Safe and sensible baseline correction of pupil-size data

Sebastiaan Mathôt[1,2]*, Jasper Fabius[3], Elle Van Heusden[3], Stefan Van der Stigchel[3]

[1] Department of Psychology, Heymans Institute, University of Groningen, Groningen, The Netherlands

[2] Aix-Marseille University, CNRS, LPC UMR 7290, Marseille France

[3] Department of Experimental Psychology, Helmholtz Institute, Utrecht University, The Netherlands

* Corresponding author

Address for correspondence:
Sebastiaan Mathôt
Department of Experimental Psychology
University of Groningen
Grote Kruisstraat 2/1
9712 TS Groningen
Netherlands

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2725v1 | CC BY 4.0 Open Access | rec: 13 Jan 2017, publ: 13 Jan 2017

Abstract

Measurement of pupil size (pupillometry) has recently gained renewed interest from psychologists, but there is little agreement on how pupil-size data is best analyzed. Here we focus on one aspect of pupillometric analyses: baseline correction, that is, analyzing changes in pupil size relative to a baseline period. Baseline correction is useful in experiments that investigate the effect of some experimental manipulation on pupil size. In such experiments, baseline correction improves statistical power by taking into account random fluctuations in pupil size over time. However, we show that baseline correction can also distort data if unrealistically small pupil sizes are recorded during the baseline period, which can easily occur due to eye blinks, data loss, or other distortions. Divisive baseline correction (corrected pupil size = pupil size / baseline) is affected more strongly by such distortions than subtractive baseline correction (corrected pupil size = pupil size - baseline). We make four recommendations for safe and sensible baseline correction of pupil-size data: 1) use subtractive baseline correction; 2) visually compare your corrected and uncorrected data; 3) be wary of pupil-size effects that emerge faster than the latency of the pupillary response allows (within ±220 ms after the manipulation that induces the effect); and 4) remove trials on which baseline pupil size is unrealistically small (indicative of blinks and other distortions).

Keywords: pupillometry, pupil size, baseline correction, research methods

Safe and sensible baseline correction of pupil-size data

Pupil size is a continuous signal: a series of values that indicate how pupil size changes over time. In this sense, pupil-size data is similar to electroencephalographic (EEG) data, which indicates how electrical brain activity changes over time; and it is different from most behavioral measures, such as response times, that generally provide only a single value for each trial of the experiment, such as a single response time.

Psychologist are often interested in how pupil size is affected by some experimental manipulation (reviewed in Beatty & Lucero-Wagoner, 2000; Mathôt & Van der Stigchel, 2015). To give a classic example, Kahneman & Beatty (1966) asked participants to remember a varying number (3-7) of digits. They found that participants' pupils dilated (i.e. became larger) when the participants remembered seven digits, compared to when they remembered only three; that is, memory load caused the pupil to dilate (become bigger).

As was common for pupil-size studies of the time, Kahneman & Beatty (1966) expressed their results in millimeters of pupil diameter; that is, they used absolute pupil-size values. But expressing pupil size in absolute values has a disadvantage: It is affected by slow, random fluctuations of pupil size. These fluctuations are a source of noise that reduce statistical power and make it more difficult to detect the effects of interest (in the case of Kahneman & Beatty (1966) the effect of memory load). To deal with these fluctuations, researchers often look at the *difference* in pupil size compared to a baseline period, which is typically the start of the trial. By looking at pupil-size changes, rather than absolute pupil sizes, differences in pupil size that already existed before the trial are taken into account, are no longer a source of noise, and no longer reduce statistical power. This is baseline correction.

Baseline correction is thus a way to reduce the impact of random pupil-size fluctuations from one trial to the next; it is *not* a way to control for overall differences in pupil size between participants. Of course, some participants have larger pupils than others (see Tsukahara, Harrison, & Engle, 2016 for a fascinating study on the relationship between pupil size and intelligence); and the distance between camera and eye, which varies slightly from participant to participant, also affects pupil size, at least as measured by most eye trackers. But such between-subject differences are better taken into account statistically, through a repeated measures ANOVA or a linear mixed-effects model with by-participant random intercepts (e.g. Baayen, Davidson, & Bates, 2008)—just like between-subject differences in reaction times are usually taken into account. Phrased differently, baseline correction is a way to turn a between-trial design, in which pupil sizes are compared between trials, into a within-trial design, in which pupil sizes are compared between different moments within a single trial.

There are two main ways to apply baseline correction: *divisive*, in which pupil size is converted to a proportional difference from baseline pupil size (corrected pupil size = pupil size / baseline), and *subtractive*, in which pupil size is converted to an absolute difference from baseline pupil size (corrected pupil size = pupil size - baseline). There are variations of these approaches, such as using percentage rather than proportion change, or converting absolute differences from baseline pupil size to z-scores; but these are all minor variations of these two general approaches. Here we will therefore focus on the difference between divisive and subtractive baseline correction.

How do researchers choose between divisive and subtractive baseline correction? We cannot be certain, because a reason for choosing one or the other is never given, at least not in any paper that we've seen. But we are free to speculate.

Divisive baseline correction is attractive because it provides an intuitive measure: proportion change. If a paper states that an eye movement caused a 10% pupillary constriction (Mathôt, Melmi, & Castet, 2015), you can easily judge the size of this effect: substantial but not enormous. In contrast, if a paper states that a manipulation caused a 0.02 mm diameter change (Bombeke, Duthoo, Mueller, Hopf, & Boehler, 2016), you need a moment to remember (or look up) that human pupils are 2 to 8 mm in diameter, and that a 0.02 mm effect is therefore tiny. This is, in our view, less intuitive. And if the the eye tracker reports pupil size in arbitrary units (typically based on a pixel count of the camera image), then absolute pupil-size differences become even harder to interpret. However, despite these disadvantages, subtractive baseline correction may be the natural choice for some researchers because it is the standard approach in EEG research (e.g. Gross et al., 2013; Woodman, 2010).

