Assessing the Effects of Technical Variance on the Statistical Outcomes of Web Experiments Measuring Response Times

Andrew Brand[1] and Michael T. Bradley[2]

Abstract

A simulation was conducted to assess the effect of technical variance on the statistical power of web experiments measuring response times. The results of the simulation showed that technical variance reduced the statistical power and the accuracy of the effect size estimate by a negligible magnitude. This finding therefore suggests that researchers' preconceptions concerning the unsuitability of web experiments for conducting research using response time as a dependent measure are misguided.

Keywords

web experiments, response times, technical variance, simulation, statistical power, effect sizes

Introduction

For just over a decade, psychologists have been conducting their research online, and this trend is reflected in the emergence of websites that provide links to a growing number of web-based studies (e.g., Psychological Research on the Net—http://psych.hanover.edu/research/exponnet.html, Web Experiment List—http://www.wexlist.net/, Online Psychology Research UK—http://www.onlinepsychresearch.co.uk/). Although web surveys are commonly used to conduct research, the adoption of web experiments to study perceptual and cognitive processes that have typically been investigated in traditional lab experiments has been considerably less. A possible reason for this reluctance to conduct web experiments is because experiments that investigate perceptual and cognitive processes commonly use response latencies as a dependent measure, and the measurement of response time in a web experiment is perceived to be problematic.

A main appeal of web experiments is that large sample sizes can be obtained (Birnbaum, 2001, 2004; Hewson, 2003; Reips, 2000, 2002). Hence, web experiments typically tend to have more

[1]King's College London, London, UK
[2]University of New Brunswick, Saint John, NB, Canada

Corresponding Author:
Andrew Brand, Institute of Psychiatry, King's College London, London, SE5 8AF, UK
Email: andrew.brand@kcl.ac.uk

statistical power and yield more accurate effect size estimates of true effect sizes than lab experiments. However, the increase in statistical power and accuracy of the effect size estimate due to the web experiment's large sample size may be offset to some degree by an increase in unexplained error variance. This increase in error variance may be due to variation in environmental factors, for example, variation in background noise, lighting conditions, viewing angle, and presence of distractions in the testing environment (Hecht, Oesker, Kaiser, Civelek, & Stecker, 1999). This increase in error variance may also be due to technical variance (i.e., variation in technical factors), for example, variation in computers, monitors, keyboard, operating systems, and web browsers (Reips, 2000). In most cases, the variation in environmental and technical factors is not knowable; hence, variation in these factors may not be statistically controlled for.

For a web experiment employing within subjects manipulation, variation in environmental and technical factors between subjects will have no effect on statistical power and effect size estimate. The influence of variation in environmental and technical factors between participants, however, will be disruptive when an experiment involves manipulating a between-subjects factor. This is because any effect of the between-subject factor manipulation on the dependent measure will be masked by variation in environmental and technical factors between the participants. Consequently, variation in environmental and technical factors between the participants will reduce statistical power and the accuracy of the effect size estimate, when a between-subject factor is manipulated.

A response time dependent measure from a web experiment employing a between-subjects manipulation is most likely to be affected by technical variance. This is because the accuracy in measuring response times can vary markedly between different computer setups. For instance, it has been shown that response times can vary considerably between different models of mice and keyboards (Plant, Hammond, & Whitehouse, 2003; Plant & Turner, 2009). Other factors may also cause variation in response times such central processing unit (CPU) speed, operating systems, device drivers, number of background processes, and applications concurrently running (see Plant & Turner, 2009).

Studies have also assessed the accuracy of response times from experiments using web-based client-side technologies by comparing response time measurements for a known interval. Keller, Gunasekharan, Mayo, and Corley (2009) assessed the accuracy of response time measurements using Java, by using keystroke repetition to generate an interval of a known length. Their findings showed that a laptop running Windows XP and several other applications overestimated the expected interval of 300 ms by approximately 20 ms, whereas a desktop personal computer (PC) running Linux and several applications overestimated the interval of 100 ms by approximately 1 ms. Similarly, Tew and McGraw (2002) assessed the accuracy of response time measurements using Authorware, by employing a DaqPad as a keystroke generator to transmit a keystroke at a known interval. However, the data from Tew and McGraw (2002) showed that the accuracy of response times could vary with regard to different machine configurations (e.g., CPU speed, memory, operating system).

