

A reminder on millisecond timing accuracy and potential replication failure in computer-based Psychology experiments:

An open letter

Richard R. Plant

The Black Box ToolKit Ltd

There is an ongoing 'replication crisis' across the field of Psychology in which researchers, funders and members of the public are questioning the results of some scientific studies and the validity of the data they are based upon. However few have considered that a growing proportion of research in modern Psychology is conducted using a computer. Could it simply be that the hardware and software, or experiment generator, being used to run the experiment itself be a cause of millisecond timing error and subsequent replication failure? This article serves as a reminder that millisecond timing accuracy in Psychology studies remains an important issue and that care needs to be taken to ensure that studies can be replicated on current computer hardware and software.

Keywords: Replication failure, Experiment generator, Millisecond timing, Timing error, Computer-based experiment

The Black Box ToolKit Ltd, PO Box 3802, Sheffield, S25 9AG, UK

r.plant@blackboxtoolkit.com

+44 (0)114 3030056

Author Note

Richard R. Plant is a director of The Black Box ToolKit Ltd.

There is an ongoing 'replication crisis' across the field of Psychology in which researchers, funders and members of the public are questioning the results of some scientific studies and the validity of the data they are based upon (Pashler & Wagenmakers, 2012). Areas for concern range from experimenter expectancy and statistical power through to publication bias and the file drawer problem to outright research fraud. Some have gone so far as to suggest that Bayesian tests might be applied to quantify the results or efficacy of replication attempts so that the field might know which studies are more valid and go some way to ameliorate the issue (Verhagen & Wagenmakers, 2014). However few have considered that a growing proportion of research in modern Psychology is conducted using a computer. Whether that be under 'controlled conditions' in a laboratory or more widely online across the web. Could it simply be that the hardware and software, or experiment generator, being used to run the experiment itself be a locus for replication failure?

With an increasing number of publications making use of complex computer-based experimental methods I feel now is an appropriate juncture to remind researchers that if they present stimuli, synchronize between equipment or report response times in units of a millisecond, they should consider whether what they do is always reliable, accurate and valid. Secondly can they honestly state what the timing accuracy of their study was; and thirdly are they confident they could replicate their experimental effect in another lab using different hardware and software. As replication forms the cornerstone of the scientific method requests such as these should not prove unwelcome or overly onerous and can only enhance the standing of our field.

Unfortunately faster hardware has not improved accuracy and one might argue that the degree of experimental control on offer today is worse than 20 years ago. This could be attributed to changing display technologies, multi-tasking operating systems and manufacturers striving to reduce component cost and quality by offloading previous hardware tasks to software alternatives. Timing artifacts caused by factors such as input lag on TFT monitors, soundcard start-up latency, and polling delays on non-specialist response devices are now more common than ever. In sum accuracy has continued to decrease but our confidence in the equipment and the perception of accuracy has risen as computers have become faster and ubiquitous.

In fields such as EEG, MEG, fMRI and in complex cognitive paradigms that demand millisecond accurate stimulus presentation, synchronization and event marking, timing consistency is becoming a pressing problem for researchers. In fact the issue is so acute that Psychology Software Tools (PST), the vendors of E-Prime, are launching a new device called Chronos in 2015 which will allow researchers to move sound presentation away from the PC altogether and on to custom designed hardware in order to attempt to circumvent the timing variability inherent in modern soundcards (Zuccolotto, Babjack, Cernicky, Sobotka, Basler & Struthers, 2014). It is not uncommon for soundcard start-up latency to introduce timing lags running into hundreds of milliseconds regardless of which experiment generator is used (<http://www.pstnet.com/eprimestartup.cfm>). Similarly in vision the inherent design of commodity TFT monitors and their electronics introduces input lag and refresh rate uncertainties that means certain models may suffer stimulus presentation delays into the tens of milliseconds (<http://www.displaylag.com>). Driven by the need to counter this in certain studies Thurgood, Whitfield and Patterson (2011) have produced millisecond accurate display devices using LEDs that allowed them to successfully run studies that revealed humans can recognize outlines of animals with 83% accuracy at exposure times down to one millisecond. Later authors have decreased visual presentations still further to around 250 microseconds (1/4 of a millisecond) using custom LCD Tachistoscopes and found that ultrabrief stimuli while not detectable on a conscious level still evoke a brain response when measured using EEG (Sperdin, Spierer, Becker, Michel & Landis, 2014). In both sensory modalities as a result of ensuring accurate timing new avenues of research have opened up and traditional views have been brought into question.

