even though power was notoriously low in pre-web psychology research and promises to become more adequate as web-based research is more easily scalable, overpowering studies is just as bad (e.g., Faul et al., 2009; Reips 1997). There has been a bit of bragging in some articles relying on web-based data collection in terms of how many participants had been reached – “millions and millions” –, some of which then report pitiful effect sizes, but really for methodological and ethical reasons to request such large numbers of people to devote their time usually is not needed and thus may fall back on the authors of such articles as questionable practices.

Methods in Web-Based Research

While most methods in principle have been transferred to the web, many needed to be modified and adapted to the online format, so often, new challenges arose. Psychological tests, for example, should not simply be transferred from paper-and-pencil to the computerized format and then be handled similarly as usual on the web. In principle, they need to be evaluated and validated as web-based instruments (Buchanan, 2001; Buchanan & Smith, 1999). For reasons of space, the selection of methods presented here will be limited to those that have large effects and decent impact.

Design: One-Item-One-Screen (OIOS)

When designing a web-based instrument that consists of a number of items, a researcher has to decide how many items are going to be presented on one screen – or whether this may vary by device or participant. The OIOS or one-item-one-screen design has several advantages, namely “Context effects (interference between items) are reduced (Reips, 2002a), meaningful response times and drop out can be measured and used as dependent variables, and the frequency of interferences in hidden formatting is vastly reduced” (Reips, 2010, pp. 33–34). The design thus is routinely used in research studies, for example, in urban planning (Roth, 2006), where even a variant was developed, the “Two-item-one-screen design” (Lindquist et al., 2016).

OIOS has been proposed as a recommended strategy in a general framework for technology-based assessments by Kroehne and Goldhammer (2018). They write, “Item-level response times from questionnaire items (e.g., Likert type) are an interesting source of information above and beyond item responses.” (p. 543) and go on to criticize current implementations of the Programme for International Student Assessment (PISA) and other large-scale assessments as missing out on such opportunities in web-based assessments.

Including other data (para-data, meta-data) apart from the response itself, especially behavioral data that indicate timing, navigation, and switching of answers (Stieger & Reips, 2010) and thus can be important in identifying issues during test and questionnaire construction as well as diagnostic indicators beyond content responses, likely will become much more important in future surveying. Issues with adaptivity (e.g., in responsive design) and ethical application will continue to be discussed (see e.g., Hilbig & Thielmann, 2021).

Seriousness Check

The seriousness check is a technique that can be used in all reactive types of web-based research to significantly improve data quality (Aust et al., 2013; Bayram, 2018; Musch & Klauer, 2002; Reips, 2000, 2009; Verbree et al. 2020). Revilla (2016) found only limited evidence supporting the technique, but hers was an underpowered study with online panelists and a wording geared towards commitment that resulted in very few “non-serious” participants.

Nowadays, online studies are accessible to a large diversity of participants. However, many people just click through a questionnaire out of curiosity, rather than providing well-thought-out answers (Reips, 2009). This poses a serious threat to the validity of online research (Oppenheimer et al., 2009; Reips, 2002b, 2009). The seriousness

How do you intend to browse the Web pages of this study?

I would like to seriously participate now.

I would like to look at the pages only.

Please click on the following button to start with the experiment.

Figure 2. Seriousness check technique (adapted from Reips, 2009).

check addresses this problem (Figure 2): In this approach, the respondents are asked about the seriousness of their participation or for a probability estimate that they will complete the entire study or experiment (Musch & Klauer, 2002). Thus, by using the seriousness check, irrelevant data entries can be easily identified and be excluded from the data analysis. To provide a rough idea of how large the seriousness check's effect can be: It routinely was observed that of those answering "I would like to look at the pages only" around 75% will drop, while of those answering "I would like to seriously participate now" only ca. 10–15% will drop during the study. Overall, about 30–50% of visitors will fail the seriousness check, that is, answer "I would like to look at the pages only" (Reips, 2009).

Figure 2 depicts a possible seriousness check proposed by Reips (2009). The participants are asked whether or not they want to participate seriously in the experiment ("I would like to seriously participate now."/"I would like to look at the pages only."). Seriousness checks can be implemented before (e.g., Bernecker & Job, 2011) and after (Aust et al., 2013) participation in the study. Reips (2002b, 2008, 2009) has argued to conduct seriousness checks on the first page of the experiment because this is the best predictor of dropout rates and thus a measure of motivation. Additionally, conducting a seriousness check before the completion of the study can reduce the dropout rates (Reips, 2002a). The participant's answer to the check question serves as self-commitment, as predicted by dissonance theory (Frick et al., 2001). Bayram (2018) experimentally showed that emphasizing the seriousness increased the degree of information participants accessed and their time they spent on the study. The technique has been shown to predict dropout and control for motivational confounding (Aust et al., 2013; Bayram, 2018; Musch & Klauer, 2002; Reips, 2002b, 2008, 2009). Some tools for web-based research, for example, WEXTOR (Reips & Neuhaus, 2002; <https://wextor.eu>), implement the seriousness check by default.

In a study with 5,077 participants representative of the Dutch population in education, gender, and age over 15 years, Verbree and colleagues (2020) recently confirmed that self-reported seriousness and motivation significantly

predict multiple data quality indicators. In preparing a study that includes a seriousness check, it may be important to know that their results showed that self-reported seriousness varies with demography.

Instructional Manipulation Check (IMC) and Other Attention Checks

The IMC (Oppenheimer et al., 2009) was created with the same intention in mind as the seriousness check, to identify and avoid data from participants who are not as attentive in online studies as is required to gather quality data. In the case of the IMC, the focus is on attention during instructions. Of course, if participants do not properly attend to instructions, it is unlikely they will provide valid data during any subsequent task. Hence, screening them out at the beginning makes sense. However, there are several other attention checks that were designed to verify attention during tasks – we will also briefly look at those further below.

In the IMC, a request for unusual navigation is hidden within the instructions, for example, to *not* click on the submit button at the end of the page, but rather on the title of the text or a small blue dot. Only those who read the instruction carefully and comply will follow the navigation, and only their data will be analyzed later.