In pupillometry, there is no established standard for applying baseline correction. Based on our experience, most researchers now apply some form of baseline correction (but see e.g. Gamlin et al., 2007), and variations of subtractive baseline correction (Binda, Pereverzeva, & Murray, 2013; Hupé, Lamirel, & Lorenceau, 2009; Jainta, Vernet, Yang, & Kapoula, 2011; Knapen et al., 2016; Koelewijn, Zekveld, Festen, & Kramer, 2012; e.g. Laeng & Sulutvedt, 2014; Murphy, Moort, & Nieuwenhuis, 2016; Porter, Troscianko, & Gilchrist, 2007; Privitera, Renninger, Carney, Klein, & Aguilar, 2010) seem somewhat more common than variations of divisive baseline correction (Bonmati-Carrion et al., 2016; Herbst, Sander, Milea, Lund-Andersen, & Kawasaki, 2011; Mathôt, van der Linden, Grainger, & Vitu, 2013; H. Wilhelm, Lüdtke, & Wilhelm, 1998). (One paper is listed per research group. This list is anecdotal, and not a comprehensive review.)

As far as we know, no-one has systematically studied baseline correction of pupil-size data. However, baseline correction has been studied in the context of EEG/ MEG data, as shown by a recent debate about whether or not baseline correction of EEG/ MEG data should be abandoned in favor of high-pass filtering (Maess, Schröger, & Widmann, 2016; Tanner, Morgan-Short, & Luck, 2015; Tanner, Norton, Morgan-Short, & Luck, 2016). However, pupil-size data is different from EEG/ MEG data. For example, although pupil size fluctuates in cycles of 1-2 s (Mathôt, Siebold, Donk, & Vitu, 2015; Reimer et al., 2014), it does not show the slow systematic drift shown by EEG/ MEG voltages (Tanner et al., 2016). Also, pupil-size data is strongly affected by eye blinks, which result in full signal loss, preceded and followed by sharply distorted recordings (e.g. Mathôt, 2013). Of course, EEG/ MEG data is also strongly affected by blinks (Hoffmann & Falkenstein, 2008), but not as catastrophically as pupil size is.

Our aim is therefore to do study baseline correction specifically for pupil-size data. We will use real and simulated data to see how robust subtractive and divisive baseline corrections are to noise, and how they affect statistical power. We will not apply any other techniques for improving data quality, such as blink reconstruction or smoothing, because we feel that baseline correction should be safe and sensible on its own. We will end by making several recommendations for baseline correction of pupil-size data.

Simulated data

We first investigate the effect of divisive and subtractive baseline correction in simulated data. The advantage of simulated data over real data is that it allows us to control noise and distortion, and therefore to see how robust baseline correction is to imperfections of the kind that also occur in real data. In addition, simulated data allows us to simulate two experimental conditions that differ in pupil size by a known amount, and therefore to see how baseline correction affects the power to detect this difference.

Data generation

We started with a single real 3 s recording of a pupillary response to light, recorded at 1000 Hz with an EyeLink 1000 (SR Research). This recording did not contain blinks or recording artifacts, but did contain the slight noise that is typical of pupil-size recordings. Pupil size was measured in arbitrary units as recorded by the eye tracker, ranging from roughly 1600 to 4200.

Based on this single recording, 200 trials were generated. To each trial, a constant value, randomly chosen for each trial, between -1000 and 2000 was added to each sample. In Figure 1b, this is visible as a random shift of each trace up or down the y-axis. In addition, one simulated eye blink was added to each trial. Eye blinks were modeled as a period of 10 ms during which pupil size linearly decreased to 0, followed by 50 to 150 ms (randomly varied) during which pupil size randomly fluctuated between 0 and 100, followed by a period of 10 ms during which pupil size linearly increased back to its normal value. This resembles real eye blinks as they are recorded by video-based eye trackers (e.g. Mathôt, 2013).

To simulate two conditions that differed in pupil size, we added a series of values that linearly increased from 0 (at the start of the trial) to 200 (at the end of the trial) to half the trials (i.e. the same slight increase from 0 to 200 was applied to half the trials). These trials are the Red condition; the other trials are the Blue condition. As shown in Figure 1a, pupil size is slightly larger in the Red condition than in the Blue condition, and this effect increases over time.

Crucially, in the Blue condition there were two trials in which a blink started at the first sample, and therefore affected the baseline period. In none of the other trials did the baseline period contain a blink.