With growing interest in online browser-based experimentation due to developments in HTML5, CSS3 and JavaScript, new experiment vendors such as Cognilab (www.cognilab.com) are promoting the use of the web for reaction time studies. However as with standalone experiments there are timing issues inherent in the technologies employed here too in terms of presentation and response accuracy (Garaizar, Vadillo & López-de-Ipiña D, 2014). Critically web delivery mechanisms may be built on shifting sands as (Garaizar et al, 2014) point out, '... rapid development cycles of the technologies used, make it difficult, if not impossible, to offer results and conclusions about the latest versions of those technologies. The competition among the main developers of user-agents (Google, Microsoft, Mozilla, Apple, Opera) has resulted in a plethora of updates, which are hard to keep current (e.g., Mozilla Firefox took longer than seven years to

pass from version 1.0 to 5.0, but in the last three years has published over twenty-five new versions). In vogue research delivered on touch screen mobile phones and tablets also suffers from presentation and response timing inaccuracy. Whether app-based, or on the web, the technology itself cannot be relied upon (e.g. touch screens typically have at least 100ms of latency).

It is almost impossible to circumvent all possible sources of error due to the complexities inherent in the hardware and software used today as this example from Garaizar & Vadillo, 2014, clearly illustrates when benchmarking the timing accuracy of PsychoPy using a Black Box Toolkit:

"Results of the tests conducted with Ubuntu Linux running on the MacBook Pro Mid 2009. Before gathering these data, we found a problem in the execution of our tests: Preliminary tests showed that the stimulus durations registered by the BBTK photosensors doubled the expected values (e.g., white and black frames lasted 200 ms in the 100 ms condition). Surprisingly, this error was not reported in the PsychoPy log file. After commenting these results with the developers of PsychoPy, they informed us that in some configurations of Linux the graphics card is being told twice to wait for a vertical blank before proceeding, so every frame actually takes two frames. Because the frame time remains consistent, PsychoPy assumes that the frame rate of the monitor is 30 Hz (and not 60 Hz). Therefore, it does not report any missed frames (all frames look like the expected period by this measure). Fortunately, there was a simple solution. PsychoPy includes a property option to disable the wait for the next vertical blank (`win.waitBanking = False`). After implementing this change, we tested the 200 ms condition and found no timing errors."

The sobering fact of the matter is that without using external chronometry there would be little chance of noting that stimuli had been presented for double the amount of time! Worryingly the log files suggested that the test had run normally imbuing an unwarranted confidence in the experiment, equipment and software used. How would a researcher know there was a timing issue without using external chronometry? How would they know to alter some arcane setting to remedy timing error on their specific equipment? Where does this leave replication? It is worth noting that I firmly believe benchmarking various experiment generators using idealized equipment and simplified scripts is counterproductive and lulls researchers into a false sense of security and is unrepresentative of what they do and may achieve in the field. I personally discounted this approach after our 2001 & 2002 papers in which we benchmarked the three leading experiment generators of the day, E-Prime, SuperLab and ERTS, on over simplistic presentation and response tasks.

For over a decade my colleagues and I have consistently highlighted the causes of inaccurate timing across a range of sensory modalities and commonly used hardware and software (Plant et al, 2002, 2003, 2004, 2009, 2013, 2014) with the issues and implications collectively summarized in, 'Could millisecond timing errors in commonly used equipment be a cause of replication failure in some neuroscience studies?' (Plant & Quinlan, 2013). In other scientific fields equipment is routinely calibrated and error limits stated in publications (e.g., instrumented and calibrated laboratories in Chemistry). One would not wish for our field to be regarded unfavorably simply because of a failure to acknowledge that artifacts can and do reside within equipment.

