Could millisecond timing errors in commonly used equipment be a cause of replication failure in some neuroscience studies?

Richard R. Plant and Philip T. Quinlan
The University of York

Neuroscience is a rapidly expanding field where complex studies and equipment setups are the norm. Often these push boundaries in terms of what technology can offer and increasingly make use of a wide range of stimulus materials and interconnected equipment (e.g., MRI, EEG, MEG, Eye trackers, biofeedback etc.). The software that bonds the various constituent parts together itself allows for ever more elaborate investigations to be carried out with apparent ease.

However research over the last decade has suggested a growing, yet under acknowledged, problem in obtaining millisecond accurate timing in some computer-based studies. Crucially timing inaccuracies can affect not just response time measurement but also stimulus presentation and synchronization between equipment. This is not a new problem but rather one that researchers may have assumed had been solved with the advent of faster computers, state of the art equipment and more advanced software. In this paper we highlight potential sources of error, their causes and their likely impact on replication.

Unfortunately in many applications inaccurate timing is not something easily resolved by utilizing ever faster computers, newer equipment or post hoc statistical manipulation. To ensure consistency across the field we advocate that researchers self-validate the timing accuracy of their own equipment whilst running the actual paradigm in situ.

Keywords: Replication, Millisecond timing accuracy, Millisecond timing error, Experiment generators, Equipment error

Richard R. Plant
The Black Box ToolKit Ltd, PO Box 3802, Sheffield, S25 9AG, UK
r.plant@blackboxtoolkit.com
+44 (0)114 3030056

Philip T. Quinlan
Department of Psychology, The University of York, Heslington, York, YO10 5DD, UK
philip.quinlan@york.ac.uk
+44 (0)1904 323190

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

Introduction

There would appear to be a growing unease within the field of neuroscience, and across psychology in general, that some findings may not be as stable, repeatable or as valid as the academic literature describing them might indicate. Some have suggested that a proportion of studies might not be repeatable at all (Pashler & Wagenmakers, 2012). Given published studies typically represent positive findings due to the infamous “file drawer problem” this could be a cause for concern (Rosenthal, 1979).

We acknowledge that replication failure is a multifaceted issue with many potential causes and a myriad of solutions especially in a field as complex as neuroscience. Factors including choice of equipment, software tools, statistical analysis, overstated effect sizes, and inferences that go beyond the available data can have a compounding effect.

However, based on our own research carried out over the last decade we feel we can account for at least a proportion of the problem. We believe that millisecond timing errors residing within researcher’s equipment, closely followed by what might be termed, ‘human error’, could account for some unstable findings.

Given continual advances in available hardware and software combined with the complexity of many paradigms and equipment setups, where computers are used to present stimuli, synchronize with other equipment and record responses, it is without doubt that there will be some degree of unquantified timing error. This can mean that stimuli are not presented when requested, that different pieces of equipment are not synchronized correctly, that event markers are not temporally aligned and that responses may be longer than indicated. As a consequence, the conclusions drawn by researchers may not be valid. An extremely troubling thought is that such problems may go unnoticed because there is no obvious indication that anything has gone or is going wrong. On these grounds we suggest that the scale of such timing errors is considerably greater than might be suspected.

As regards researcher fallibility we believe that some of the current generation may lack in-depth knowledge on all of the equipment they use on a daily basis as studies can now be constructed with relative ease. This is especially true when using some of the more advanced software packages and newer techniques in the context of functional brain imaging. There is typically an isolating layer that conceptually separates the researcher and the software used from the hardware. Because an experiment generator lets a researcher dial in various stimulus presentation timings this does not mean the hardware can physically match what is requested; often it cannot.

More traditional human error can also be an issue in that stimulus materials may be incorrectly prepared, sequence timings wrongly specified or external equipment suboptimally utilized. Such oversights are understandable because on the surface all may appear well. A deeper understanding of how the hardware operates raises an awareness of potential pitfalls and forearms the researcher in tempering expectations. By reflecting on such possibilities we urge caution in approaching the literature because extant experimental effects may in part be due to such mistimings.

Three basic questions crystallize the issue:

1. Are researchers always carrying out the experiments they assume they are in terms of stimulus presentation, synchronization and response timing accuracy?
2. Are researchers aware of any mistimings that affect stimulus presentation, synchronization and response timing accuracy and can they quantify such together with the affect they may have on the validity of results?
3. Are researchers confident that they can replicate all experimental parameters so as to ensure that comparable results are obtained when different hardware and software in another laboratory are used?

Unfortunately we are unable to state with certainty what the exact magnitude of timing errors were within any specific published study or what the impact was on the results obtained. Such issues can only be settled at the time a given study is carried out by using external chronometry (i.e., to independently validate timing accuracy whilst the study was running in a live environment). All we can state with absolute certainty is that all studies are likely to suffer from varying degrees of presentation, synchronization and response time errors if timing error was not specifically controlled for. The aim here is to raise awareness of the issues and spread good practice across the field and not to attempt to raise uncertainty.

An Exemplar Study Illustrating Potential Sources of Timing Error and their Magnitude

By using an exemplar study from the neuroscience literature it is possible to highlight potential sources of timing error and their magnitude in the absence of carrying out actual bench tests with external chronometry (e.g., oscilloscopes, photodiodes, signal generators and logic analyzers). We are able to accomplish this because the timing errors discussed are inherent in each category of equipment used and are well recognized within the electronics industry. Moreover, they have been reliably observed in the field of experimental psychology through empirical means (Plant & Hammond, 2002; Plant, Hammond & Whitehouse, 2003; Plant & Turner, 2004, 2009, 2012).

Although the study described below is based on a recently published article, it should be noted that this was chosen purely for its complexity and use of equipment and not because we suspect any timing error per se. It remains anonymous as it is impossible to establish retrospectively to what degree any specific study was, or was not, affected by timing errors. This can only be done at the time a study was run. We do not intend to discuss the actual studies methodology, results or conclusions drawn and have removed identifying references to specific equipment manufacturers.

The exemplar study chosen relates to using both fMRI and EEG to study sentence processing and spatiotemporal dynamics of argument retrieval and reordering. In the fMRI testing sessions a three Tesla scanner with a 12-channel head coil was used with a commercial experiment generator. Visual stimuli were presented using an LCD XGA mirror-projection system with a reported refresh rate of 100 Hz[1] and auditory stimuli using air-conduction headphones. Responses were collected using a scanner safe response box. In contrast, in the EEG portion a 64 channel/DC amplifier was used with a commercial experiment generator. Visual stimuli were presented using a CRT monitor with a refresh rate of 75 Hz and auditory stimuli using bookshelf speakers (located 100 cm from participant). Responses were collected using a standard response box. A summary of equipment used is shown in table 1.

fMRI	EEG
Three Tesla scanner (12-channel head coil)	64 channel/DC amplifier
Commercial experiment generator	Commercial experiment generator
LCD XGA mirror-projection system (100 Hz)	CRT monitor (75 Hz)
Air-conduction headphones	Bookshelf speakers
Scanner safe response pad	Standard response pad

The following were unspecified in the methodology section: Computer system; computer operating system; audio device/soundcard; graphics card; revision of experiment generator software; response pad make and model; scanner software; and EEG software.