Several web experiments using a response time measure have reported comparable results to lab experiments but none of these studies have made a direct statistical comparison between data from a web experiment and a lab experiment (e.g., Corley & Scheepers, 2002; McGraw, Tew, & Williams, 2000; Nosek, Banaji, & Greenwald, 2002; Reimers & Maylor, 2005). However, a study by Reimers and Stewart (2007) did directly compare response times for a web and lab version of a simple binary choice task. The response times were on average 30 ms longer in the web version than in the lab version with millisecond accurate timing. Although this difference in accuracy between the lab and web version of the experiment is encouragingly small, it is difficult to gauge the effect that different machines have on the accuracy of the response time measure. This is because the study's sample size was small ($N = 22$) and participants when performing the web version were allowed to use "student machines" at the University that are likely to be similar in specification, hence, any general assessment of the possible variation in response time measurements between different machines is likely to be an underestimation.

Although it is clear that technical variance will reduce the statistical power and the accuracy of the effect size estimate from a web experiment that employs a between-subjects manipulation with a response-time-dependent measure, the *extent* to which the statistical power and the accuracy of the effect size estimate will be reduced by technical variance is unclear. To gain an appreciation of the degree that statistical power and the accuracy of the effect size estimate is reduced as a result of technical variance, a simulation was conducted.

Method

The simulation was conducted using the statistical package R (R Development Core Team, 2009). The R packages used were Generalized Additive Models for Location Scale and Shape (GAMLSS; Stasinopoulos & Rigby, 2007) and Methods for the Behavioral, Educational, and Social Sciences (MBESS) (Kelley, 2007). The majority of web experiments use client-side based languages (e.g., Java, JavaScript, Flash, and Authorware), where the timing measurement is executed locally on the client's machine rather than on the web server. Hence, the simulation assumed that the speed of the Internet connection of a participant's machine would not affect the accuracy of the response time measure.

The simulation of the effect of technical variance on the statistical outcome of a web experiment is intended to present a worst-case scenario, where conditions realistically maximize the adverse effects of technical variance on the statistical outcome of a web experiment.

To focus exclusively on the effects of technical variance on the statistical outcomes of a between-subjects experiment, between-subjects experiments were simulated where the dependent measure was a response time from a single trial as opposed to where the dependent measure is based on the average response time across multiple trials. There are two reasons for basing the simulation on a single trial. First, Brand, Bradley, Best, and Stocia (2011) have shown that statistical power and standardized effect size estimates based on an average from multiple trials will vary in relation to the correlation between trial data. Hence, the influence of technical variance on statistical power and standardized effect size estimates measured from a single trial will not be masked by effects from correlations between trial data. Second, the effects of technical variance will be greatest for a single trial experiment because the variance in the timing error associated with different hardware/software configurations will diminish as the number of trials in a web experiment increase. This is because timing error statistically averages out across multiple trials (Damian, 2010). Therefore, a simulation based on a single trial experiment will maximize the deleterious effect of technical variance on the statistical outcomes of a between-subjects experiment.

An ex-Gaussian distribution is commonly used to model human response times (see, Hohle, 1965; Luce, 1986; McGill, 1963), therefore ex-Gaussian distributions were used to create the data distributions for each condition. The control condition data distribution was created where the mean of the normal component (μ) is 500 ms, the standard deviation of the normal component (σ) is 50 ms, and the mean of the exponential component (τ) is 100 ms. Note that these values of μ, σ, and τ are commonly used when modeling human response times for cognitive task (Damian, 2010; Heathcote, Popiel, & Mewhort, 1991). Because response time data are commonly transformed by applying a log transformation to reduce the positive skewness of the response time distribution (e.g., Box & Cox, 1964), the three experimental conditions were created so that the difference between the log transformed data from the control distribution and log transformed data from an experimental distribution corresponded to one of Cohen's (1977) definitions of small ($d = 0.20$), medium ($d = 0.50$), and large ($d = 0.80$) effect sizes. To achieve the experimental condition distributions for the three different effects sizes, the mean of the normal component (μ) of the control distribution was adjusted. For the small effect size ($d = 0.20$) μ was adjusted to 520 ms, for medium effect size ($d = 0.50$) μ was 550 ms, and for the large effect size ($d = 0.80$) μ was 580. For a probability density function plot of the ex-Gaussian response times distributions that were created see Figure 1.
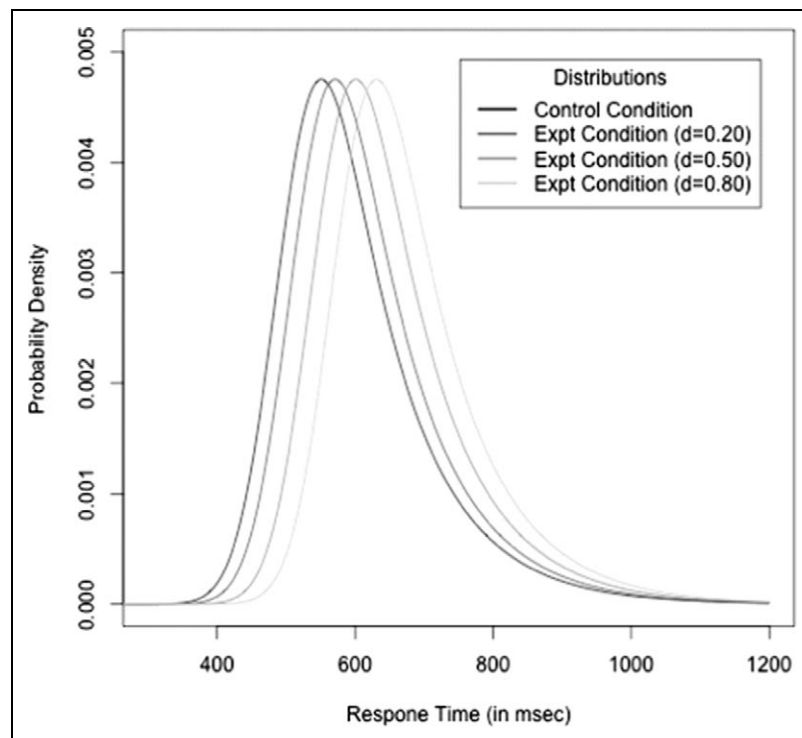
Figure 1. Ex-Gaussian response times distributions.