These examples are neither mutually exclusive nor exhaustive. The common thread being that the software or script being run can know nothing of the errors introduced by the hardware, rendering engine or plug-in. It, just like the experimenter, will blindly act on and report the timings given without question. However in the real world a human participant, or other piece of equipment where synchronization is called for, are likely to succumb to such timing errors. All major experiment generator vendors acknowledge these issues and make their best endeavors to circumvent them where possible. However singling out one experiment generator, web platform or technology is counterproductive as the number of permutations at the individual researcher level is exponential. Nor is attempting to highlight a published study that may have achieved significant results due to bad timing. Readers should bear in mind that post-hoc statistical correction cannot hope to completely ameliorate presentation or synchronization issues as often there can be a systematic bias between conditions (e.g., a sound stimulus in one condition and not another). Neither can one ignore the possibility of human error on the part of the experimenter when constructing paradigms.

Worryingly in the literature there seems to be a growing trend toward 'checking' timing accuracy by proxy by running a study and if the results generally tally with those expected or are inline with previously published findings then the timing 'must' have been acceptable. Indeed this is how the authors of QRTengine, an online experiment generator and delivery system (Barnhoorn, Haasnoot, Bocanegra & Steenbergen, 2014) validated its efficacy rather than using external chronometry to check presentation and response timings. In no other field of science would this be viewed as acceptable.

The current landscape is best summarized by the vendors of Paradigm, a commonly used platform for running experiments, when they say, "How much will your experiments timing differ from ours? Honestly, it's almost impossible to tell. There are many sub-optimal combinations of hardware and software that could negatively affect your experiments timing." (www.paradigmexperiments.com/timing.html).

Personally I would wish to put the onus on the researcher to ensure that they continually meet the high standards our field expects of them and check their own equipment's accuracy. Furthermore that they employ the correct methodology and technology for a given study: if they are intent on web-delivery that they fully understand the limitations of the specific technology they are using. For example, the web-based experiment generator QRTengine may not offer reliable presentations or ISI's due to the need to contact the server and download and run scripts live on each trial (Barnhoorn, Haasnoot, Bocanegra & Steenbergen, 2014). Simply put, in this example reliability may decrease as server load increases and the same experiment run at different times, or days of the week, may yield different results. In addition because scripts and materials are not cached globally there is also likely a startup cost in terms of execution that may unduly affect presentation, synchronization and response timings depending on which web-browser is used.

In ensuring consistency I feel there is also a role to be played by journal editors, publishers and funders as well as each researchers host institution. Notwithstanding timing inaccuracy may be acceptable where timings are not reported in milliseconds or where a specific method does not call for a high degree of precision. It should also be noted that increasing the sample size by moving online cannot hope to solve issues that are intertwined with inaccurate or variable stimulus presentation timings. In short there is no 'standard computer' upon which an experiment might be run.

It would be remiss of me to suggest what hardware and software researchers should use to run their experiments or crucially what levels of timing error is acceptable. Determining the levels of accuracy needed in a given experiment, or research area, is solely the responsibility of the researcher and to some extent the academic publications that publish their scholarly works. Knowing ones error at the outset is something that vast majority of researchers cannot currently attest to. As I have alluded to placing blind faith in modern equipment is risky at best with recognized experiment generator vendors such as PST, the makers of E-Prime, needing to produce their own custom hardware to achieve accurate audio presentation for example. Such recognition somewhat begs the question of where this leaves previously published articles which used standard commodity hardware?

As in other fields it is likely that retractions due to computer harbored error become more commonplace. For example, Crosse & Lalor, 2014, retracted their paper, 'The cortical representation of the speech envelope is earlier for audiovisual speech than audio speech.' from the Journal of Neurophysiology due to them retrospectively discovering presentation errors as a result of their equipment choice. They summarize their retraction as follows: "...we detected a subtle yet consistent misalignment in the timing of our audiovisual stimuli. Thus, the latency shift we reported for audiovisual speech in the article cannot be trusted to be accurate. Latency shifts have previously been reported for discrete audiovisual speech in humans and for discrete non-human primate vocalizations. Whether similar latency shifts also occur in the context of continuous audiovisual human speech requires further investigation." (<http://dx.doi.org/10.1152/jn.z9k-2710-retr.2014>). It should be noted that this retraction should be regarded as highly laudable and must have taken immense courage when they could have easily remained silent.