Table 1 Summary of equipment used in the exemplar study

An overview of the fMRI methodology was as follows: subsequent to the presentation of a visual fixation cross, an auditory stimulus was presented after a random interval of either 0, 500, 1000 or 1500 ms. Given this timing variation, the trials were padded with silence so that they were a consistent length of eight seconds. Next a comprehension question was visually presented for 1500 ms. Finally visual feedback given by an emoticon was presented for 1000 ms. Yes/No responses to comprehension questions were collected via button presses. The EEG methodology was essentially the same apart from the fact that the initial visual fixation cross was presented for a random duration in the range of 2000 to 3500 ms and participants instructed to blink only in this period.

In analyzing the possible sources of timing error in each hardware and software component, it is useful to begin with those that might be associated with the most significant cause for concern.

Soundcard Startup Latency

Soundcards used in modern computers suffer from something known as startup latency. This relates to the delay between the time a sound is requested to be played to the moment it can be physically detected at the speakers. It is matter of fact that all soundcards suffer from some degree of startup latency and that manufacturers rarely specify such timings. Latency can range from a few milliseconds to several seconds and can vary markedly between different soundcards and their electronics. Unfortunately there is no way to ascertain the delay through software alone and external chronometry is needed so as to quantify it exactly. In engineering terms, it is apparent that such lags are more frequently observed in modern computer systems due to operating system changes over time (e.g., the soundcard driver model of Microsoft Windows Vista/7/8 typically produces worse timing than earlier versions of this operating system). The integration of electronic components and offloading of sound processing to the host computer in order to reduce manufacturing costs are also likely to have had a detrimental effect.

In an experimental setting, such as in the exemplar outlined, this can have a major impact. For example, in fMRI where the experiment generator software typically waits for a scanner sync pulse before presenting the audio, it could mean that the sounds are actually played long after their intended onset. Such delays have implications for the presentation and synchronization of other stimuli and events. In relation to response time (RT), latencies are likely to be artificially elongated because the experiment generator will take onset times from when it requested a sound be played and not from when the actual stimulus occurred at the speakers. If the identical study is run again in the same laboratory using different soundcards it is likely that an untoward and unnoticed confound would be introduced across the two cases.

A similar set of problems can arise in the EEG setting where event markers are sent when a sound is assumed to begin. This can mean that markers may be made before the actual sound is truly presented. Thus an evoked potential may not be temporally yoked to the correct stimulus in the manner expected.

In relation to soundcards, Psychology Software Tools (PST), the vendors of E-Prime, summarized the issue as follows, “E-Prime reports millisecond accurate timing. This does not mean that E-Prime is capable of making hardware do things it cannot. For experiment paradigms that require auditory stimuli, much care and concern should be considered with the hardware being used. Sound cards can have good or poor startup latency, which is the time from when E-Prime tells the sound card to play sound to when the sound actually emits from the sound card or speakers. Not all sound cards are created equal.” (http://www.pstnet.com/eprimestartup.cfm).

Using PST’s freely available E-Prime soundcard timing data for E-Prime version 1.2 and 2.0 we have plotted three graphs to illustrate the issue more clearly (Fig. 1, 2 and 3). Note that the same soundcards were repeatedly tested on the same computer hardware but with different operating systems and driver revisions (hence apparent duplication). The key thing to note is that different soundcards can have very different timing characteristics and that most researchers will be unaware of what the startup latency of their own system will be at any given point. To help give a historical prospective and illustrate how soundcard startup latency has worsened: Microsoft Windows XP was released in 2001, Vista in 2006, Windows 7 in 2009 and Windows 8 in 2012.

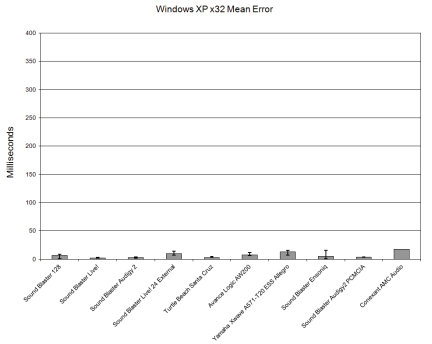

Fig. 1 Mean soundcard startup latency for Microsoft Windows XP when using E-Prime 1.2 – error bars show min and max errors

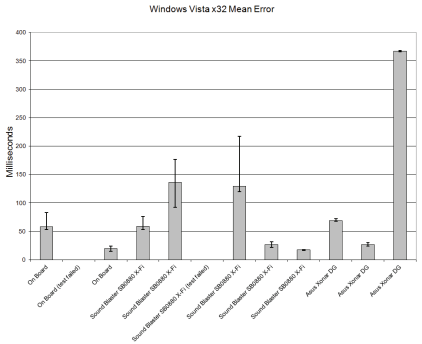

Fig. 2 Mean soundcard startup latency for Microsoft Windows Vista when using E-Prime 2.0 – error bars show min and max errors

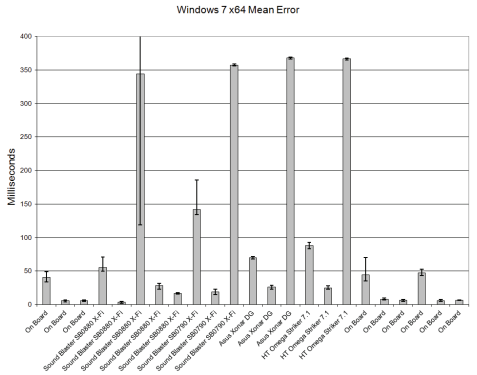

Fig. 3 Mean soundcard startup latency for Microsoft Windows 7 x64 when using E-Prime 2.0 – error bars show min and max errors (missing error bar 3001 ms)

It is important to note that PST have been at the forefront of making such timing data freely available and should be applauded for doing so. Such timing variation will occur with any experiment generator running on the same hardware as it will with any in-house software the researcher themselves have written.

Soundcard latency is something that is easily observable if two soundcards are available for direct comparison. For a sobering example of how soundcard latency can affect professional musicians as well as researchers see, http://www.youtube.com/watch?v=LcIhbAxlXSM. The effect is also often noticeable on certain laptops where there can be an unpredictable delay between clicking on the interface and receiving auditory feedback.

Air-conduction Headphones and the Speed of Sound

In brain scanning environments air-conduction headphones are often used as a matter of routine. Critically, the length of the actual tube can have an affect on when the sounds are delivered from the transducer (Killion, 1984). Largely this can be attributed to the speed of sound in air which adds approximately 1 ms per foot (30 cm) to latency. Such latency is over an above that introduced due to soundcard startup latency and that of the transducer itself. Often it is not stated whether any correction has been made for such factors in published studies, what the length of tube was or the distance between speakers and participant. In terms of EEG and other laboratory-based environments the use of standard headphones or desktop speakers will give rise to

different time delays. Again this may introduce unintended conditional biases that if uncorrected will corrupt stimulus delivery timings, synchronization and accurate response registration.

Input Lag on Data Projectors and TFT Monitors

Input lag is a well-understood phenomenon within electronics and is the delay between when an image appears on screen from when the signal was sent by the computer into the monitor cable (for a definition and generic testing methodology see, http://www.tftcentral.co.uk/articles/input_lag.htm). A traditional CRT monitor is considered to have a zero input lag, whereas TFTs can have input lags of upwards of 50 ms. Delays for data projectors are typically worse with the type listed as being used in the fMRI part of the exemplar study giving rise to lags of between 50-150 ms.