Another attention check, the Cognitive Reflection Test (CRT; Frederick, 2005) is a frequently used measure of cognitive vs. intuitive reflection, sometimes also discussed as a measure of numerical ability. A typical task it includes is the bat-and-the-ball problem: "A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost? ...cents". The intuitive answer most inattentive participants go for is "10 cents", but the correct answer is "5 cents". With the widespread use of the task as a nice riddle and as an attention check in web-based research, many people have become familiar with this task. Indeed, Stieger and Reips (2016) found in a large study that 44% of the more than 2,200 participants were familiar with the task or a similar task and scored substantially higher on the test (Cohen's $d = 0.41$). They also found that familiarity varies with sociodemographics. Web researchers should, therefore, better use lesser-known attention check items or the methods discussed above.

Multiple Site Entry Technique

The multiple site entry technique I first proposed in the mid-1990s (Reips, 1997, 2000) is a strategy used in web-based research to target different samples via different recruitment sites and compare their data. The method can be used in behavioral and social research to assess

the presence and impact of self-selection effects. Self-selection effects can be considered a major challenge in social science research. With the invention of online research in the 1990s, the multiple site entry technique became possible because the recruitment of participants via different links (URLs) is very easy to implement. It can be assumed that there is no or very limited self-selection bias if the data sets coming from different recruitment sites do not differ systematically (Reips, 2000). This implies that the results from studies that use the multiple site entry technique and find no sample differences indicate high generalizability.

Implementing the multiple site entry technique works as follows: Several links to the study are placed on different websites, in Internet forums, social media platforms, or offline media that are likely to attract different types of participants or are mailed out to different mailing lists. In order to identify the recruitment sources, the published URLs contain source-identifying information, and the HTTP protocol is analyzed by different referrer information (Schmidt, 2000). This means a unique string of characters is appended to the URL for each recruitment source, for example, "...index.html?source=studentlist" for a study announcement mailed to a list of students. The data file will have a column ("source" in the example) containing an entry of the referring source for each participant ("studentlist" in the example). The collected datasets can then be compared for differences in results and differences in relative degree of appeal (measured via dropout), demographic data, central results, and data quality (Reips, 2002b). Figure 3 illustrates the multiple-site entry technique.

Several studies have shown that the multiple site entry technique is useful for determining the presence and impact of self-selection in web-based research (Reips, 2000, 2002a; Roth, 2006). Now, in the spring of 2021, Google Scholar shows more than 75 publications that mention the technique. It has been used in memory (Kristo et al., 2009), personality (Bluemke & Zumbach, 2012; Buchanan et al., 2005; Trapnell & Campbell, 1999), trauma surveys (Hiskey & Troop, 2002), cross-cultural music-listening (Boer & Fischer, 2011), landscape research (Roth, 2006), criminological psychology (Buchanan & Whitty, 2014), and political psychology (Kus et al., 2014) and it has entered the methodological discussion in the fields of experimental survey research (Holbrook & Lavrakas, 2019), sex research (Mustanski, 2001), and health science (Whitehead, 2007). Rodgers and colleagues (2003) used the multiple site entry technique to detect biased responses to their web questionnaire on drug use (subsequently validated by finding discussions of their research in forums on that particular recruitment website). The multiple site entry technique thus helps to detect potential sampling problems, which in turn ensures the quality of the data

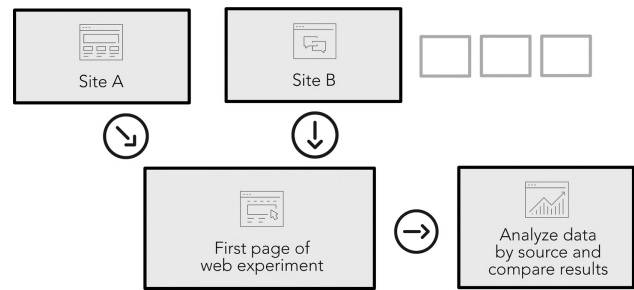


Figure 3. Illustration of the multiple site entry technique (adapted from Reips, 2009).

collection over the Internet (Dillman et al., 2010; Reips, 2002b). Therefore, the generalizability of the research results can be estimated when using the multiple-site entry technique (Reips, 2002a).

Subsampling Technique

Data quality may vary with a number of factors (e.g., whether personal information is requested at the beginning or end of a study, Frick et al., 2001, or whether participants are not allowed to leave any items unanswered and, therefore, show psychological reactance, Reips, 2002b). Subsampling analysis is a verification procedure. For a random sample drawn from all data submitted (e.g., from 50 out of 1,500 participants), every possible measure is taken to verify the responses, resulting in an estimate for the whole dataset. This technique can help estimate the prevalence of wrong answers by checking verifiable responses (e.g., age, gender, occupation) both by specific item and via the aggregate for any item. Ray and colleagues (2010) consider this technique as one possible option to better verify age and thus protect children on the Internet.

Buchanan and colleagues (2005, p. 120) noted that such "ways of estimating the degree of potential data contamination" along with other control procedures would need to be developed and researched more in the future. While a lot has happened in data analysis, for example, in the application of Benford's law (Benford, 1938) to survey data (Judge & Schechter, 2009), the advances in design and procedure of web-based research (or all types of research, for that matter) need further research and development.

Warm-Up Technique

The warm-up technique in Internet-based experimenting, first proposed by Reips (2001), is a method that can be used to avoid dropout during the experimental phase of a study to maximize the quality of the data. This technique is based

on the finding that most dropout will occur at the beginning of an online study (Reips, 2002c). Therefore, participants are presented with tasks and materials similar to the experimental materials before the actual experimental manipulation is introduced.

By using this technique, one can counter three main problems in web-based research: Firstly, the dropout during the actual experiment will be lower (Reips, 2001). Secondly, the dropout cannot be attributed to the experimental manipulation (Reips, 2002a, 2002b). Thirdly, only highly committed participants will stay in the experiment, and thus, the quality of the collected data is improved. Reips and colleagues (2001) showed that by using the warm-up technique the dropout during the actual experiment was extremely low (< 2%). In comparison, the average dropout rate in web-based research is much higher, in a review summarizing previous web-based experimental research, Musch and Reips (2000) found it to average at 34%.