For each of the three effect sizes, 100,000 web experiments were simulated as follows. First, 79 response times from the control distribution were randomly sampled. Plant (personal communication, September 10, 2009) suggests that the timing error for a response time measure in a web experiment would be typically in the range of 10–100 ms. This range reflects the findings of Plant and Turner (2009) that showed the timing error associated with a variety of different mice range from 10 ms to 49 ms and also that other factors (e.g., CPU speed, memory, operating system, CPU load) may also contribute to timing error. For each of the response times, a randomly sampled value from a uniform distribution ranging from 10 to 100 was added to simulate the timing error associated with the various hardware/software configurations. Next, a log transformation was applied to all the response times. Then this procedure was repeated for the experimental distribution. The overall sample size of 158 for a simulated experiment was obtained from a survey of web experiments by Musch and Reips (2000). For each simulated web experiment, the observed effect size was calculated and a two-tailed between-subjects *t*-test was computed to determine whether a statistically significant difference ($p < .05$) was obtained. To determine the statistical power of a web experiment, the percentage of the 100,000 web experiments that obtained statistical significance was calculated. For each set of 100,000 simulated experiments, the mean effect size estimate and the mean percentage of the differences between the effect size estimate and the true effect size were calculated. This therefore enables the reduction in the accuracy of the effect size estimate due to the technical variance in response timing accuracy to be assessed.

To calculate the statistical power when there is no timing error, the above procedure was repeated but this time no timing error was added to the raw response time data. Then to assess the loss of statistical power due to the technical variance in response timing accuracy, the statistical power

Table 1. Statistical Power of a Web Experiment and the Loss of Statistical Power due to Technical Variance for the Small, Medium, and Large True Effect Sizes

	Small Effect ($d = 0.20$), %	Medium Effect ($d = 0.50$), %	Large Effect ($d = 0.80$), %
Statistical power	22	85	100
Statistical power loss	1	3	0

Table 2. Mean Effect Size Estimate of a Web Experiment and the Reduction in Accuracy of the Mean Effect Size Estimate due to Technical Variance for the Small, Medium, and Large True Effect Sizes

	Small Effect ($d = 0.20$)	Medium Effect ($d = 0.50$)	Large Effect ($d = 0.80$)
Mean effect size estimate	0.19	0.48	0.77
Reduction in the accuracy of the mean effect size estimate (%)	1	3	4

of the web experiments with timing error was subtracted from the statistical power of the web experiments with no timing error.

The results of the simulation are summarized in Tables 1 and 2. Table 1 shows floor effects on power loss with almost none and no power loss with 0.20 and 0.80 effect sizes. The maximal power loss was with the 0.50 effect size, but even then it was negligible with a 3% drop due to technical variance. The reduction in the accuracy of the effect size estimate due to technical variance was again negligible across the three benchmark true effect sizes. The small error in estimates tended to increase in a more linear fashion such that with rounding error the largest effect size of 0.80 was found to be reduced by almost 4%.