An indication of the timing variance of data projectors is provided in table 2. This data was collated from specialist websites that focus on audio-visual equipment in relation to home cinema and computer gaming. Online computer gaming is one area where the timing characteristics of display equipment matters and is easily observable. Small delays can mean that an opponent can win due to them having equipment with better timing characteristics. This is because gamers commonly make use of the 'twitch reflex' and the speed of their reactions is tied to the timing latency of their display hardware. It is worth noting that computer gamers, and consumers in general, have access to the same display equipment as that used to present stimuli and measure RTs in studies of human performance in the typical laboratory.

Using different projectors in other laboratories could produce different results as participants would be exposed to stimulus materials with non-identical timing characteristics. Even if identical computers and experiment generators were used, this would still be the case. One should note that when the same projectors were retested using default settings, results were markedly worse (e.g., the Acer H9500's delay rose from 41ms with all image processing turned off to 150 ms with motion smoothing turned on: a 109 ms increase in image onset timing error). Image processing is typically used to smooth fast moving video input such as in field sports and is usually turned on by default.

Image processing off	Image processing off	CFI/motion smoothing on
Sanyo Z200 – 16 ms	Mitsubishi HC7800 – 33 ms	Panasonic PT-AE7000U – 66 ms
Espon 8350 – 18.5 ms	**Acer H9500 – 41 ms**	Mitsubishi HC7800 – 83 ms
Sony HW10 – 10-20 ms	Panasonic PT-AE7000U – 41 ms	Epson 5010 – 141 ms
Sony HW20 – 16-32 ms	JVC RS1 – 50 ms	**Acer H9500 – 150 ms**
Infocus X10 – 25 ms	BenQ W5000 – 50 ms	BenQ W7000 – 150 ms
Panasonic AR100U – 25 ms	BenQ W7000 – 50 ms	
Optoma HD800x – 30 ms	JVC RS40 – 70 ms	
Panasonic AE300U – 30 ms	Espon 5010 – 81.4 ms	
Sanyo Z3000 – 30 ms	Espon 3010 – 100 ms	
Panasonic AE4000 – 33 ms		

Table 2 Input lag measures for commonly available data projectors
(timing data collated from http://www.avsforum.com :
http://www.avsforum.com/t/1427732/video-game-lag-time-and-projectors-which-is-best-bet
http://www.avsforum.com/t/1377652/input-lag-a-scientific-experiment-epson-8350-3010-5010-more
http://www.avsforum.com/t/1068844/input-lag-of-various-projectors)

Input lag is caused by the quality and processing speed of the projectors or TFT monitors electronics. In an experimental setting this can make presentation timing and accurate synchronization with other equipment, such as in fMRI or EEG, problematic. A CRT if driven at 200 Hz can display an image in 5 ms, or one refresh. Whereas typical projectors might take over 10-30 times as long. Such onset delays can mean late presentation in a scanning environment relative to a sync pulse and early event marking in EEG. Early event marking in EEG is typically observed as the presentation computer event marks when it sends the image to the display device and not when the actual image physically appears.

Commercial experiment generators

All experiment generators are at the mercy of the operating system they run on as is any bespoke software written by the researcher. A typical Microsoft Windows based experiment generator could run on Windows XP, Vista, Windows 7 or the newly released Windows 8 and across a wide variety of hardware using a multitude of possible driver versions. There is an almost infinitely large number of ways in which IT equipment can be configured with the variations being dictated by operating system, processor type, amount and type of memory, soundcard, graphics card, and drivers for each piece of hardware ad infinitum. In short there is no standard computer upon which an experiment generator might be run. This 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." (http://www.paradigmexperiments.com/timing.html).

In terms of our exemplar study, some experiment generators are considered more suited to fMRI work whereas others are better matched to EEG. Therefore it is likely that at least a proportion of timing variability will be attributable to them. Different experiment generators when setup to run conceptually the same experiment on identical equipment will often produce different results. For example, Plant and Hammond (2001; 2001a, b), found differences between the three leading experiment generators of that period. Although, at that time ERTS, Superlab and E-Prime were tested, we have no reason to suppose that if you compared any two modern experiment generators they would produce exactly the same results when running an ostensibly identical paradigm. Different underlying code and methodologies will often produce varying results in terms of timing accuracy even when running on the same hardware and operating system. So it could be that the actual choice of experiment generator could have an unwanted effect especially if different ones are used in different settings.

Operating Systems

As all software must run on a compatible operating system it is logical to assume that the operating system itself will in part determine timing accuracy. Based on the experience of professional programmers, we would propose that one reason why more modern operating systems struggle to achieve accurate presentation, synchronization and response timing is because of the layers of complexity that have been added over time. Layers and layers of operating system software and Application Programming Interfaces (APIs) make it more difficult for application software to interact directly with the physical hardware and obtain accurate timing (Fig. 4 and 5). If one examines the date stamped layers in the .NET framework v4 we can see this laid bare.

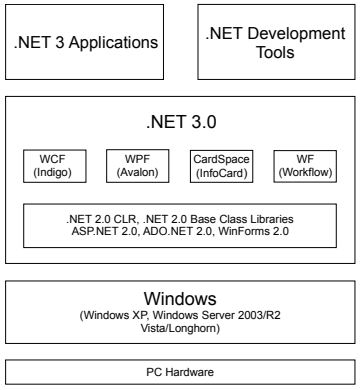

Fig. 4 Microsoft .NET Framework version 3.0

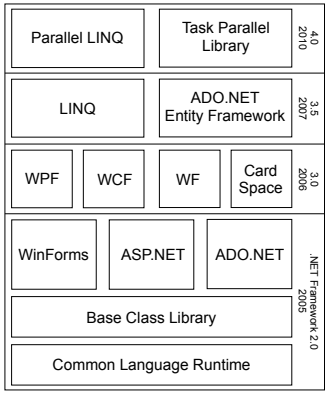

Fig. 5 Microsoft .NET Framework version 4.0

The Microsoft Windows Presentation Foundation (WPF) that is built into the .NET framework is a classic example of this layering effect (for more information see, http://msdn.microsoft.com/en-

us/library/ms754130.aspx). In essence, the WPF deals with everything that is drawn on screen. Even basic actions take much more processing time than you might imagine. If you make use of development software such as Microsoft Visual Studio then the application software you write is at a high level of abstraction. When the program is executed the code has to be interpreted and drilled down through various layers, next it is converted into a common runtime language and only then is it that the desired action takes place. Before the .NET framework and ring fenced driver models it was possible to contact the hardware directly more easily and achieve much more reliable and accurate timing. This is one reason why older but slower computers and less complex operating systems are often capable of more consistent and accurate timing despite the hardware being on the face of it an order of magnitude slower. Evidence for this can be seen in relation to soundcard startup latency discussed earlier where older operating systems produced less timing error with identical hardware.

Microsoft Windows 8 has yet another abstraction layer in addition to .NET called WinRT (for more information see, http://en.wikipedia.org/wiki/Windows_Runtime). Supposedly this enables developers to write a consistent looking application (Fig. 6) that will run on Windows 8 and also on Windows Phone 8 in much the same way as iOS from Apple supports multiple devices. However to make use of the latest standardized visual effects and techniques programmers often have to utilize libraries that might produce slower executing code which could be less timing sensitive.