To close I would propose that researchers self-validate their own timing accuracy using external chronometry (e.g. using a Black Box Toolkit or oscilloscope) and state confidence intervals on publication. The only way to be sure of timing accuracy is to check it when paradigms are running in-situ, on the experimenter's own equipment, at the time a given study is carried out. Generalization across apparently similar hardware types or categories of study is folly. Where event marking is critical researchers could make use of devices such as the Black Box Toolkit's mBBTK (Plant, 2014), the

Cedrus StimTracker or Psychology Software Tools upcoming Chronos hardware. One should be wary of the temptation to increase the number of trials as a substitution for running a well controlled experiment.

References

- Barnhoorn, Jonathan, S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2014). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, November 2014. <http://dx.doi.org/10.3758/s13428-014-0530-7>.
- Crosse M.J. & Lalor E.C. (2014). The cortical representation of the speech envelope is earlier for audiovisual speech than audio speech. *Journal of Neurophysiology* 111: 1400–1408, 2014. <http://dx.doi.org/10.1152/jn.00690.2013> / Retraction <http://dx.doi.org/10.1152/jn.z9k-2710-retr.2014>.
- 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 (10): e109812. <http://dx.doi.org/10.1371/journal.pone.0109812>.
- Garaizar P, Vadillo MA (2014) Accuracy and Precision of Visual Stimulus Timing in PsychoPy: No Timing Errors in Standard Usage. *PLoS ONE* 9(11): e112033. <http://dx.doi.org/10.1371/journal.pone.0112033>.
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7, 528-530. <http://dx.doi.org/10.1177/1745691612465253>.
- Plant, R., & Hammond, N. V. (2001). Benchmarking the timing characteristics of tools used by behavioural scientists. *Abstracts of the Psychonomic Society (42nd Annual Meeting)*, 6, 109.
- Plant, R., & Hammond, N. V. (2002). Towards an Experimental Timing Standards Lab: Benchmarking precision in the real world. *Behavior Research Methods, Instruments, and Computers*, 34, 218-226.
- Plant, R., Hammond, N. V., & Whitehouse, T. (2003). How choice of mouse may effect response timing in psychological studies. *Behavior Research Methods, Instruments and Computers*, 35, 276-284.
- Plant, R., & Turner, G. (2004). Self-validating presentation and response timing in cognitive paradigms: How and why? *Behavior Research Methods, Instruments and Computers*, 36, 291-303.
- Plant, R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behavior Research Methods*, 41, 598-614. <http://dx.doi.org/10.3758/BRM.41.3.598>.
- Plant, R. R. & Quinlan, P. T. (2013), Could millisecond timing errors in commonly used equipment be a cause of replication failure in some neuroscience studies?, *Cognitive, Affective, & Behavioral Neuroscience*, <http://dx.doi.org/10.3758/s13415-013-0166-6>.
- Plant, R. R. (2014). Quick, quick, slow: Timing inaccuracy in computer-based studies means we may need to make use of external chronometry to guarantee our ability to replicate. *Paper presented at the 44th Annual Meeting of the Society for Computers in Psychology (SCiP)*, Long Beach, California, November 20.
- Sperdin HF, Spierer L, Becker R, Michel CM, & Landis T (2014). Submillisecond unmasked subliminal visual stimuli evoke electrical brain responses. *Human Brain Mapping*. <http://dx.doi.org/10.1002/hbm.22716>.
- Thurgood, C., Whitfield, T. W., & Patterson, J. (2011). Towards a visual recognition threshold: new instrument shows humans identify animals with only 1ms of visual exposure. *Vision Research*, 51, 1966-1971. <http://dx.doi.org/10.1016/j.visres.2011.07.008>.

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143, 1457-1475. <http://dx.doi.org/10.1037/a0036731>.

Zuccolotto, A., Babjack, D., Cernicky, B., Sobotka, S. S., Basler, L. & Struthers, D. (2014). Methods for assessing and standardizing audio stimulus presentation latencies across heterogeneous hardware and operating system platforms. *Poster presented at the 44th Annual Meeting of the Society for Computers in Psychology (SCiP)*, Long Beach, California, November 20.