The technique was used in personality and health behavior research (Hagger-Johnson & Whiteman, 2007), motivation (Bernecker & Job, 2011), landscape perception (Lindquist et al., 2016; Roth, 2006), gender research (Fleischmann et al., 2016), polar research (Summerson & Bishop, 2012), and has been implemented in software to conduct web experiments (Naumann et al., 2007). It entered the methodological discourse in Poland (Siuda, 2009) and China (Wang et al., 2015), where it is called 热身法.

Pitfalls, Best Practices

In 2010, I wrote, “At the core of many of the more important methodological problems with design and formatting in Internet-based research are interactions between psychological processes in Internet use and the widely varying technical context.” (p. 32). The interaction between psychology and technology can lead to advances, but foremost in research, many pitfalls were discovered that subsequently led to recommendations for best practices, which I both will introduce in this section.

Measurement

Psychology has a long history of finding strategies to measure behaviors, mental processes, attitudes, emotions, self-reported inner states, or other constructs. Pronk and colleagues (2020), following-up on various others (Garaizar et al., 2014; Plant, 2016; Reimers & Stewart, 2015; Reips, 2007a; Schmidt, 2001; van Steenbergen & Bocanegra, 2016), investigated timing accuracy of web applications, here with a comparative focus on touchscreen and keyboard devices. Their results confirm what theoretically

was expected from the technical structure and limitations of the web (Reips, 1997, 2000):

... very accurate stimulus timing and moderately accurate RT measurements could be achieved on both touchscreen and keyboard devices, though RTs were consistently overestimated. In uncontrolled circumstances, such as researchers may encounter online, stimulus presentation may be less accurate. ... Differences in RT overestimation between devices might not substantially affect the reliability with which group differences can be found, but they may affect reliability for individual differences (p. 1371).

An example of how the combination of computerized measurement and large sample sizes achievable on the web provided a shift in measurement accuracy that was long needed but overlooked in light of traditions and measurement burden is the switch from Likert-type rating scales to visual analogue scales. The latter has become much easier to administer via the web than on paper (where distances have to be measured with a ruler) and on computers (where software often lacks that type of scale as an option). Figure 4 shows an example of both types of scales. Note the visual analogue scale turns 100 this year (Hayes & Patterson, 1921).

As Reips and Funke (2008) point out, as a result of their experiment, data collected with web-based visual analogue scales provide better measurement than Likert-type scales and offer more options for statistical analysis. Importantly, visual analogue scales are not to be confused with slider scales, for which we know its handle causes potential problems because its default position may cause anchoring effects or erroneous ratings (Funke, 2016). Slider scales may also cause more problems for the less educated (Funke et al., 2011).

Dropout and Other Nonresponse

Dropout is more prevalent in web-based research and may have detrimental effects (Reips, 2002b; Zhou & Fishbach, 2016). Zhou and Fishbach describe how unattended selective dropout can lead to surprising yet false research conclusions. However, avoiding dropout (e.g., via high hurdle and warm-up techniques, and low tech principle) or controlling it (e.g., via seriousness check) are not the only strategies to deal with the higher prevalence of dropout on the web. Rather, dropout can be used as a dependent variable (Reips, 2002a). Bosnjak (2001) describes seven types of non-response behavior that can be observed in surveys, these all are good measures in web-based research.

In an early and widely cited study, Frick and colleagues (2001) experimentally manipulated the influence of announcing an incentive (or not), anonymity, and

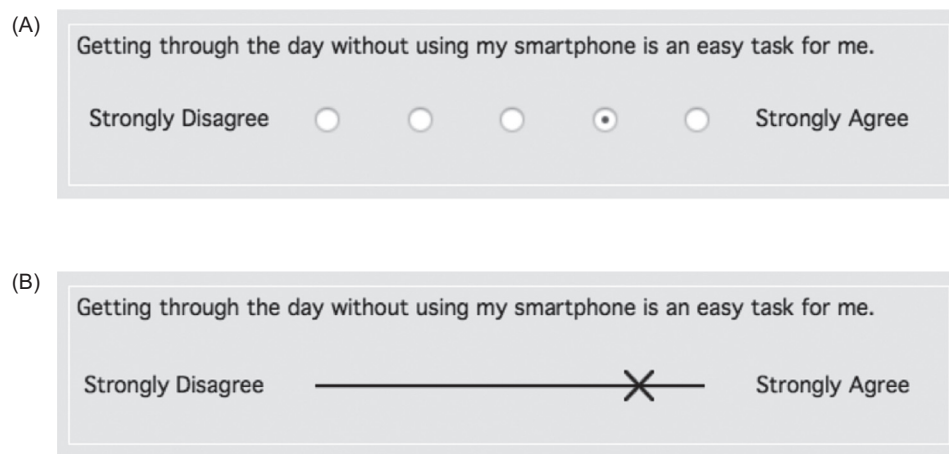


Figure 4. Web-based Likert type scale (A) and Visual Analogue scale (B).

placement of demographic questions (beginning vs. end) on dropout. Incentive announcement and position of demographic questions showed large main effects on dropout, it ranged from 5.7% in the incentive known and demographic questions at the beginning condition to 21.9% in the incentive unknown and demographic questions at the end condition. They also found a strong effect (> 100 min difference in reported TV consumption per week) of the order of questions – asking both for time devoted to charity work and TV consumption had created a context that evoked socially desirable responses. Birnbaum (2021) also discusses dropout and reflects on early discussions and findings by those who adopted the web for research and developed the associated methodology.

DropR (<http://dropr.eu>) is ShinyApp and R software that we created to meet the increased need to calculate dropout rates because dropout is much more frequent in web-based research than in laboratory research. In the analysis and reporting of web experiments, the commonly high dropout makes it necessary to provide an analysis and often also visualize dropout by condition. DropR supports web researchers in both providing manuscript-ready Figures specifically designed for accessibility (see Vereenooghe, 2021) and all major survival and common dropout analyses on the fly, including Kaplan-Meier, chi-square, odds ratio, and rho family tests. Visual inspection allows for quick detection of critical differences in dropout. DropR is Open Source software available from Github (Reips & Bannert, 2015).