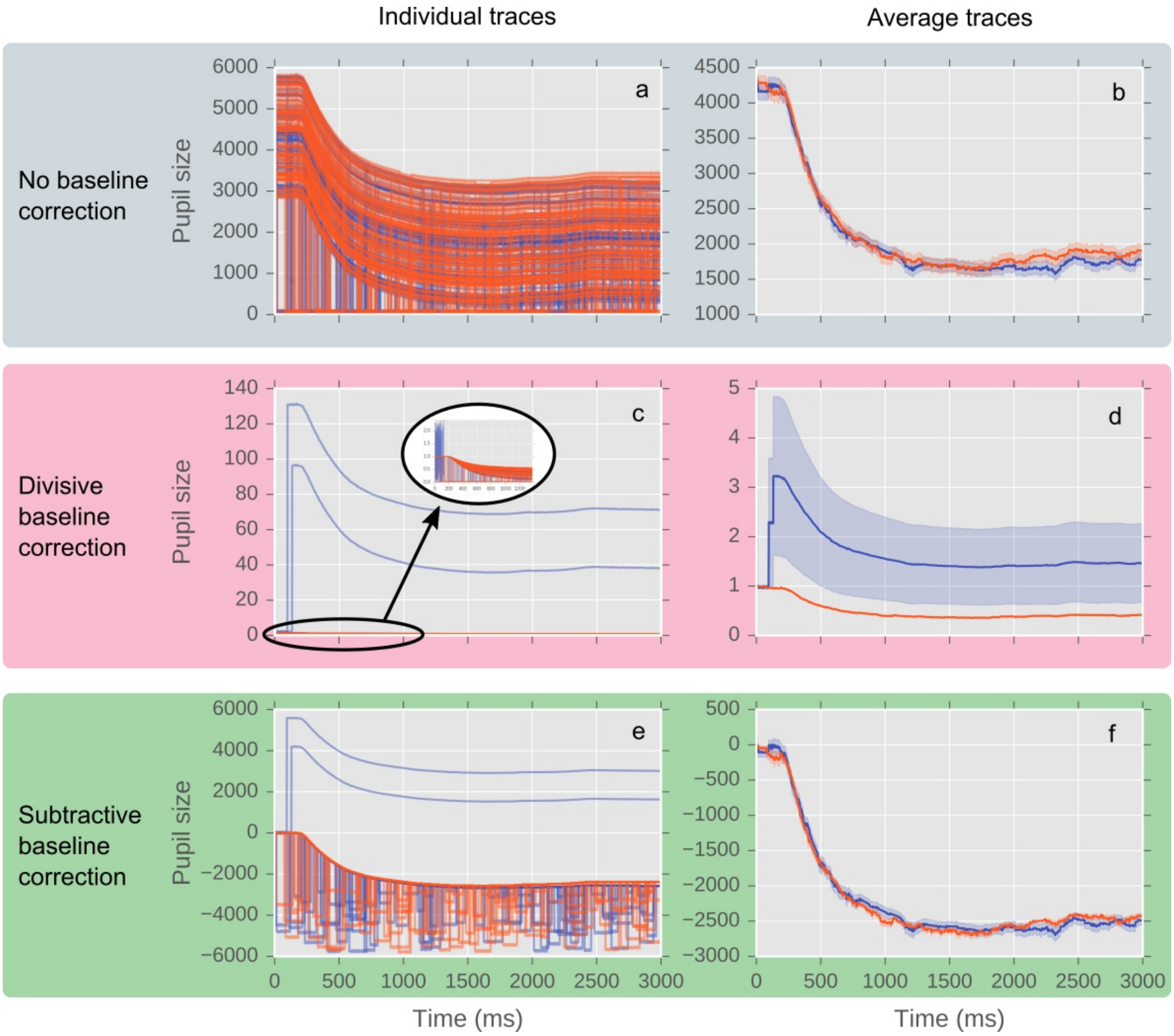

Figure 1. The effects of divisive and subtractive baseline correction in a simulated dataset. a, b) No baseline correction. Y-axis reflects pupil size in arbitrary units. c, d) Divisive baseline correction. Y-axis reflects proportional pupil-size change relative to baseline period. e, f) Subtractive baseline correction. Y-axis reflects difference in pupil-size from baseline period in arbitrary units. Individual pupil traces: a, c, e). Average pupil traces: b, d, f).

Divisive baseline correction

First, median pupil size during the first 10 samples (corresponding to 10 ms) was taken as baseline pupil size. Next, all pupil sizes were divided by this baseline pupil size. This was done separately for each trial.

(The length of the baseline period varies strongly from study to study. Some authors prefer long baseline periods of up to 1000 ms (e.g. Laeng & Sulutvedt, 2014), which have the disadvantage of being susceptible to pupil-size fluctuations during the baseline period. Other authors, including ourselves (e.g. Mathôt et al., 2015), prefer short baseline periods, which have the disadvantage of being susceptible to recording noise. But the problems that we highlight in this paper apply to long and short baseline periods alike.)

The results of divisive baseline correction are shown in Figure 1c,d. In the two Blue trials in which there was a blink during the baseline period, baseline pupil size was very small; consequently, baseline-corrected pupil size was very large. These two trials are clearly visible in Figure 1c as unrealistic baseline-corrected pupil sizes ranging from 40 to 130 (as a proportion of baseline), whereas in this dataset realistic baseline-corrected pupil sizes tend to range from 0.3 to 1 (see zoom panel in Figure 1c). Pupil sizes on these two trials are so strongly distorted that they even affect the overall results: As shown in Figure 1d, the overall results suggest that the pupil is largest in the Blue condition, whereas we had simulated an effect in the opposite direction.

Subtractive baseline correction

Subtractive baseline correction was identical to divisive baseline correction, except that baseline pupil size was subtracted from all pupil sizes.

The results of subtractive baseline correction are shown in Figure 1e,f. Again, the two Blue trials with a blink during the baseline period are clearly visible in Figure 1e. However, their effect on the overall dataset is not as catastrophic as for divisive baseline correction.

Statistical power

The results above show that the effects of blinks during the baseline period can be catastrophic, and much more so for divisive than subtractive baseline correction. However, it may be that divisive baseline correction nevertheless leads to the highest statistical power when there are no blinks during the baseline period (even though this is unlikely to happen in real data). To test this, we generated data as described above, while varying the following:

- Effect size, from 0 (no difference between Red and Blue) to 500 (Red larger than Blue) in steps of 50
- Baseline correction: no correction, divisive, or subtractive
- Blinks during baseline in Blue condition: yes (2 blinks) or no

We generated 10,000 datasets for each combination, giving a total of (10,000 × 11 × 3 × 2 =) 660,000 datasets. For each dataset, we conducted an independent-samples t-test to test for a difference in mean pupil size between the Red and Blue conditions during the last 50 samples. We considered three possible outcomes (for convenience, we treat the 0 effect size as if it still constitutes a true positive effect):

- Detection of a true effect: $p < .05$ and pupil size smallest in the Blue condition
- Detection of a spurious effect: $p < .05$ and pupil smallest in the Red condition
- No detection of an effect: $p \geq .05$