Discussion

Although it is appreciated that technical variance in a web experiment with a between-subjects design measuring response times will reduce the statistical power and the accuracy of the observed effect size estimate, the findings from the simulation show that even under conditions that realistically maximize technical variance the loss of statistical power and the reduction in the accuracy of the effect size estimate of a web experiment due to technical variance is minimal across small, medium, and large effect sizes. These findings reflect previous assessments of the effect that timing errors have on the statistical power of a typical lab experiment, where one machine is used to test all participants. For instance, Damian (2010) showed that error variance due to a lag in polling the input device is minimal relative to the error variance as a result of human task performance. Damian concluded that a lag in polling the input device has a minimal effect on the statistical power of an experiment. Similarly, Ulrich and Giray (1989) showed that an internal clock with a low timing resolution negligibly affects statistical power.

The results from the simulation showed that the statistical outcomes of a web experiment would be minimally affected by technical variance in a web experiment with a between-subjects design measuring response times. However, one may argue that environmental variance with regard to variation in distractions in the immediate testing environment might have a substantial negative impact on the statistical outcomes of a web experiment measuring response times. But since distractions in the immediate testing environment (e.g., due to a phone ringing) are likely to result in outlier response times they can be excluded from statistical analysis hence the effects of distractions on the statistical outcome may be easily minimized. Other aspects of environmental variance such as

variation in screen size, viewing angle, and screen luminance may subtly increase error variance. In most cases, the increase in error variance is probably relatively small and is therefore likely to be subsumed into the simulated timing error associated with technical variance (10–100 ms). However, it is possible that for some experimental paradigms the error variance due to variation in screen size, viewing angle, and screen luminance may be expected to be sizable. In these cases, the simulation could be conducted with an additional variable to simulate timing error due to environmental variance.

One factor that might undermine the statistical outcome of a web experiment with a between-subjects design measuring response times is participants dropping out of the web experiment. This is because selective dropout in one of the between-subjects condition may mask an effect (Birnbaum, 2001, 2004; Reips, 2000, 2002). For example, Brand (2004) documents a prime example of how a differential dropout between conditions in a web experiment employing a between-subjects design can effectively nullify an effect. It is therefore highly recommended that drop-out rates for conditions be monitored, in order to assess whether selective dropout in a condition may have affected the statistical findings of the web experiment. Another factor that might impair the quality of data from web experiment data, in general, is that participants might make multiple submissions. Fortunately, there are several methods to prevent and control for multiple submissions (see Birnbaum, 2004; Reips, 2002).

The findings from the simulation suggest that that technical variance with regard to variation in the accuracy of response time measurements has a negligible detrimental impact on the statistical outcomes of a between-subjects web experiment. However, other forms of technical variance may have a greater detrimental effect on the statistical outcomes of a web experiment employing a between-subjects design. For instance, in a web experiment investigating a subliminal priming effect, variation in screen refresh rates may reduce the statistical power and accuracy of effects size estimate. Encouragingly, Weinberger and Westen (2008) have successfully demonstrated a subliminal priming effect in a web experiment involving a within subjects design. However, the possible extent that variation in screen refresh rates will influence the statistical outcomes of a web experiment with a between-subjects design is not clear, conducting simulations may help shed some light on this issue as they have done so here.

Ultimately, the validity of the findings of the simulation conducted in this article, and any simulation in general, rest largely on the legitimacy of its underlying assumptions with regard to the choice of parameter values. For the simulation described in this article, the choice of parameters values was based on previous research (e.g., the sample size for the web experiment) and informed opinion (e.g., range of the timing error). Changing these parameters values will obviously alter the results from a simulation. For example, increasing the range of the timing error for the simulated web experiment will further reduce its statistical power and the accuracy of its effect size estimate.

To improve the concision and clarity of the simulation's findings, parameters (e.g., sample size for the web experiment) were based on a single value, which was deemed to be reasonably legitimate, as opposed to being based on a set of values. Similarly, the decision to conduct a simulation based on one trial as opposed to multiple trials was to preserve the clarity of the simulation's findings. This is because Brand et al. (2011) showed that standardized effect size estimates and statistical power can be considerably distorted in multiple trial experiments and, as a consequence, the effects of the technical variance in timing accuracy on the statistical outcomes of the web experiment would have been masked.

Given that the simulation presented in this article focuses exclusively on the effects of technical variance on the statistical outcomes of web experiments measuring response times, researchers must be careful not to over generalize from these findings. This is because different experimental paradigms may need to take into account different sources of error or bias, when a web experiment is employed to investigate an effect. For instance, for some experimental paradigms selective dropout in a between-subjects condition may be an issue. Hence, the value and validity of conducting web experiments must ultimately be judged by the researcher on a paradigm-to-paradigm basis.

However, such judgments may be invaluably informed by findings from simulations similar to the one presented here.