Dropout is particularly useful in detecting *motivational confounding* in experiments (Reips, 2000, 2002b). Whenever conditions differ in motivational aspects, there is the danger that this confound may explain any between-condition findings. On the web this would likely show in differential dropout rates. On the contrary, “because there is usually very minor dropout in offline experiments, the lab-

oratory setting is not as sensitive as the Internet setting in regard to detecting motivational confounding.” (Reips, 2009). In a secondary analysis of two studies, Shapiro-Luft and Cappella (2013) confirm that motivational confounding can be detected in web-based studies with video content. It has been of consideration in conducting real-time multiplayer experiments on the web (Hawkins, 2015) and is now routinely used as an argument to conduct studies on the web rather than in the lab (e.g., Lithopoulos et al., 2020; Sheinbaum et al., 2013).

Interaction of Technology and Psychology

A chief issue in web-based research is interactions between technology and human factors, such as the type of device and personality of user. Which device someone uses is by no way randomly determined, it depends on one’s preferences – either directly, for example, because one may be an “early adopter” and likes to buy and use the newest technology or indirectly because one’s personality and demographics drive one to follow a certain education and end up in a certain profession where some type of technology is more common than in other professions. Buchanan and Reips (2001) analyzed responses of 2,148 participants to a web-based Five-Factor personality inventory and compared demographic items for users of different computing platforms. The responses of participants whose web browsers were Javascript-enabled were also compared with those whose web browsers were not. Macintosh users were significantly more “Open to Experience” than were PC users, and users using Javascript-enabled browsers had significantly lower education levels.

For research, an immediate consequence is to expect larger direct and indirect self-selection and coverage biases. If a web-based study relies on certain specific technologies, it will not reach every person with the same probability. For

theory-guided and experimental basic research, this is less of an issue, but it may be for any research that tries to generalize from samples to populations. Technical and situational variance in itself in the presence of an effect strengthens the case for its generalizability (Reips, 1997, 2000), as it diminishes the probability that an effect is an artifact resulting from a specific technological setup in the laboratory. Krantz (2021) further notes that “technical variance also suggests a potential for modification of the theoretical understanding of the phenomenon” (p. 233) in web-based research and shows how this can be done with the famous illusion named after the founder of this journal, Ebbinghaus.

Recruitment

Once one understands the willingness of people to participate in web-based research (Bosnjak & Batinic, 2002), the question arises where to find them. Participants for web-based research in psychology can be found via various types of sources beyond any that also work for laboratory-based research: Mailing lists, forums/newsgroups, online panels, social media (Facebook, Twitter, Instagram, Snapchat, Tuenti, ...), frequented websites (e.g., for news), special target websites (e.g., by genealogists), Google ads. Further, there are dedicated websites like “Psychological research on the net” (<https://psych.hanover.edu/research/exponnet.html>) by John Krantz or the “web experiment list” at <https://wexlist.uni-konstanz.de>

Within just a few years, it has become common to recruit workers as participants for “mini jobs” on crowdsourcing platforms like Clickworker, Prolific Academic, Cloudflare, or Amazon Mechanical Turk (AMT). Anderson and colleagues (2019) show social and personality psychology as an example of how dominant recruitment via AMT became a recruitment platform just within a few years. However, this proliferation of its use among researchers stands in stark contrast with much criticism about data quality from AMT workers (“MTurkers”) and the site’s limitations. Reips (in press) writes,

Workers respond to be paid, whereas other research participants respond to help with research. A second reason why MTurkers provide lower quality data may be tied to the forums they have established where jobs are discussed, including online studies. It may well be that rumors and experiences shared in these forums lead to decreased data quality. A third reason is artificial MTurkers that have appeared on the site, these are computer scripts or “bots”, not humans (Dreyfuss, 2018), ironically replacing the “hidden human” in the machine with machines. Stewart and colleagues (2015) calculated that with more and more laboratories in the Behavioral sciences moving to

MTurk the total size of the actual participant pool for all studies approaches just 7,300 people rather than the hundreds of thousands in the past.

Ironically, a service that was developed to employ people who appear to work as a machine because the machine (computer) cannot do the task as well is now going down the drain because human beings have programmed scripts that pose as human workers.

Conclusion

Web-based research has enabled psychologists to explore new topics and do their research with previously impossible options to reach large heterogeneous samples and people with rare characteristics, run studies easily in several samples and cultures simultaneously (see Subsampling Technique section), and go deeper into multiple fine-grained measurements that may bring a revival of behavior instead of self-report and other measures in psychology.

At the same time, technological factors became more dominant in the ways psychological research is conducted, various dangers include (1) the dependency on non-scientific agents like big player companies who provide only selected data access to “embedded scientists”, (2) lack of reproducibility because of the many hardware and dynamic software factors involved and the quickly changing technology, and (3) the distance between the researcher and participants, who may not even be human but vague “agents” provided by commercial recruitment services.

Web methodologists are, of course, trying to keep up with the fast development and provide multiple solutions to the challenges posed by the web as a route for research. For those of us who actively have experienced research before the web revolution, it will be an important task to describe our insights from comparing pre-web with web-based research and teach a new generation of researchers in psychology. We will be the only generation to have witnessed and experienced the change.