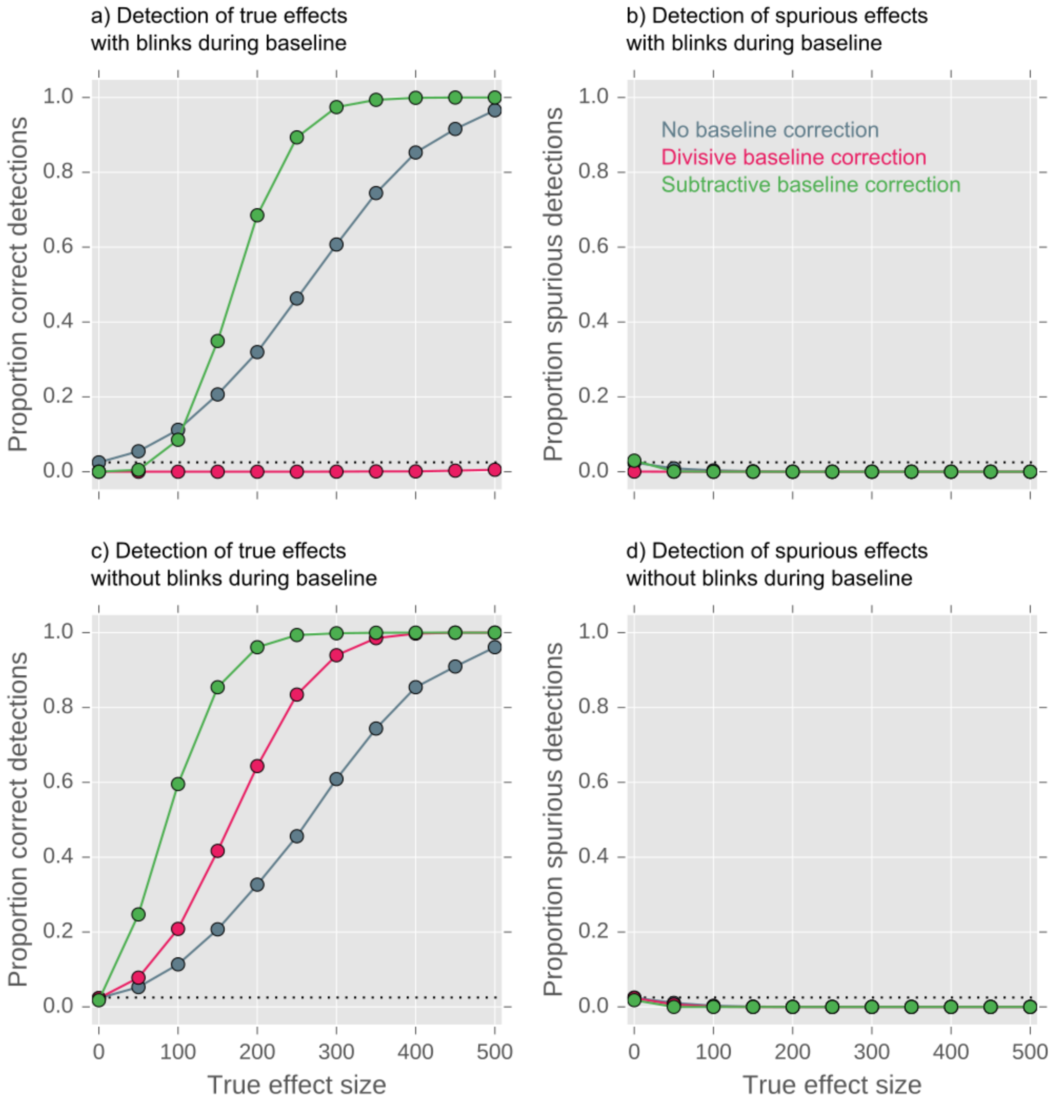

Figure 2. *Proportion of detected real (a,c) and spurious (b,d) effects when applying subtractive baseline correction (green), divisive baseline correction (pink), or no baseline correction at all (gray). Data with different effect sizes (x-axis: 0-500) and with (a,b) or without (c,d) blinks during the baseline was generated,*

The proportion of datasets in which a true effect was detected is shown in Figure 2a,c; the proportion on which a spurious effect was detected is shown in Figure 2b,d. By chance (i.e. if there was no effect and no systematic distortion of the data) and given our two-sided $p < .05$ criterion, we would expect to find a .025 proportion of detections of true and spurious effects; this is indicated by the horizontal dotted lines.

First, consider the data with blinks in the baseline (Figure 2a,b). With divisive baseline correction (pink), neither true nor spurious effects are detected (except for a handful of true effects for the highest effect sizes, but these are so few that they are hardly visible in the figure), because the blinks introduce so much variability in the signal that it is nearly impossible for any effect to be detected.

With subtractive baseline correction (green), true effects are often observed, and spurious effects are not. However, for weak effects, subtractive baseline correction is less sensitive than no baseline correction at all (gray). This is because, for weak effects, the blinks introduce so much variability that true effects cannot be detected; however, the variability is less than for divisive baseline correction, and for medium-to-large effects subtractive baseline correction is actually more sensitive than no baseline correction at all—despite blinks during the baseline.

Now consider the data without blinks in the baseline (Figure 2b). An effect in the true direction is now detected in all cases, but there is a clear difference in sensitivity: subtractive baseline correction is most sensitive, followed by divisive baseline correction, in turn followed by no baseline correction at all. The three approaches do not differ markedly in the number of detected spurious effects.

Real data

The simulated data highlights problems that can occur when applying baseline correction, especially divisive baseline correction, if there are blinks in the baseline period. But you may wonder whether these problems actually occur in real data. To test this, we looked at the effects of baseline correction in one representative set of real data.

Data

The data was collected for a different study, and consisted of 2520 trials (across 15 participants) in two conditions, here labeled Blue and Red. All participants signed informed consent before participating and received monetary compensation. The experiment was approved by the ethics committee of Utrecht University.