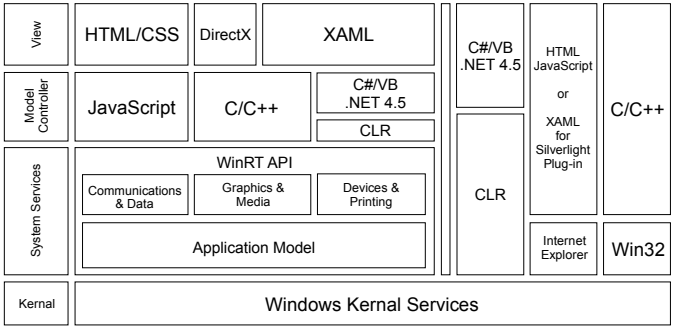

Fig. 6 Microsoft WinRT for the Windows 8 platform

Such abstraction layers are produced to help maintain backward compatibility and may account for why newer operating systems are not necessarily better at timing. Modern Apple platforms also suffer from this multilayered approach to development and as such is one reason why one operating system, or programming language, is not necessarily better than another.

In relation to our fMRI and EEG based exemplar it is feasible the one portion of the study might be run on one operating system and the other on another. Thus timing error might be introduced purely as a result of the operating system chosen even if utilizing the same experiment generator. Perhaps surprisingly a currently sizable number of laboratories still run and maintain Microsoft Windows XP machines as there is a perception that it is inherently better for timing accuracy. Based on our findings, and those from PST, this would seem to be a shrewd move.

Hardware Device Drivers

Device drivers provided by manufacturers can also be a source of timing error. For example Digital Signal Processing (DSP), or effects settings such as “generic”, “concert hall”, “sports” etc., on soundcards can add tens through to hundreds of milliseconds to the startup latency. Most computer users will leave the settings at their default unaware of the consequences. Plant and Turner (2009), for example, found that on one soundcard they tested, turning off the default generic DSP immediately improved the startup latencies by 27.5 milliseconds on average (DSP on latency [M = 37.09 ms, SD = 1.33 ms, Min = 34.38 ms, Max = 39.48 ms]; DSP off latency [M = 9.61 ms, SD = 1.04 ms, Min = 7.63 ms, Max = 12.63 ms]). Much the same processing overhead was

shown in table 2 when image processing was left turned on some data projectors tripled their input lag. Any additional processing whether done in hardware or software is likely to increase timing latency. In relation to the exemplar study it is not certain what settings were used or what impact drivers had on timing accuracy.

Scanner Safe Response Pads/Standard Response Pads

In general as long as purpose built response devices are used (i.e., known quality button boxes, or response pads, from reputable manufacturers) the measurement of RTs is often subject to relatively little variance. Even so, there still might be a large and unacceptable absolute response time error because of onsets being incorrectly marked due to temporal misalignment with a visual or auditory presentation. Examples from across the literature suggest that there can be large variations when using different types of more generic response device that are purely artifacts of the devices electronics. For example, Plant et al. (2003; Plant & Turner, 2009) showed that the range of timing error when using different makes and models of computer mice produced statistically significant effects when tested using the same simple visual RT paradigm and otherwise identical computer system (Fig. 7).

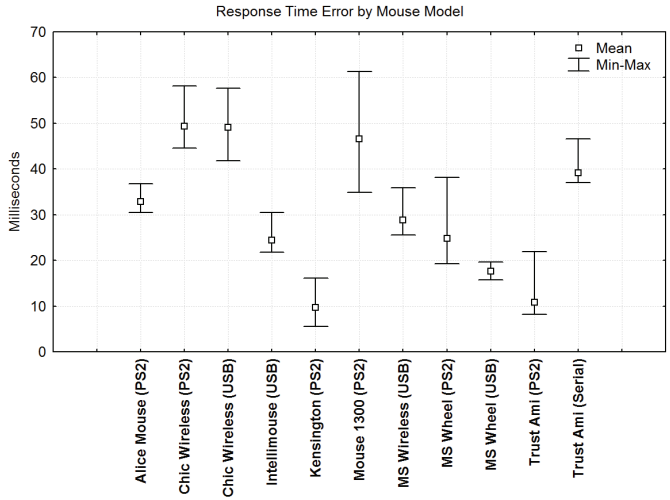

Fig. 7 Response Time error from target for various mice when measured with fixed responses (Adapted from Plant & Turner, 2009)

Psychology Software Tools have reached similar conclusions when testing keyboards as summarized in table 3.

Operating System	Response Device	Mean Error (ms)	SD (ms)
Windows XP	Keyboard (USB Belkin)	39.84	13.98
Windows XP	Keyboard (USB Dell)	8.30	0.74

Table 3 RT errors for two types of keyboard
(excerpt from http://www.pstnet.com/eprimedevice.cfm)

All generic or stock keyboards and mice will suffer from timing error irrespective of computer platform, speed or operating system used. This is because input devices generally check for keys or buttons being pressed in what is known as a polling loop. When they register a response this is sent to the computer and they carry on looping around. The speed and efficiency of this loop can vary between devices and account for the errors seen. In the context of the exemplar study, using different response devices in each experimental setting could have a negative impact on response time and synchronization.

Bugs in Software

Whilst we cannot be certain whether there were any specific bugs that affected presentation, synchronization and response timing in the exemplar study it is implausible to suppose that any software is perfect. McDonnell (2004) suggested that when developing computer code the industry average is between 15-50 defects per KLOC (1,000 lines of code). This translates to 0.1 to 0.5 defects per 1,000 lines when using a clean room environment (e.g., avionics software development). It is worth noting that commercial software can run to the millions of lines of code (due to the use of libraries) even if the user code itself is relatively short.

Even the 'best' software can have defects as is witnessed in the Northeast Power Blackout in the USA (programming bug), the NASA Mars Climate Orbiter (metric to imperial measurement miscommunication amongst designers), or something as mundane as the Denver Airport baggage-handling system (see http://www.scientificamerican.com/article.cfm?id=2003-blackout-five-years-later, http://mars.jpl.nasa.gov/msp98/orbiter/ and http://calleam.com/WTPF/?page_id=2086, respectively, for each of these examples).

It is almost impossible to estimate what percentage of defects would result in timing errors within our discipline but there are examples in the literature where bugs may have introduced unwanted issues. For example, the well publicized case of certain software potentially miscalculating diffusion tensor MRI results (Basser & Jones, 2002).

More recently, in relation to eye trackers, it has been found that some may be prone to timing and other errors as Morgante, Zolfaghari and Johnson (2011) discovered when evaluating the temporal and spatial accuracy of the Tobii T60XL eye tracker. They found that, "systematic delays and drifts were revealed in oculomotor response times"; that the "system's spatial accuracy was observed to deviate somewhat in excess of the manufacturer's estimates"; and "the experimental flexibility of the system appears dependent on the chosen software". In one of their four studies they concluded: "In summary, data from the E-Prime output file and the Tobii Studio .avi were significantly different. There was a clear, systematic bias in the .avi that appears to reflect a reduction across trials in oculomotor latencies, so much so that for most participants, it seems that they came to predict the location of the second object's appearance after a dozen or so trials. In our view, this does not provide an accurate reflection of participants' behavior, because it is not reasonable to assume that they could know this location in advance. Instead, this effect is artifactual and stems from error in the system". Of course it is not possible to tell whether or not such programming errors were present in the exemplar study. Nonetheless, should such errors be present they could adversely affect timing and may remain hidden unless an experiment is systematically tested using external chronometry.