References

- Akay, A., & Martinsson, P. (2009). *Sundays are blue: Aren't they? The day-of-the-week effect on subjective well-being and socioeconomic status*. (No. 4563; IZA Discussion Paper). <http://hdl.handle.net/10419/36331>
- Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, 45(6), 842–850. <https://doi.org/10.1177/0146167218798821>
- Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, 53, 1407–1425. <https://doi.org/10.3758/s13428-020-01501-5>

- Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, 45(2), 527–535. <https://doi.org/10.3758/s13428-012-0265-2>
- Back, M. D., Küfner, A. C. P., & Egloff, B. (2011). “Automatic or the people?” Anger on September 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science*, 22(6), 837–838.
- Barras, G. (2009). Gallery: Flickr users make accidental maps. *New Scientist*. <http://www.newscientist.com/article/dn17017-gallery-flickr-user-traces-make-accidental-maps.html>
- Bayram, A. B. (2018). Serious subjects: A test of the seriousness technique to increase participant motivation in political science experiments. *Research and Politics*, 5(2). <https://doi.org/10.1177/2053168018767453>
- Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4), 551–572.
- Bernecker, K., & Job, V. (2011). Assessing implicit motives with an online version of the picture story exercise. *Motivation and Emotion*, 35(3), 251–266. <https://doi.org/10.1007/s11031-010-9175-8>
- Birnbaum, M. H. (2001). A web-based program of research on decision making. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 23–55). Pabst Science.
- Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, 55, 803–832. <https://doi.org/10.1146/annurev.psych.55.090902.141601>
- Birnbaum, M. H. (2021). Advanced training in web-based psychology research: Trends and future directions. *Zeitschrift für Psychologie*, 229(4), 260–265. <https://doi.org/10.1027/2151-2604/a000473>
- Bluemke, M., & Zumbach, J. (2012). Assessing aggressiveness via reaction times online. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 6(1), Article 5. <https://doi.org/10.5817/CP2012-1-5>
- Boer, D., & Fischer, R. (2011). Towards a holistic model of functions of music listening across cultures: A culturally decentred qualitative approach. *Psychology of Music*, 40(2), 179–200. <https://doi.org/10.1177/0305735610381885>
- Bosnjak, M. (2001). Participation in non-restricted web surveys: A typology and explanatory model for item-nonresponse. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 193–207). Pabst Science.
- Bosnjak, M., & Batinic, B. (2002). Understanding the willingness to participate in online-surveys. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 81–92). Hogrefe & Huber.
- Buchanan, T. (2000). Potential of the Internet for personality research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 121–140). Academic Press.
- Buchanan, T. (2001). Online personality assessment. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 57–74). Pabst Science.
- Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the Internet. *European Journal of Psychological Assessment*, 21(2), 115–127. <https://doi.org/10.1027/1015-5759.18.1.115>
- Buchanan, T., & Reips, U.-D. (2001). Platform-dependent biases in online research: Do Mac users really think different? In K. J. Jonas, P. Breuer, B. Schauenburg, & M. Boos (Eds.), *Perspectives on Internet research: Concepts and methods (Proceedings of the 4th German Online Research Conference (GOR), May 17–18, Göttingen, Germany)* University of Göttingen. https://www.uni-konstanz.de/iscience/reips/pubs/papers/Buchanan_Reips2001.pdf
- Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the world wide web. *British Journal of Psychology*, 90(1), 125–144. <https://doi.org/10.1348/000712699161189>
- Buchanan, T., & Whitty, M. T. (2014). The online dating romance scam: Causes and consequences of victimhood. *Psychology, Crime & Law*, 20(3), 261–283. <https://doi.org/10.1080/1068316X.2013.772180>
- Callegaro, M., Manfreda, K. L., & Vehovar, V. (2015). *Web survey methodology*. Sage.
- Cipora, K., Soltanlou, M., Reips, U.-D., & Nuerck, H.-C. (2019). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behavior Research Methods*, 51, 1676–1692. <https://doi.org/10.3758/s13428-019-01213-5>
- Coomber, R. (1997). Using the Internet for survey research. *Sociological Research Online*, 2. <http://www.socresonline.org.uk/2/2/2.html>
- Daikeler, J., Bosnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: An updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, 8(3), 513–539. <https://doi.org/10.1093/jssam/smz008>
- Deshpande, P. S., Tauber, S., Chang, S. M., Gago, S., & Jameson, K. A. (2016). Digitizing a large corpus of handwritten documents using crowdsourcing and Cultural Consensus Theory. *International Journal of Internet Science*, 11(1), 8–32.
- Dillman, D. A., Reips, U.-D., & Matzat, U. (2010). Advice in surveying the general public over the Internet. *International Journal of Internet Science*, 5(1), 1–4.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160.
- Fleischmann, A., Sieverding, M., Hespeneheide, U., Weiß, M., & Koch, S. C. (2016). See feminine – think incompetent? The effects of a feminine outfit on the evaluation of women’s computer competence. *Computers & Education*, 95, 63–74. <https://doi.org/10.1016/j.compedu.2015.12.007>
- Fokkema, T., Kooiman, T. J., Krijnen, W. P., van der Schans, C. P., & de Groot, M. (2017). Reliability and validity of ten consumer activity trackers depend on walking speed. *Medicine and Science in Sports and Exercise*, 49(4), 793–800. <https://doi.org/10.1249/MSS.0000000000001146>
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Frick, A., Bächtiger, M. T., & Reips, U.-D. (2001). Financial incentives, personal information and drop-out in online studies. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 209–219). Pabst Science.
- Funke, F. (2016). A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review*, 34(2), 244–254. <https://doi.org/10.1177/0894439315575477>
- Funke, F., Reips, U.-D., & Thomas, R. K. (2011). Sliders for the smart: Type of rating scale on the web interacts with educational level. *Social Science Computer Review*, 29, 221–231.
- Garaizar, P., & Reips, U.-D. (2019). Best practices: Two web browser-based methods for stimulus presentation in behavioral experiments with high resolution timing requirements. *Behavior Research Methods*, 51, 1441–1453. <https://doi.org/10.3758/s13428-018-1126-4>
- Garaizar, P., Vadillo, M. A., & López-de-Ipiña, D. (2014). Presentation accuracy of the web revisited: Animation methods in the HTML5 era. *PLoS One*, 9, Article e109812. <https://doi.org/10.1371/journal.pone.0109812>

- Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19(5), 847–857. <https://doi.org/10.3758/s13423-012-0296-9>
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012–1015.
- Hagger-Johnson, G. E., & Whiteman, M. C. (2007). Conscientiousness facets and health behaviors: A latent variable modeling approach. *Personality and Individual Differences*, 43(5), 1235–1245. <https://doi.org/10.1016/j.paid.2007.03.014>
- Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, 47(4), 966–976. <https://doi.org/10.3758/s13428-014-0515-6>
- Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*, 18, 98–99.
- Heidingsfelder, M. (1997). *Der Internet-Rogator* [The Internet-Rogator]. Software demonstration at first German Online Research (GOR) Conference, Cologne. https://www.gor.de/archive/gor97/fr_13.html
- Hilbig, B. E., & Thielmann, I. (2021). On the (mis)use of deception in web-based research: Challenges and recommendations. *Zeitschrift für Psychologie*, 229(4), 225–229. <https://doi.org/10.1027/2151-2604/a000466>
- Hiskey, S., & Troop, N. A. (2002). Online longitudinal survey research: Viability and participation. *Social Science Computer Review*, 20(3), 250–259.
- Holbrook, A., & Lavrakas, P. J. (2019). Vignette experiments in surveys. In P. Lavrakas, M. Traugott, C. Kennedy, A. Holbrook, E. de Leeuw, & B. West (Eds.), *Experimental methods in survey research* (pp. 369–370). Wiley. <https://doi.org/10.1002/9781119083771.part8>
- Honing, H. (2021). Lured into listening: Engaging games as an alternative to reward-based crowdsourcing in music research. *Zeitschrift für Psychologie*, 229(4), 266–268. <https://doi.org/10.1027/2151-2604/a000474>
- Janetzko, D. (2008). Objectivity, reliability, and validity of search engine count estimates. *International Journal of Internet Science*, 3, 7–33.
- Judge, G., & Schechter, L. (2009). Detecting problems in survey data using Benford's Law. *Journal of Human Resources*, 44(1), 1–24. <https://doi.org/10.1353/jhr.2009.0010>
- Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of quality of answers in a web survey. *Public Opinion Quarterly*, 73(2), 349–360. <https://doi.org/10.1093/poq/nfp031>
- Klein, B., & Reips, U.-D. (2017). Innovative social location-aware services for mobile phones. In A. Quan-Haase & L. Sloan (Eds.), *Handbook of social media research methods* (pp. 421–438). Sage.
- Krantz, J. H. (2021). Ebbinghaus illusion: Relative size as a possible invariant under technically varied conditions? *Zeitschrift für Psychologie*, 229(4), 230–235. <https://doi.org/10.1027/2151-2604/a000467>
- Krantz, J., & Reips, U.-D. (2017). The state of web-based research: A survey and call for inclusion in curricula. *Behavior Research Methods*, 49(5), 1621–1629. <https://doi.org/10.3758/s13428-017-0882-x>
- Krantz, J. H., Ballard, J., & Scher, J. (1997). Comparing the results of laboratory and world-wide web samples on the determinants of female attractiveness. *Behavior Research Methods, Instruments, & Computers*, 29, 264–269. <https://doi.org/10.3758/BF03204824>
- Kristo, G., Janssen, S. M. J., & Murre, J. M. J. (2009). Retention of autobiographical memories: An Internet-based diary study. *Memory*, 17(8), 816–829. <https://doi.org/10.1080/09658210903143841>
- Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, 45(2), 527–563. <https://doi.org/10.1007/s41237-018-0063-y>
- Kuhlmann, T., Garaizar, P., & Reips, U.-D. (2021). Smartphone sensor accuracy varies from device to device: The case of spatial orientation. *Behavior Research Methods*, 53, 22–33. <https://doi.org/10.3758/s13428-020-01404-5>
- Kuhlmann, T., Reips, U.-D., & Stieger, S. (2016, November 11). Smartphone tilt as a measure of well-being? Results from a longitudinal smartphone app study. In U.-D. Reips (Ed.), *20 years of Internet-based research at SciP: Surviving concepts, new methodologies* [Symposium]. 46th Society for Computers in Psychology (SciP) Conference, Boston, MA, USA.
- Kus, L., Ward, C., & Liu, J. (2014). Interethnic factors as predictors of the subjective well-being of minority individuals in a context of recent societal changes. *Political Psychology*, 35(5), 703–719. <https://doi.org/10.1111/pops.12038>
- Lindquist, M., Lange, E., & Kang, J. (2016). From 3D landscape visualization to environmental simulation: The contribution of sound to the perception of virtual environments. *Landscape and Urban Planning*, 148, 216–231. <https://doi.org/10.1016/j.landurbplan.2015.12.017>
- Lithopoulos, A., Grant, S. J., Williams, D. M., & Rhodes, R. E. (2020). Experimental comparison of physical activity self-efficacy measurement: Do vignettes reduce motivational confounding? *Psychology of Sport and Exercise*, 47, Article 101642. <https://doi.org/10.1016/j.psychsport.2019.101642>
- Loomis, D. K., & Paterson, S. (2018). A comparison of data collection methods: Mail versus online surveys. *Journal of Leisure Research*, 49(2), 133–149. <https://doi.org/10.1080/00222216.2018.1494418>
- Mangan, M. A., & Reips, U.-D. (2007). Sleep, sex, and the web: Surveying the difficult-to-reach clinical population suffering from sexomnia. *Behavior Research Methods*, 39, 233–236.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Lieberman Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182.
- Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, 7(3), 221–237.
- Musch, J., & Klauer, K. C. (2002). Psychological experimenting on the world wide web: Investigating content effects in syllogistic reasoning. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 181–212). Hogrefe & Huber.
- Musch, J., & Reips, U.-D. (2000). A brief history of web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 61–88). Academic Press.
- Mustanski, B. S. (2001). Getting wired: Exploiting the Internet for the collection of valid sexuality data. *Journal of Sex Research*, 38(4), 292–301.
- Naumann, A., Brunstein, A., & Krems, J. F. (2007). DEWEX: A system for designing and conducting web-based experiments. *Behavior Research Methods*, 39(2), 248–258. <https://doi.org/10.3758/BF03193155>
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872. <https://doi.org/10.1016/j.jesp.2009.03.009>