Conclusions

In summary, the findings of the simulation strongly suggest that the effects of technical variance on the statistical outcomes of web experiments measuring response times are minimal. Similarly, to Gosling, Vazire, Srivastava, and John (2004) who demonstrated that preconceptions of Internet-based questionnaires were unfounded, our findings in conjunction with previous findings investigating the accuracy of response times in web experiments, suggest that researchers' preconceptions about the unsuitability of web experiments for conducting response time based research are also misguided.

Therefore, in light of these findings, researchers should not dismiss the use of web experiments to investigate perceptual and cognitive processes measuring response times. This is especially so if it is understood that web experiments can have potentially far greater statistical power than traditional lab experiments. Researchers should not consider survey-based research as the only type of research that is suitable to conduct on the web. Other more traditional lab experiments may also be effectively implemented online. Moreover, given the potential advantages of web experiments, in terms of statistical power and truly voluntary motivated participants, researchers should be actively pursing this approach rather than shunning it.

Declaration of Conflicting Interests

Funding

References

Birnbaum, M. H. (2001). *Introduction to behavioral research on the Internet.* Upper Saddle River, NJ: Prentice Hall.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, *55*, 803-832.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, *26*, 211-246.

Brand, A. (2004). *A web experiment based enquiry into the verbal overshadowing effect (PhD thesis).* Cardiff: Cardiff University.

Brand, A., Bradley, M. T., Best, L. A., & Stocia, G. (2011). Multiple trials may yield exaggerated effect size estimates. *The Journal of General Psychology*, *138*, 1-11.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (revised edition). New York: Academic Press.

Corley, M., & Scheepers, C. (2002). Syntactic priming in English sentence production: Categorical and latency evidence from an Internet-based study. *Psychonomic Bulletin and Review*, *9*, 126-131.

Damian, M. F. (2010). Does variability in human performance outweigh imprecision in response devices such as computer keyboards? *Behavior Research Methods*, *42*, 205-211.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, *59*, 93-104.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*, 340-347.

Hecht, H., Oesker, M., Kaiser, A., Civelek, H., & Stecker, T. (1999). A perception experiment with time-critical graphics animation on the World-Wide Web. *Behavior Research Methods, Instruments and Computers*, *31*, 439-445.

Hewson, C. (2003). Conducting research on the Internet. *The Psychologist*, *16*, 290-293.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, *69*, 382-386.

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, *41*, 1-12.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, *39*, 979-984.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York, NY: Oxford University Press.

McGill, W. J. (1963). Stochastic latency mechanisms. In R. D. Luce & R. R. Bush (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 309-360). New York, NY: Wiley.

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of web-delivered experiments: Can you trust the data? *Psychological Science*, *11*, 502-506.

Musch, J., & Reips, U. D. (2000). A brief history of web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 61-87). San Diego, CA: Academic Press.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics*, *6*, 101-115.

Plant, R. R., Hammond, N., & Whitehouse, T. (2003). How choice of mouse may affect response timing in psychological studies. *Behavior Research Methods*, *35*, 276-284.

Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a word of commodity computers: New hardware, new problems? *Behavior Research Methods*, *41*, 598-614.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Reimers, S., & Maylor, E. A. (2005). Task switching across the life span: Effects of age on general and specific switch costs. *Developmental Psychology*, *41*, 661-671.

Reimers, S., & Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of RT measurement capabilities. *Behavior Research Methods*, *39*, 365-370.

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). SanDiego, CA: Academic Press.

Reips, U. D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*, 243-256.

Stasinopoulos, D. M., & Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, *23*, 1-46.

Tew, M. D., & McGraw, K. O. (2002). *The accuracy of response timing by Authorware programs*. Unpublished manuscript. Retrieved from http://www.psych.uni.edu/psychexps/Scrapbook/Timing_2002.pdf

Ulrich, R., & Giray, M. (1989). Time resolution of clocks: Effects on reaction time measurement: Good news for bad clocks. *British Journal of Mathematical & Statistical Psychology*, *42*, 1-12.

Weinberger, J., & Westen, D. (2008). RATS, we should have used Clinton: Subliminal priming in political campaigns. *Political Psychology*, *29*, 631-651.

Bios

Andrew Brand is a software developer for the Department of Psychology at King's College London. He is also the creator of iPsychExpts (www.ipsychexpts.com), a website that encourages and promotes the use of web experiments for conducting psychological research. Email address: andrew.brand@kcl.ac.uk.

Michael T. Bradley teaches cognition and social psychology at the University of New Brunswick. He has conducted research on information and memory tests with the polygraph and has a long-standing interest in effect size and power issues. Email address: bradley@unbsj.ca.