Different Versions of the Same Software

There may be subtle and apparently imperceptible differences between different versions of the same software that could potentially impact on timing. For example, Wang, Vaidyanathan, Haake and Pelz (2012), when analyzing the SMI RED 250 remote eye tracker discovered that it was some 45% outside the manufacturers specifications in relation to the level of spatial uncertainty in participants gaze relative to a target area of interest. It was also found to exceed the calibration error on about half the trials with a two dimensional Kolmogorov-Smirnov test suggesting a statistical difference between fixation distributions recorded by two minor revisions of the same analysis software (SMI BeGaze 3.0 and BeGaze 3.1). It is sobering to think that similar bugs or differences between different versions of software used in the neuroscience field might also exist. In terms of the exemplar study it is feasible to suppose that the fMRI and EEG portions of the study might use different versions of the same software for operational reasons (e.g., only certain versions of E-Prime support extensions for fMRI, http://www.pstnet.com/support/kb.asp?TopicID=5345).

Human Error

With the complexity of experimental paradigms increasing, it is unrealistic to suppose that all studies are error free in terms of software settings, experimental scripts or synchronization between other hardware. One such parallel can be drawn from Physics where in 2012 researchers working on the Opera project at CERN announced that they thought they had managed to accelerate neutrinos to a speed faster than light

(http://www.nature.com/news/2011/110922/full/news.2011.554.html). It finally emerged that they had left a crucial cable untightened and also fell victim to a faulty GPS clock and that these had accounted for the scarcely believable results. It is not entirely unreasonable to suppose that researchers in our own field could miswire equipment or enter incorrect figures into software. It is also plausible that erroneously interfacing equipment, whether by software or hardware means, could result in incorrect use of scanner sync pulses or production of invalid EEG event markers. Due to the nature of such errors they may go unnoticed by the experimenter. In terms of the exemplar study, as timing was not independently validated it is hard to know if human error in constructing the paradigms played any role.

Other Sources of Timing Error

Although we have discussed the majority of potential sources of hardware and software based timing errors in relation to the exemplar study, for the sake of completeness it is useful to briefly outline other key factors that could apply in a wider experimental setting.

Image Intensity and Color

It is not widely appreciated that TFT monitors fade over time, meaning that they will not be as bright as they once were. This is due to wear on the fluorescent back light (CCFL) and to yellowing of the internal plastic screen and other components as a consequence of exposure to ultraviolet light. As can be seen from an industry standard-decay graph for a 14-in. TFT panel, the reduction in brightness can be striking (Fig. 8). In total, 10,000 hours equates to around 3.5 years at 8 h of use per day.

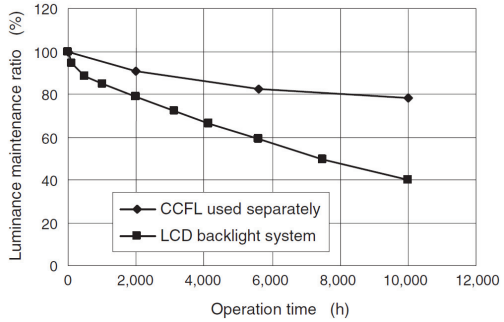

Fig. 8 Results of Normal Temperature Backlight Test
Data from Sanken Electric Co Ltd, Japan (www.sanken-ele.co.jp)
CCFL backlight manufacturers for major laptop brands

As displays age this natural fading can make images harder to perceive by participants. Even at the same settings images presented in the future may be significantly harder to perceive and react to. This fading will vary between make and model and between two apparently identical TFT monitors at ostensibly the same brightness setting. This could be problematic in studies that present images or text for short durations (e.g., in priming studies). To ensure consistency as devices age they should be calibrated using a professional TFT calibration device. This applies not just to brightness and contrast but also to color hue and intensity.

TFT Panel Response Time

By 'panel' we mean the physical TFT panel as opposed to the whole monitor, which includes the electronics needed to process the image. Panels with a slow response time are prone to blurring moving images which may affect how the stimulus appears and disappears. Panels with a quoted 5 ms response time will not actually display an image in 5 ms because of input lag nor in reality are they likely to have response times of 5 ms. It should be noted that quoted panel response time is not the same as input lag but is instead a measure of how long an individual pixel on the panel takes to switch between two colors (usually from grey to grey). Response

time also varies between make and model and manufacturers typically cherry pick the tests which make their panel appear fastest. For studies involving fast moving images (e.g. RSVP or priming) this could be a potential problem as it is likely that panel response time could have an impact on timing accuracy (for more information on panel response time see, http://www.tftcentral.co.uk/articles/response_time.htm).

Voice Keys

In certain studies voice keys are used for recording vocal responses and measuring RTs. Typically these operate on the basis of 'crossing thresholds' or the number of times an analog signal reaches a certain threshold or volume. This threshold is also susceptible to whether fricatives, plosives, voiced or unvoiced sounds are made. The researcher can be unaware of exactly what this threshold is and whether the setting is the same as one used previously. As with other response devices the electronics and the interface used can vary considerably from device to device which can add tens if not hundreds of milliseconds to actual response times. Poorly calibrated voice keys can add a significant amount of variation within an experiment. For a more in-depth discussion on voice keys and the magnitude of error they can introduce, see Kessler, Treiman & Mullennix, 2002; and Tyler, Tyler & Burnham, 2005. Striving to obtain reliably accurate responses, Tyler et al. went so far as to build their own specialized voice keys and calibrate them against hand coding of the recorded waveforms. This laborious process involved examining recorded sound waves to determine the true onsets, offsets, and durations relative to stimulus events.

Automatic Updates and Over-the-Air Push

Nearly all institutional PCs, Macs and Linux boxes will be set to automatically update their operating systems as bug fixes are released and security holes patched. However this can have unintended consequences on timing accuracy. For instance, Apple recently removed the ability to take full control of the operating system on some versions of iOS as part of an update. This will have undoubtedly affected the timing of some existing apps which used such methods by default to try and achieve better timing accuracy.

Unless the impact of an update has a visibly catastrophic effect on a study a researcher is unlikely to notice. Whether participants might is another issue. It is quite feasible for machines across a campus or hospital to be on different revisions of the same operating system and to possess different timing characteristics as a result. More savvy researchers run their machines off network and do not apply updates to help ensure consistency.

Discussion

In the previous sections we have shown how a typical neuroscience study might be negatively affected by timing errors brought about as a result of the hardware and software used. Although the causes of many of the potential timing errors highlighted are well known and understood within computer science and electronics they are perhaps less well appreciated in our discipline; hence the pressing need to raise awareness.

When computers were first used in studies of human performance there was much debate in the psychological literature as to whether millisecond timing accuracy was really needed at all. Logically if humans are much more variable than relatively small random timing errors present in the equipment then this may be true. Ulrich and Giary (1989) for example, are often cited as supporting the notion that absolute timing accuracy may not be needed. They suggested, "the effect of time resolution on detecting a true mean RT difference is negligible if the variance of the true RT is relatively large" (p. 1). Primarily they cited a timing resolution of 30 ms or worse as the point where action need be taken. They went on to propose how the RTs from relatively low resolution computer clocks of the era could be corrected post hoc via statistical methods. Crucially however Ulrich and Giary focused exclusively on RT measurement using older and effectively more accurate equipment. TFT input lag did not exist as CRTs were the only display devices available 25 years ago, soundcard startup latency was unlikely to be an issue and fMRI was only on the horizon as was high density EEG. As we have alluded to RT measurement is probably the easiest component to correct either statistically or electronically. However even if we sped up polling rates in response devices so that they registered RTs near instantly and with zero variability this does not address the question of whether RTs are measured from the true physical onset of a stimulus or

event marker. For example, due to soundcard startup latency the actual RT measure would be taken from when the experiment generator requested the sound be played and not from when the actual sound emerged from the speakers. This could be hundreds or even thousands of milliseconds adrift and is equally as bad as having an inaccurate clock. The implication for replication, especially where different equipment is used, should be obvious.