- Plant, R. R. (2016). A reminder on millisecond timing accuracy and potential replication failure in computer-based psychology experiments: An open letter. *Behavior Research Methods*, 48, 408–411. <https://doi.org/10.3758/s13428-015-0577-0>
- Pronk, T., Wiers, R. W., Molenkamp, B., & Murre, J. (2020). Mental chronometry in the pocket? Timing accuracy of web applications on touchscreen and keyboard devices. *Behavior Research Methods*, 52(3), 1371–1382. <https://doi.org/10.3758/s13428-019-01321-2>
- Ratcliffe, J., Soave, F., Bryan-Kinns, N., Tokarchuk, L., & Farkhatdinov, I. (2021, May 8–13). *Extended reality (XR) remote research: A survey of drawbacks and opportunities*. Paper presented at the CHI Conference on Human Factors in Computing Systems (CHI '21), Yokohama, Japan. <https://doi.org/10.1145/3411764.3445170>
- Ray, J. V., Kimonis, E. R., & Donoghue, C. (2010). Legal, ethical, and methodological considerations in the Internet-based study of child pornography offenders. *Behavioral Sciences and the Law*, 28(1), 84–105. <https://doi.org/10.1002/bsl.906>
- Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript web experiments. *Behavior Research Methods*, 47, 309–327. <https://doi.org/10.3758/s13428-014-0471-1>
- Reips, U.-D. (in press). Internet-based studies. In M. D. Gellman & J. Rick Turner (Eds.), *Encyclopedia of behavioral medicine* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4614-6439-6_28-2
- Reips, U.-D. (1996, October). *Experimenting in the world wide web*. Paper presented at the Society for Computers in Psychology conference, Chicago, IL.
- Reips, U.-D. (1997). Das psychologische Experimentieren im Internet [Psychological experimenting on the Internet]. In B. Batinic (Ed.), *Internet für Psychologen* (pp. 245–265). Hogrefe.
- Reips, U.-D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–117). Academic Press. <https://doi.org/10.1016/b978-012099980-4/50005-8>
- Reips, U.-D. (2001). The web experimental psychology lab: Five years of data collection on the Internet. *Behavior Research Methods, Instruments, & Computers*, 33, 201–211. <https://doi.org/10.3758/BF03195366>
- Reips, U.-D. (2002a). Internet-based psychological experimenting: Five dos and five don'ts. *Social Science Computer Review*, 20(3), 241–249. <https://doi.org/10.1177/089443930202000302>
- Reips, U.-D. (2002b). Standards for Internet-based experimenting. *Experimental Psychology*, 49(4), 243–256. <https://doi.org/10.1026/1618-3169.49.4.243>
- Reips, U.-D. (2002c). Theory and techniques of conducting web experiments. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 229–250). Hogrefe & Huber.
- Reips, U.-D. (2006). Web-based methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 73–85). American Psychological Association. <https://doi.org/10.1037/11383-006>
- Reips, U.-D. (2007a, November). *Reaction times in Internet-based research*. Invited symposium talk at the 37th Meeting of the Society for Computers in Psychology (SCiP) Conference, St. Louis, MO.
- Reips, U.-D. (2007b). The methodology of Internet-based experiments. In A. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford handbook of Internet psychology* (pp. 373–390). Oxford University Press.
- Reips, U.-D. (2008). How Internet-mediated research changes science. In A. Barak (Ed.), *Psychological aspects of cyberspace: Theory, research, applications* (pp. 268–294). Cambridge University Press. <https://doi.org/10.1017/CBO9780511813740.013>
- Reips, U.-D. (2009). Internet experiments: Methods, guidelines, metadata. *Human Vision and Electronic Imaging XIV, Proceedings of SPIE*, 7240, Article 724008. <https://doi.org/10.1117/12.823416>
- Reips, U.-D. (2010). Design and formatting in Internet-based research. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 29–43). American Psychological Association. <https://doi.org/10.1037/12076-003>
- Reips, U.-D., & Bannert, M. (2015). *dropR: Analyze dropout of an experiment or survey* [Computer software] (R package version 0.9). Research Methods, Assessment, and iScience, Department of Psychology, University of Konstanz. <https://cran.r-project.org/package=dropR>
- Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, 40(3), 699–704.
- Reips, U.-D., & Garaizar, P. (2011). Mining Twitter: Microblogging as a source for psychological wisdom of the crowds. *Behavior Research Methods*, 43, 635–642. <https://doi.org/10.3758/s13428-011-0116-6>
- Reips, U.-D., & Krantz, J. H. (2010). Conducting true experiments on the web. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 193–216). American Psychological Association. <https://doi.org/10.1037/12076-01>
- Reips, U.-D., & Lengler, R. (2005). The web experiment list: A web service for the recruitment of participants and archiving of Internet-based experiments. *Behavior Research Methods*, 37, 287–292. <https://doi.org/10.3758/BF03192696>
- Reips, U.-D., Morger, V., & Meier, B. (2001). "Fünfe gerade sein lassen": Listenkontexteffekte beim Kategorisieren ["Letting five be equal": List context effects in categorization]. https://www.uni-konstanz.de/iscience/reips/pubs/papers/re_mo_me2001.pdf
- Reips, U.-D., & Neuhaus, C. (2002). WEXTOR: A web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, and Computers*, 34(2), 234–240. <https://doi.org/10.3758/BF03195449>
- Revilla, M. (2016). Impact of raising awareness of respondents on the measurement quality in a web survey. *Quality and Quantity*, 50(4), 1469–1486. <https://doi.org/10.1007/s11135-015-0216-y>
- Rodgers, J., Buchanan, T., Scholey, A. B., Heffernan, T. M., Ling, J., & Parrott, A. (2001). Differential effects of ecstasy and cannabis on self-reports of memory ability: A web-based study. *Human Psychopharmacology: Clinical and Experimental*, 16(8), 619–625. <https://doi.org/10.1002/hup.345>
- Rodgers, J., Buchanan, T., Scholey, A. B., Heffernan, T. M., Ling, J., & Parrott, A. C. (2003). Patterns of drug use and the influence of gender on self-reports of memory ability in ecstasy users: A web-based study. *Journal of Psychopharmacology*, 17(4), 389–396. <https://doi.org/10.1177/0269881103174016>
- Roth, M. (2006). Validating the use of Internet survey techniques in visual landscape assessment: An empirical study from Germany. *Landscape and Urban Planning*, 78(3), 179–192. <https://doi.org/10.1016/j.landurbplan.2005.07.005>
- Schmidt, W. C. (2000). The server-side of psychology web experiments. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 285–310). Academic Press.
- Schmidt, W. C. (2001). Presentation accuracy of web animation methods. *Behavior Research Methods, Instruments, & Computers*, 33, 187–200. <https://doi.org/10.3758/BF03195365>
- Shapiro-Luft, D., & Cappella, J. N. (2013). Video content in web surveys: Effects on selection bias and validity. *Public Opinion Quarterly*, 77(4), 936–961. <https://doi.org/10.1093/poq/nft043>