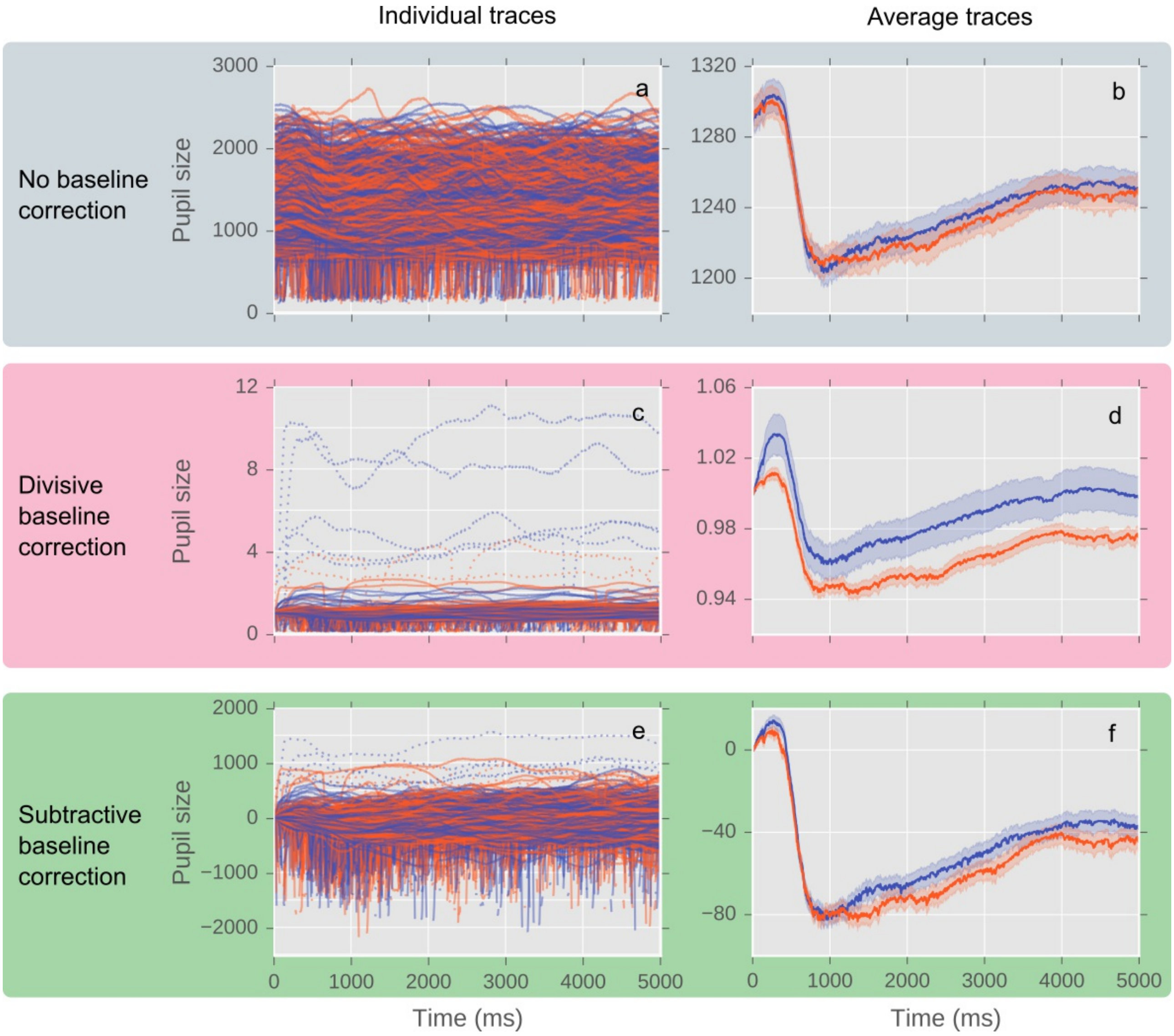

Figure 3. *The effects of divisive and subtractive baseline correction in a real dataset. a, b) No baseline correction. Y-axis reflects pupil size in arbitrary units. c, d) Divisive baseline correction. Y-axis reflects proportional pupil-size change relative to baseline period. e, f) Subtractive baseline correction. Y-axis reflects difference in pupil-size from baseline period in arbitrary units. Individual pupil traces: a, c, e). Average pupil traces: b, d, f).*

First, consider the uncorrected data (Figure 3a,b). The trial starts with a pronounced pupillary constriction, followed by a gradual redilation. Overall (Figure 3b), the pupil is slightly larger in the Blue than the Red condition, but this difference is small. (Whether or not the difference between the two conditions is reliable is not relevant in this context.) The individual traces (Figure 3a) show a lot of variability between trials, as well as frequent blinks, which correspond to the vertical spines protruding downward.

Divisive baseline correction

Figure 3c,d shows the data after applying divisive baseline correction (applied in the same way as for the simulated data). Overall, the data now suggests that the pupil is markedly larger in the Blue than the Red condition. But to the expert eye, the pattern is odd, because the difference between Blue and Red is mostly due to a sharp (apparent) pupillary dilation in the Blue condition immediately following the baseline period; afterwards, the difference remains more-or-less constant. This is odd if you know that, because of the latency of the pupillary response, real effects on pupil size develop at the earliest about 220 ms after the manipulation that caused them (e.g. Mathôt, van der Linden, Grainger, & Vitu, 2015); in other words, there should not be any difference between Blue and Red before 220 ms.

If we look at the individual trials, it is clear where the problem comes from: Because of blinks during the baseline period, baseline-corrected pupil size is unrealistically large in a handful of trials (dotted lines). Most of these trials are in the Blue condition, and this causes overall pupil size to be overestimated in the Blue condition. (It is not clear why there are more blinks in the Blue than the Red condition. This may well be due to chance. But even if the two condition systematically differ in blink rate—which would be interesting—this difference should not confound the pupil-size data!)

Subtractive baseline correction

Figure 3e,f shows the data after applying subtractive baseline correction (applied in the same way as for the simulated data). Overall, the difference in pupil size between Blue and Red is exaggerated compared to the raw data (compare Figure 3f to Figure 3b). If we look at the individual trials, this is again due to the same handful of trials (dotted lines), mostly in the Blue condition, in which pupil size is overestimated because of blinks in the baseline period. In other words, subtractive baseline correction suffers from the same problem as divisive baseline correction, but to a much lesser extent.