Later authors such as Neath, Earle, Hallett and Surprenant (2011) have tested modern computer setups using a photodiode and custom hardware to measure visual presentation onset and subsequent RTs when using Apple Macintosh computers and stock keyboards. They found that by using this method RTs could be as much as 100 ms too long and that different keyboards could vary by as much as 20 ms. Tellingly RTs collected were faster when tested using a CRT than a standard TFT (i.e., when using Psychtoolbox under MATLAB to present stimuli and register response times). In psychophysics terms Neath et al concluded, "if a researcher tests all subjects using the exact same hardware, if the focus is on relative rather than absolute RTs, if the differences in RT in the conditions to-be-examined are expected to be fairly large (e.g., at least 20-40 ms), if only certain software is used, and if many properties of the visual display are not of critical importance, then the conclusions drawn from RT data collected on a stock iMac are likely to be the same as those drawn from RT data collected on custom or high-end hardware." (p. 362) We (Plant & Hammond 2001a, 2002; Plant et al., 2003) have used virtually identical methods and have reached similar conclusions as have other authors (e.g., Damian, 2010; Reimers & Stewart, 2007).

However we do not feel testing the accuracy of response devices alone addresses the underlying issue or timing errors potential impact on replication. Such findings can lull the researcher into a false sense of security in that they are unrepresentative of the studies they carry out and findings only apply to that specific piece of equipment, on that specific computer setup at that point in time. They are not generalizable in a readily usable way that applies to a real neuroscience study that uses a different and usually more complex methodology, different hardware and different software. One reason why response device timing error may be overrepresented in the literature could be because it is relatively straightforward to test. It is far more difficult to test whole-system timing – that is, to measure multimodal presentation, synchronization, and response timing across multiple hardware devices when they are running in situ – in an empirical and ethologically valid way.

Nor do we believe post hoc statistical treatments and averaging remove the need to strive for accurate timing. Unfortunately such methods do not address systematic conditional biases nor can they solve the issue of late presentation of stimuli or synchronization errors between equipment. As regards replication, two ostensibly identical paradigms may actually present stimuli for very different durations despite the intentions of the experimenter. In our view post hoc statistical manipulation should only address timing issues in very specific circumstances. The negative impact of timing error on replicability might be compounded still further if different equipment is used in the field as compared with a University laboratory setting due to cost and availability issues.

In our view there is only one sure way to address timing error and that is to empirically determine its magnitude in an ethologically valid way using external chronometry and then to take appropriate action to correct it, work with it or replace the equipment or software that is being used (Plant & Hammond, 2002). Neath et al. (2011) would seem to agree as they concluded, "Should researchers conduct experiments on stock Apple Macintosh computers when the dependent variable is RT? Given the variability in RTs observed [above], we strongly recommend that researchers using any computer to collect RTs should assess the accuracy and reliability of their chosen platform. It is always desirable to minimize sources of error, and therefore one should validate the system on which one is collecting data. Given our findings [above], we can recommend using the particular hardware/software combinations tested in only some situations." (p. 362)

Another, more costly option is to produce specialist hardware that can actually do what researchers assume that commodity hardware already does (e.g., the VPixx CRT replacement TFTs, which can display images with little or no input lag). The VPixx display hardware for example has been used successfully to run studies that have revealed humans can recognize outlines of animals with 83% accuracy at exposure times down to one

millisecond (Thurgood, Whitfield & Patterson, 2011). Without expensive custom hardware this would have been impossible to achieve. As a result new avenues of research have opened up and traditional views of visual processing speed have been brought into question. Currently such solutions can be prohibitively expensive for mainstream use due to the complexity of the electronics required (http://www.vpixx.com/).

Our own approach over the last decade has been to advise researchers to check the timing of their own equipment whilst running their own paradigm in situ. That is, to empirically evaluate every aspect of their study that is timing critical using external chronometry (i.e., stimulus presentation, synchronization and response timing). This notion provided the driving ethos behind the launch of the Black Box ToolKit in 2003 and which enables the researcher to check all aspects of millisecond timing accuracy (www.blackboxtoolkit.com). Alternatives such as the StimTracker from Cedrus perform a similar function in relation to event marking (www.cedrus.com) and both solutions are used in hundreds of labs worldwide where timing is considered critical.

Unfortunately an ethologically valid approach to testing ones own timing without specialist turnkey solutions can be difficult requiring the use of oscilloscopes, logic analyzers, function generators and accompanying electronics and engineering skillsets. In addition this rarely results in a clear picture of what the whole system timing error is as such an approach does not represent what a calibrated human respondent would do. It is acknowledged that some laboratories currently make use of oscilloscopes and photodiode based approaches for example to check visual stimulus-response timing. Despite this the majority consistently fail to state this in published articles.

How has this Happened and why does it Continue to Happen?

Most researchers have a tacit awareness of the importance of millisecond accurate timing but relatively few are skilled enough to assess their own error rates and attempt to circumvent them. Many have assumed that faster computers mean more accurate timing and that this issue need no longer be addressed. Unfortunately nothing could be further from the truth. Today's technology is fundamentally different from what went before: TFTs are not drop in replacements for CRTs for example.

At a deeper level we would suggest that the reason researchers have not brought it on themselves to improve their own accuracy is two fold. The first is that journals do not request researchers state that they have checked their equipment accuracy in order to publish.

The second is by the application of AntiPatterns from software engineering. AntiPatterns are defined as: "A commonly occurring pattern or solution that generates negative consequences. An AntiPattern may be a pattern in the wrong context. When properly documented, an AntiPattern comprises a paired AntiPattern solution with a refactored solution." (Brown, Malveau, McCormick & Mowbry, 1998, p. 275.). Put simply researchers do not generally strive for as near zero timing error as possible as results often match their expectations. They continue to do this even though a well documented solution already exists, i.e., measuring timing error using external chronometry and then applying various workarounds. This is somewhat akin to people not backing up their computer until at some point they suffer catastrophic data loss. In both cases this might be considered cognitive dissonance writ large.

It is useful at this point to note that there is a sobering and subtle difference between precision and accuracy. experiment generator vendors and hardware suppliers quote "millisecond precision". The subtly is that milliseconds are quoted as being the units of measurement but there is no mention of whether the timings are accurate to within one millisecond. Precision and accuracy in our view are two very different things: an atomic clock is incredibly precise but if you read it at the wrong instant then the time read off is not accurate relative to the event you were trying to time. In the same way in an experimental setting if a sound or image is displayed later than intended it is irrelevant at what rate the clock ticks.

Researchers should also be wary of placing blind faith in software based time audits that some experiment generators produce. Logically they can know nothing of the display onset error due input lag in a specific TFT

panel's electronics. All they can tell the researcher is the time they sent the image for rendering on the screen not when it physically appeared.

Consequences of Failing to Address these Issues

Replication and the scientific method should be at the cornerstone of our discipline and we postulate that timing error accounts for at least some unstable findings reported across the literature. Undoubtedly the file drawer problem compounds the situation as does the willingness of many journals to publish findings where the methodology is poorly described and calibration data and confidence limits for equipment never mentioned. 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. Neuroscience is not unique in having to grapple with these issues. In fact anywhere where humanistic computer-based research methods are applied and measurements reported in units of a millisecond caution is warranted.