- Sheinbaum, T., Berry, K., & Barrantes-Vidal, N. (2013). Proceso de adaptación al español y propiedades psicométricas de la Psychosis Attachment Measure [Spanish version of the Psychosis Attachment Measure: Adaptation process and psychometric properties]. *Salud Mental*, 36(5), 403–409. <https://doi.org/10.17711/sm.0185-3325.2013.050>
- Shevchenko, Y., Kuhlmann, T., & Reips, U.-D. (2021). Sampil: A user-friendly smartphone app and web-based means of scheduling and sending mobile notifications for experience-sampling research. *Behavior Research Methods*, 53, 1710–1730. <https://doi.org/10.3758/s13428-020-01527-9>
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological Momentary Assessment. *Annual Review of Clinical Psychology*, 4, 1–32. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091415>
- Siuda, P. (2009). Eksperyment w Internecie – nowa metoda badań w naukach spotecznych [Experiment on the Internet: New social sciences research method analysis]. *Studia Medioznawcze*, 3(38).
- Summerson, R., & Bishop, I. D. (2012). The impact of human activities on wilderness and aesthetic values in Antarctica. *Polar Research*, 31, Article 10858. <https://doi.org/10.3402/polar.v31i0.10858>
- Stephens-Davidowitz, S. (2017). *Everybody lies: Big data, new data, and what the Internet can tell us about who we really are*. Dey Street Books.
- Stieger, S., & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior*, 26(6), 1488–1495. <https://doi.org/10.1016/j.chb.2010.05.013>
- Stieger, S., & Reips, U.-D. (2016). A limitation of the Cognitive Reflection Test: Familiarity. *PeerJ*, 4, Article e2395. <https://doi.org/10.7717/peerj.2395>
- Stieger, S., & Reips, U.-D. (2019). Well-being, smartphone sensors, and data from open-access databases: A mobile experience sampling study. *Field Methods*, 31(3), 277–291. <https://doi.org/10.1177/1525822X18824281>
- Tan, Y. T., Peretz, I., McPherson, G. E., & Wilson, S. J. (2021). Establishing the reliability and validity of web-based singing research. *Music Perception*, 38(4), 386–405. <https://doi.org/10.1525/mp.2021.38.4.386>
- Trapnell, P. D., & Campbell, J. D. (1999). Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology*, 76(2), 284–304. <https://doi.org/10.1037/0022-3514.76.2.284>
- Van Steenbergen, H., & Bocanegra, B. R. (2016). Promises and pitfalls of web-based experimentation in the advance of replicable psychological science: A reply to Plant (2015). *Behavior Research Methods*, 48, 1713–1717. <https://doi.org/10.3758/s13428-015-0677-x>
- Verbree, A.-R., Toepoel, V., & Perada, D. (2020). The effect of seriousness and device use on data quality. *Social Science Computer Review*, 38(6), 720–738. <https://doi.org/10.1177/0894439319841027>
- Vereenooghe, L. (2021). Participation of people with disabilities in web-based research. *Zeitschrift für Psychologie*, 229(4), 257–259. <https://doi.org/10.1027/2151-2604/a000472>
- Wang, Y., Yu, Z., Luo, Y., Chen, J., & Cai, H. (2015). Conducting psychological research via the Internet: In the west and China. *Advances in Psychological Science*, 23(3), 510–519. <https://doi.org/10.3724/SP.J.1042.2015.00510>
- Weichselgartner, E. (2021, March 23). *Software for psychology experiments: Which software can be used to run experiments in the behavioral sciences (online and offline)?* https://docs.google.com/document/d/1WphZzNfwX_BWfJ4OLqN9OXXKx-EJAdI5tV_tQh9_0Sk/edit#heading=h.u9rno6vphbo4
- Whitehead, L. C. (2007). Methodological and ethical issues in Internet-mediated research in the field of health: An integrated review of the literature. *Social Science and Medicine*, 65(4), 782–791. <https://doi.org/10.1016/j.socscimed.2007.03.005>
- Wolfe, C. R. (2017). Twenty years of Internet-based research at SCiP: A discussion of surviving concepts and new methodologies. *Behavior Research Methods*, 49, 1615–1620. <https://doi.org/10.3758/s13428-017-0858-x>
- Younes, N., & Reips, U.-D. (2019). Guidelines for improving the reliability of Google Ngram studies: Evidence from religious terms. *PLoS One*, 14(3), Article e0213554. <https://doi.org/10.1371/journal.pone.0213554>
- Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111(4), 493–504. <https://doi.org/10.1037/pspa0000056>

History

Received August 5, 2021

Accepted September 2, 2021

Published online December 17, 2021

Acknowledgments

I thank Tom Buchanan for his valuable feedback on this review article. The action editor for this article was Michael Bošnjak. For further information about the author's iScience research group with their focus on experimental psychology and Internet science, please visit <https://iscience.uni-konstanz.de/en/>

Conflict of Interest

The author declares that there is no conflict of interest.

Funding

Open access publication enabled by Hogrefe Publishing.

Ulf-Dietrich Reips

Department of Psychology
University of Konstanz
Universitätsstraße 10
Fach 31
78464 Konstanz
Germany
reips@uni-konstanz.de