Identifying problematic trials

Figure 4 shows a histogram of baseline pupil sizes, that is, of median pupil sizes during the first 10 ms of the trial. In this dataset, baseline pupil sizes are more-or-less normally distributed with only a slightly elongated right tail. (But baseline pupil sizes may be distributed differently in other datasets.)

On a few trials, baseline pupil size is unusually small; but these trials are so rare that they are hardly visible in the original histogram (Figure 4a). Therefore, we have also plotted a log-transformed histogram, which accentuates bins with few observations (Figure 4b). Looking at this distribution, a reasonable cut-off seems to be 400 (arbitrary units): Baseline pupil sizes below this value are—in this dataset—unrealistic and can catastrophically affect the results as we have described above.

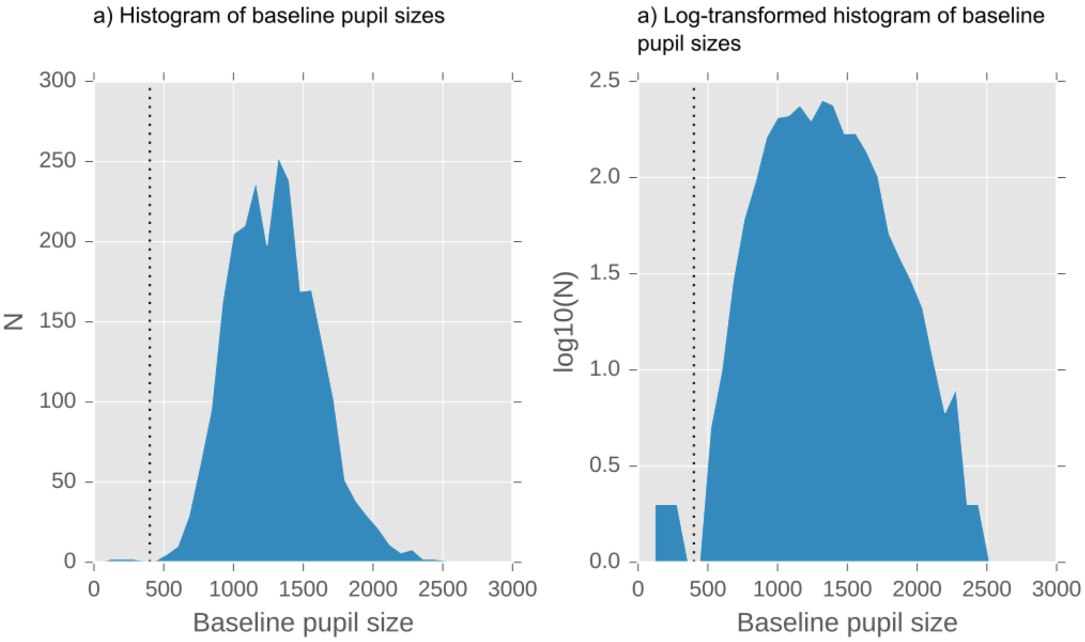

Figure 4. *A histogram of baseline pupil sizes. a) Original historgram. b) Log-transformed histogram. The vertical dotted line indicates a threshold below which baseline pupil sizes appear unrealistically small.*

In Figure 3d we have marked those trials in which baseline pupil size was less than 400 as dotted lines. As expected, those trials in which baseline-corrected pupil size is unrealistically large are exactly those trials on which baseline pupil size is unrealistically small.

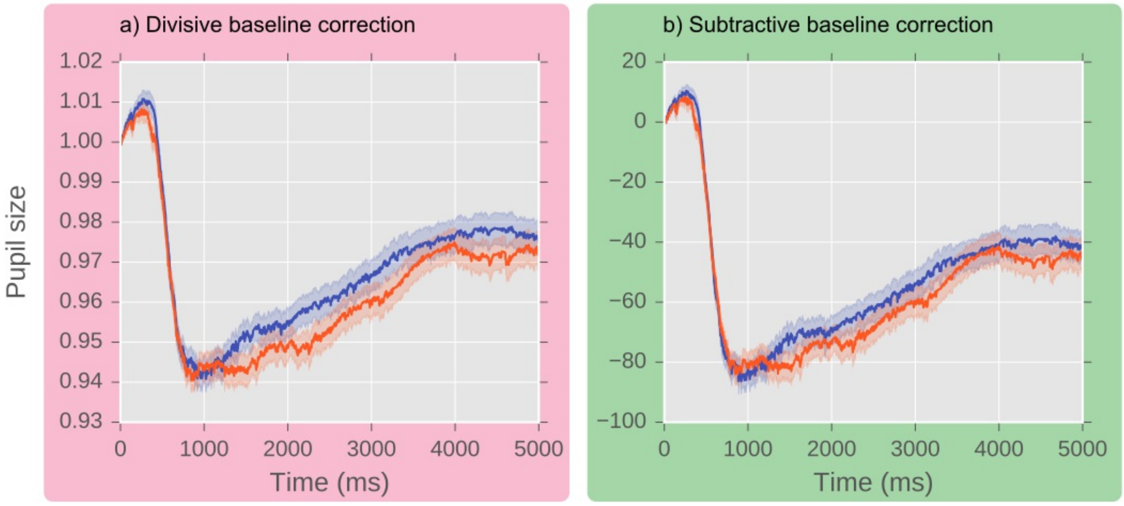

Figure 5. *Overall results for divisive (a) and substractive (b) baseline correction after removing trials in which baseline pupil size was less than 400.*

If we remove trials on which baseline pupil size was less than 400, the distortion of the overall results is much reduced. In particular, if we look at the results of the divisive baseline correction, the sharp pupillary dilation immediately following the baseline period in the Blue condition is entirely gone (compare Figure 5a with Figure 3d).

Most trials with small baselines would also have been removed if we had removed trials in which a blink was detected during the baseline. But you can think of cases in which baseline pupil size is really small while no blink is detected; for example, the eyelid can close only partly, or noisy recordings may prevent measured pupil size from going to 0 during a blink, preventing detection. Therefore, we feel that it is safer to filter based on pupil size instead of (or in addition to) filtering based on detected blinks.