If timing error remains unaddressed we would suggest that as equipment and paradigm complexity increases the number of unstable findings could rise. At the very least effect sizes may be weakened, or strengthened, or on occasion may simply disappear thus clouding the validity of an avenue of research or given approach. As has been discussed and is highlighted in Fig. 9 the researcher's expectations as regards timing accuracy can be circumvented by a multitude of causes if unchecked.

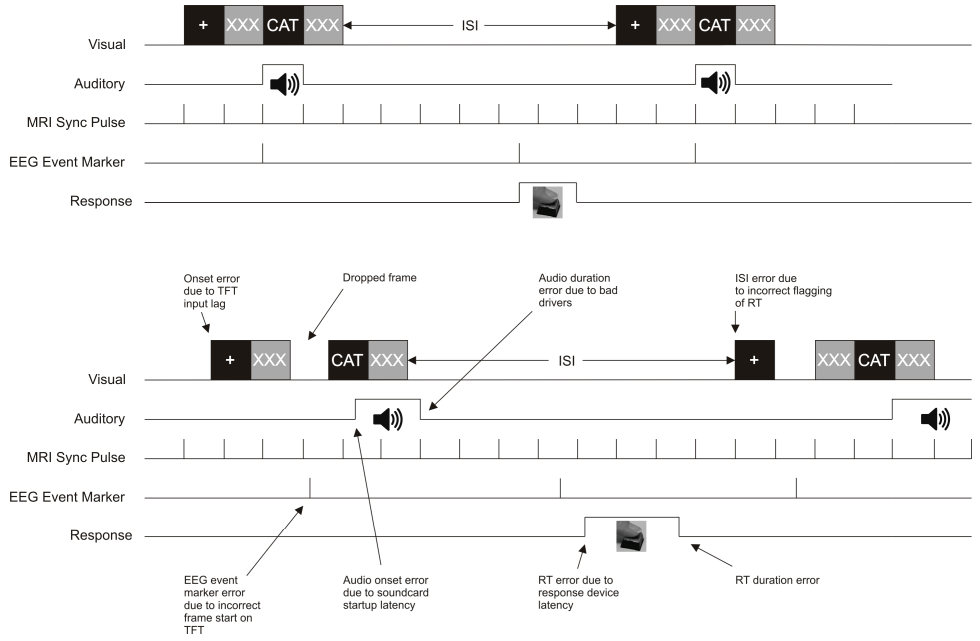

Fig. 9 An idealized model of a typical study (top) versus what may actually happen due to timing error (bottom)

Worryingly we are already beginning to see some researchers moving to unproven platforms as they are perceived to be in vogue and liked by participants. For example there is a big push for the use of Android tablets and Apple iPads in psychological research. This trend is likely to continue with the launch of Microsoft Surface and the move towards experimental deployment on tablets and phones. Few however have considered whether they actually offer a reliable experimental platform. Our own unpublished research would suggest that

much caution is needed with presentation, synchronization and response timing errors measured in the hundreds of milliseconds. Writing in his programming blog, Tyson (2011) has conducted a number of timing tests on Apple iOS (http://atastypixel.com/blog/experiments-with-precise-timing-in-ios/) as summarized in table 4 when examining different programming methods to implement accurate timers when addressing the requirement to present audio tracks in a timely fashion within music editing software. To non-programmers these represent the various options available to developers trying to obtain millisecond accurate timing on Apple hardware. Different developers may choose different methods to attempt to gain access to millisecond accurate timing depending on their own stylistic preferences. However they themselves may be unaware of the impact of every choice they make and what the wider consequence might be as they are unlikely to have used external chronometry to observe effects in the physical world.

Mechanism	Average discrepancy	Minimum discrepancy	Maximum discrepancy
NSRunLoop	16.9 ms	0.25 ms	153.7 ms
TPPreciseTimer (original)	5.5 ms	0.033 ms	72.0 ms
TPPreciseTimer (10ms spinlock)	6.0 ms	0.002 ms	76.5 ms
TPPreciseTimer (100ms spinlock)	3.7 ms	0.002 ms	44.8 ms
TPPreciseTimer (200ms spinlock)	2.91 ms	0.002 ms	74.1 ms
dispatch_after (main queue)	14.8 ms	0.16 ms	161.2 ms
dispatch_after (dedicated queue)	19.2 ms	0.1 ms	174.9 ms
dispatch_after (dedicated queue + 100ms spinlock)	22.4 ms	0.002 ms	306.8 ms

Table 4 Timing error from target using different programming techniques in iOS

As can be seen from Tyson's data, timings can be extremely variable. It should not be forgotten that Apple have some five iPad models at the last count and various versions of iOS deployed on each. Apple have also recently removed real-time mode on some versions of iOS and shifted other features to the Core Audio functions. Speculation would suggest they did this to help smooth multitasking and save battery life. This was likely done for the consumers benefit and not for researchers running experiments.

As a cautionary note toward the use of newer technologies a Microsoft Applied Sciences Group video clearly illustrating a typical 100 ms lag inherent in generic touch screens can be viewed at: http://www.youtube.com/watch?v=vOvQCPLkPt4. The implication being that any RTs in a psychology experiment that rely on touch will be late by whatever the latency of the touch registration system is on a given device. This is purely a result of the speed of the screen and touch digitizer which lies under the glass, the devices general electronics and the operating system running on it.

With the webs dominance other research groups and commercial vendors are now promoting the web browser as a credible platform for psychological research. Techniques range from Adobe Flash through to native HTML 5 and interpreted JavaScripting (e.g., http://ertslab.com/web/). Here again caution may be warranted. Reimers and Stewart (2007, 2008), have shown that, when using Adobe Flash to collect precision timed behavioral responses in a simple binary-choice experiment, RTs from uncontrolled machines used outside the laboratory were on average 20 ms slower than standardized machines inside, which in turn were approximately 10 ms slower than a calibrated Linux-based system used as a baseline.
With regard to newer technologies such as HTML 5 computer scientists recognize that different web browsers, whilst possessing the ability to display an identical webpage and execute JavaScript to run an experiment, can vary markedly in the speed in which they can do so. To help illustrate the differences between platforms and

browsers running identical code the lead author ran the SunSpider benchmark (http://www.webkit.org/perf/sunspider/sunspider.html) on a range of devices that might be used to deliver a psychological experiment. This benchmark tests commonly used real world techniques that programmers use to construct web apps that run in a browser and by association psychological experiments. To give a flavor of the levels of variation a summary of the results are as follows: For an Intel i7 desktop system running Microsoft Windows 7 the overall SunSpider score for Microsoft IE10 was 105.8 ms +/- 0.6%, for Google Chrome 25, 142.6 ms +/- 1.1% and for Mozila Firefox 19, 162 ms +/- 1.1% where faster is better (means and 95% confidence intervals). Whereas on a Google Nexus 7 tablet using Chrome the score was, 1647.9 ms +/- 1.8%, a Google Nexus 4 phone again using Chrome, 1908.8 ms +/- 3%, for Safari running on an Apple iPad, 2908 ms +/- 0.5% and on an iPhone 4, 4021.2 ms +/- 8.8%.

Even on the same desktop hardware, as the SunSpider benchmark demonstrates, there can be marked differences in the quality of each browsers web page rendering engine and it is these that could have a negative impact on supposedly 'cross platform' experiments. The bottom line is that different web browsers and hardware will in all likelihood differ in their ability to present stimuli and record responses to the same accuracy even when running identical code. Whilst such variation may be acceptable for studies that are not timing critical for others it may not.

How Can Researchers, Reviewers and Editors Address These Problems in both the Short and the Long Term?