Discussion (and four recommendations)

Here we show that baseline correction can distort pupil-size data. This happens most often when pupil size is unusually small during the baseline period, which in turn happens most often because of eye blinks, data loss, or other distortions. When baseline pupil size is unusually small, baseline-corrected pupil size becomes unusually large. This is a problem for all forms of baseline correction, but is much more pronounced for divisive than subtractive baseline correction. Therefore, subtractive baseline correction (correct pupil size = pupil size - baseline) is more robust than divisive baseline correction (corrected pupil size = pupil size / baseline).

Despite risk of distortion, it makes sense to perform baseline correction, because it increases statistical power in experiments that investigate the effect of some experimental manipulation on pupil size. In our simulations, subtractive baseline correction increased statistical power more than divisive correction; however, we simulated a fixed difference between conditions that did not depend on baseline pupil size. For such baseline-independent effects, subtractive baseline correction leads to the highest statistical power. But real pupillary responses are always somewhat dependent on baseline pupil size, if only because a baseline pupil of 2 mm cannot constrict much further, nor can a baseline pupil of 8 mm dilate much further. The more important point is therefore that baseline correction in general increases statistical power compared to no baseline correction.

Knowing the risks and the benefits, how can you perform safe and sensible baseline correction of pupil-size data? Based on our observations, we make four recommendations:

1. Use subtractive baseline correction (or some variation thereof); that is, we recommend that on the level of individual trials, baseline pupil size be subtracted from real pupil size. Other transformations can be applied as you see fit, but they should be applied to the aggregate data, and not to individual trials. For example, if you prefer to express pupil size as proportion change, you can divide pupil size by the grand mean pupil size during the baseline period averaged across all trials.
2. Visually compare your baseline-corrected data with your uncorrected data. Baseline

correction should reduce variability, but not qualitatively change the overall results.

3. Baseline artifacts manifest themselves as a rapid dilation of the pupil immediately following the baseline period. Given that real effects on pupil size emerge slowly, never within 220 ms of the manipulation, baseline artifacts can be distinguished from real effects by their timing.
4. Plot a histogram of baseline pupil sizes, and use this to visually determine a minimum baseline pupil size, and remove all trials on which baseline pupil size is smaller. We do not recommend using a fixed criterion such as 'remove all baseline pupil sizes that are more than 2.5 standard deviations below the mean'. While this may work in some cases (it would have worked in the real data used here), the distribution of baseline pupil sizes varies, and therefore a fixed criterion may not always catch all problematic trials. We also do not recommended relying on blink detection. Although blinks are the primary reason for unrealistically small baseline pupil sizes, they are not the only reason; furthermore, blinks may not be detected when the eyelid closes only partly or when the recording is noisy. As we've seen, catching all problematic trials is important, because even a handful of trials with baseline artifacts can catastrophically affect the overall results. A visually determined criterion for minimum baseline pupil size is safest.

We and many others have used baseline correction in the past, and some people, including us, have also used divisive baseline correction. In light of this paper, can we still trust these previous results? For the most part: yes. Importantly, baseline artifacts can trigger quantitatively large spurious effects, but these spurious effects are unlikely to be significant, because baseline artifacts also introduce a lot of variance. Therefore, baseline artifacts are more likely to result in false negatives (type II errors) than false positives (type I errors). We have also checked our own previous results (e.g. Blom, Mathôt, Olivers, & Van der Stigchel, 2016; Mathôt, Dalmaijer, Grainger, & Van der Stigchel, 2014) and found no signs of baseline artifacts, nor have we found obvious signs of distortion (of this kind) in other published work. Presumably, most researchers check their data visually to make sure that clear outliers are removed; therefore, catastrophic distortion (as shown in Figure 1c,d and Figure 3c,d) is likely to be noticed and, in one way or another, corrected. Nevertheless, although problems may be detected and dealt with in an ad hoc fashion, using a safe and sensible approach to begin with is preferable.

In conclusion, we have shown that baseline correction of pupil-size data increases statistical power, but can strongly distort data if there are artifacts (notably eye blinks) during the baseline period. We have made four recommendations for safe and sensible baseline correction, the most important of which are: Use subtractive rather than divisive baseline correction, and check visually whether your baseline-corrected pupil-size data makes sense.

Materials and availability

Data and analysis scripts can be found at https://github.com/smathot/baseline-pupil-size-study.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. http://doi.org/10.1016/j.jml.2007.12.005

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (Vol. 2, pp. 142–162). New York, NY: Cambridge University Press.

Binda, P., Pereverzeva, M., & Murray, S. O. (2013). Attention to bright surfaces enhances the pupillary light reflex. *Journal of Neuroscience*, *33*(5), 2199–2204. http://doi.org/10.1523/jneurosci.3440-12.2013

Blom, T., Mathôt, S., Olivers, C. N., & Van der Stigchel, S. (2016). The pupillary light response reflects encoding, but not maintenance, in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(11), 1716. http://doi.org/10.1037/xhp0000252

Bombeke, K., Duthoo, W., Mueller, S. C., Hopf, J., & Boehler, N. C. (2016). Pupil size directly modulates the feedforward response in human primary visual cortex independently of attention. *NeuroImage*, *127*, 67–73. http://doi.org/10.1016/j.neuroimage.2015.11.072

Bonmati-Carrion, M. A., Hild, K., Isherwood, C., Sweeney, S. J., Revell, V. L., Skene, D. J., … Madrid, J. A. (2016). Relationship between human pupillary light reflex and circadian system status. *PLOS ONE*, *11*(9), e0162476. http://doi.org/10.1371/journal.pone.0162476

Gamlin, P. D. R., McDougal, D. H., Pokorny, J., Smith, V. C., Yau, K., & Dacey, D. M. (2007). Human and macaque pupil responses driven by melanopsin-containing retinal ganglion cells. *Vision Research*, *47*(7), 946–954. http://doi.org/10.1016/j.visres.2006.12.015