We believe that researchers should be responsible for self-validating the timing accuracy of their own studies in order to add credibility to their findings. They might be encouraged to check each time they design a new study, change an aspect of an existing one or upgrade any hardware or software. Once a researcher knows where errors reside they can take preventative action. For example, to move an image presentation forward in time so that a TTL event marker is then aligned with the true image onset time (i.e., when it physically appears on screen) and not when it was requested to appear. This would counteract input lag on data projectors and TFT monitors. However this requires that the researcher know what the input lag was. This is something that can only be determined empirically when running the device in-situ as it will vary according to the hardware and software used. Another option would be to make use of external event marking by using something like a Black Box ToolKit or Cedrus StimTracker for example whereby when an image appears a photodiode detects the onset and sends a temporally true TTL marker.

In addition to checking timing we would suggest that researchers store an audit trail where they can prove they have done so. Whether they check using an oscilloscope, a Black Box ToolKit, StimTracker or other reliable method is irrelevant – the crucial thing is they check. It is wise to accept the notion popularized in Physics of 'known unknowns'. Without using external chronometry timing errors would remain unknown as would their likely impact.

Further we would suggest that experimental methods courses might be modified to include timing accuracy and hardware awareness at undergraduate and PhD levels. Today experiments can be constructed with such ease that the basics may be overlooked. Some of the skill of knowing how to program, or code, ones equipment together with knowledge of how the electronics work may have been lost by some of the current generation. We would suggest that reviewers might also request that researchers submit timing validation measures with submissions as a benchmark of good practice. From the point of view of journal editors we would respectfully suggest that if researchers are capable of making use of such complex experiments and equipment setups they should be capable of checking and stating the timing accuracy of their studies together with any corrective action taken. To this end we would suggest that where relevant submissions should have at least a paragraph that states how timing was checked and what the results were as some authors have begun to do (e.g., Jaekl, Soto-Faraco & Harris, 2012). One could envisage this being a recommendation initially and then becoming mandatory over the longer term. It might even be sensible for journals to have a set format for reporting such measures. For example, it would not be too onerous to produce a short protocol in the methods section which detailed timing accuracy along the lines of the one shown in Listing 1.

- Visual presentation onset error: M +32 ms, SD 2.3 ms
- Visual duration error: M +32ms, SD 2.3 ms
- EEG event marker synch error: M -32 ms, SD 2.3 ms
- Visual ISI: M +32 ms, SD 2.3 ms
- RT error relative to visual stimulus: +40-55 ms (47.5 ms subtracted post hoc)

Timing measures were self-certified using a xxxxxxxxx. Annotated timing data for this study is available at: DOI:xxxxxxxx

Listing 1. Short protocol for reporting timing accuracy

Importantly by having a DOI which links back to the timing data, there could be an audit trail that helps determine if a rigorous methodology had been followed. In much the same way more open researchers share their raw data currently. This is what we mean by self-validation and self-certification. Our personal view is that once a moratorium period had elapsed many leading researchers and laboratories would be only too happy to comply with a new gold standard for their own benefit and the benefit of the field in a wider sense. This would hopefully improve the number of successful replications and strengthen published findings. We cannot foresee any downsides.

To close it is worth noting that over the years we have consistently been asked to suggest which publications, or specific paradigms, we suspect of being impacted by timing inaccuracies. Unfortunately this is virtually impossible in retrospect as the hardware and software used would need to be assessed in the field at the time it was used. Nor are we willing to share inside knowledge of individual researcher's studies or commercial hardware and software which we have tested and was covered by Non Disclosure Agreements. We would rather the researcher themselves determine how their equipment might impact on the work they do and the results they produce.

It seems much more sensible to start with a clean slate and move forward with a coherent approach from researchers, reviewers and journal editors. Most would agree that methodology sections of articles need to be strengthened regardless. Undoubtedly the push by funding bodies such as NSF for the sharing of research data and materials is a positive step (http://www.nsf.gov/bfa/dias/policy/dmp.jsp) as is the willingness of some journals to allow submissions of supporting content. Without more experimental rigor with regard to timing, more detailed Method sections, and a closer focus on replication, we feel that the field is storing up problems for the future.

References

Basser, P. J., & Jones, D. K. (2002). Diffusion-tensor MRI: theory, experimental design and data analysis - a technical review. *NMR in Biomedicine, 15*, 456-467. http://dx.doi.org/10.1002/nbm.783

Brown, W., Malveau, R., McCormaick, S., & Mowbray, T. (1998). *AntiPatterns: Refactoring software, architectures, and projects in crisis*. New York: NY: Wiley.

Damian, M. F. (2010). Does variability in human performance outweigh imprecision in response devices such as computer keyboards? *Behavior Research Methods, 42*, 205-211. http://dx.doi.org/10.3758/BRM.42.1.205

Jaekl, P., Soto-Faraco, S., & Harris, L. R. (2012). Perceived size change induced by audiovisual temporal delays. *Experimental Brain Research, 216*, 457-462. http://dx.doi.org/10.1007/s00221-011-2948-9

Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic Biases in Voice Key Response Time Measurements. *Journal of Memory and Language*(47), 145-171. http://dx.doi.org/10.1006/jmla.2001.2835

Killion, M. C. (1984). New insert headphones for audiometry. *Hearing Instruments, 35*, 46.

McDonnell, S. (2004). *Code complete. A practical handbook of software construction*. Redmond, WA: Microsoft Press.

Morgante, J. D., Zolfaghari, R., & Johnson, S. P. (2011). A Critical Test of Temporal and Spatial Accuracy of the Tobii T60XL Eye Tracker. *Infancy, 17*, 9-32. http://dx.doi.org/10.1111/j.1532-7078.2011.00089.x

Neath, I., Earle, A., Hallett, D., & Surprenant, A. M. (2011). Response timing accuracy in Apple Macintosh computers. *Behavior Research Methods, 43*, 353-362. http://dx.doi.org/10.3758/s13428-011-0069-9

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? . *Perspectives in Psychological Science*, 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. (2001). *Towards an experimental timing standards laboratory*. Paper presented at the Society of Computers in Psychology (SCiP), Orlando, Florida, November 15.

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., & Turner, G. (2012). *Could your equipment account for your experimental effect?* Paper presented at the Society of Computers in Psychology (SCiP), Minneapolis, Minnesota, November 15.

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. http://dx.doi.org/10.3758/BF03193004

Reimers, S., & Stewart, N. (2008). Using Adobe Flash Lite on mobile phones for psychological research: Reaction time measurement reliability and interdevice variability. *Behavior Research Methods, 40*, 1170-1176. http://dx.doi.org/10.3758/BRM.40.4.1170

Rosenthal, R. (1979). The file draw problem and the tolerance for null results. *Psychological Bulletin, 83*, 638-641. http://dx.doi.org/10.1037/0033-2909.86.3.638

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

Tyler, M. D., Tyler, L., & Burnham, D. (2005). The delayed trigger voice key: An improved analogue voice key for psycholinguistic research. *Behavior Research Methods, 37*, 139-147.

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

Wang, D., Vaidyanathan, P., Haake, A., & Pelz, J. (2012). *Are eye trackers always as accurate as we assume?* Paper presented at the Society of Computers in Psychology (SCip), Minneapolis, Minnesota, November 15.

Footnotes

1. Having examined the 2003 manual for the actual data projector used in the published study, our view is that it could not have been run at 100 Hz (as that is not a supported input frequency). Our assumption is that 100 Hz and 75 Hz have been transposed.