Gross, J., Baillet, S., Barnes, G. R., Henson, R. N., Hillebrand, A., Jensen, O., … Schoffelen, J. (2013). Good practice for conducting and reporting MEG research. *NeuroImage*, *65*, 349–363. http://doi.org/10.1016/j.neuroimage.2012.10.001

Herbst, K., Sander, B., Milea, D., Lund-Andersen, H., & Kawasaki, A. (2011). Test–retest repeatability of the pupil light response to blue and red light stimuli in normal human eyes using a novel pupillometer. *Frontiers in Neurology*, *2*. http://doi.org/10.3389/fneur.2011.00010

Hoffmann, S., & Falkenstein, M. (2008). The correction of eye blink artefacts in the EEG: A comparison of two prominent methods. *PLoS ONE*, *3*(8). http://doi.org/10.1371/journal.pone.0003004

Hupé, J., Lamirel, C., & Lorenceau, J. (2009). Pupil dynamics during bistable motion perception. *Journal of Vision*, *9*(7). http://doi.org/10.1167/9.7.10

Jainta, S., Vernet, M., Yang, Q., & Kapoula, Z. (2011). The pupil reflects motor preparation for saccades - even before the eye starts to move. *Frontiers in Human Neuroscience*, *5*. http://doi.org/10.3389/fnhum.2011.00097

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585. http://doi.org/10.1126/science.154.3756.1583

Knapen, T., Gee, J. W. D., Brascamp, J., Nuiten, S., Hoppenbrouwers, S., & Theeuwes, J. (2016). Cognitive and ocular factors jointly determine pupil responses under equiluminance. *PLOS ONE*, *11*(5), e0155574. http://doi.org/10.1371/journal.pone.0155574

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*(2), 291–300. http://doi.org/10.1097/aud.0b013e3182310019

Laeng, B., & Sulutvedt, U. (2014). The eye pupil adjusts to imaginary light. *Psychological Science*, *25*(1), 188–197. http://doi.org/10.1177/0956797613503556

Maess, B., Schröger, E., & Widmann, A. (2016). High-pass filters and baseline correction in M/EEG analysis. Commentary on:“How inappropriate high-pass filters can produce artefacts and incorrect conclusions in ERP studies of language and cognition”. *Journal of*

Neuroscience Methods, *266*, 164–165. http://doi.org/10.1016/j.jneumeth.2015.12.003

Mathôt, S. (2013). *A Simple Way to Reconstruct Pupil Size During Eye Blinks*. Retrieved from http://dx.doi.org/10.6084/m9.figshare.688001

Mathôt, S., & Van der Stigchel, S. (2015). New light on the mind's eye: The pupillary light response as active vision. *Current Directions in Psychological Science*, *24*(5), 374–378. http://doi.org/10.1177/0963721415593725

Mathôt, S., Dalmaijer, E., Grainger, J., & Van der Stigchel, S. (2014). The pupillary light response reflects exogenous attention and inhibition of return. *Journal of Vision*, *14*(14), 7. http://doi.org/10.1167/14.14.7

Mathôt, S., Melmi, J. B., & Castet, E. (2015). Intrasaccadic perception triggers pupillary constriction. *PeerJ*, *3*(e1150), 1–16. http://doi.org/10.7717/peerj.1150

Mathôt, S., Siebold, A., Donk, M., & Vitu, F. (2015). Large pupils predict goal-driven eye movements. *Journal of Experimental Psychology: General*, *144*(3), 513–521. http://doi.org/10.1037/a0039168

Mathôt, S., van der Linden, L., Grainger, J., & Vitu, F. (2013). The pupillary response to light reflects the focus of covert visual attention. *PLoS ONE*, *8*(10), e78168. http://doi.org/10.1371/journal.pone.0078168

Mathôt, S., van der Linden, L., Grainger, J., & Vitu, F. (2015). The pupillary light response reflects eye-movement preparation. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(1), 28–35. http://doi.org/10.1037/a0038653

Murphy, P. R., Moort, M. L. V., & Nieuwenhuis, S. (2016). The pupillary orienting response predicts adaptive behavioral adjustment after errors. *PLOS ONE*, *11*(3), e0151763. http://doi.org/10.1371/journal.pone.0151763

Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: Insights from pupillometry. *The Quarterly Journal of Experimental Psychology*, *60*(2), 211–229. http://doi.org/10.1080/17470210600673818

Privitera, C. M., Renninger, L. W., Carney, T., Klein, S., & Aguilar, M. (2010). Pupil dilation during visual target detection. *Journal of Vision*, *10*(10), e3. http://doi.org/10.1167/10.10.3

Reimer, J., Froudarakis, E., Cadwell, C. R., Yatsenko, D., Denfield, G. H., & Tolias, A. S. (2014). Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *Neuron*, *84*(2), 355–362. http://doi.org/10.1016/j.neuron.2014.09.033

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, *52*(8), 997–1009. http://doi.org/10.1111/psyp.12437

Tanner, D., Norton, J. J., Morgan-Short, K., & Luck, S. J. (2016). On high-pass filter artifacts (they're real) and baseline correction (it'sa good idea) in ERP/ERMF analysis. *Journal of Neuroscience Methods*, *266*, 166–170. http://doi.org/10.1016/j.jneumeth.2016.01.002

Tsukahara, J. S., Harrison, T. L., & Engle, R. W. (2016). The relationship between baseline pupil size and intelligence. *Cognitive Psychology*, *91*, 109–123. http://doi.org/10.1016/j.cogpsych.2016.10.001

Wilhelm, H., Lüdtke, H., & Wilhelm, B. (1998). Pupillographic sleepiness testing in hypersomniacs and normals. *Graefe's Archive for Clinical and Experimental Ophthalmology*, *236*(10), 725–729. http://doi.org/10.1007/s004170050149

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials (ERPs) in studies of perception and attention. *Attention, Perception & Psychophysics*, *72*(8), 2031–2047. http://doi.org/10.3758/app.72.8.2031