Standardizing measurement in psychological studies: On why one second has different value in a sprint versus a marathon

Stephanie C. Goodhew¹ · Amy Dawel¹ · Mark Edwards¹

Published online: 27 April 2020
© The Psychonomic Society, Inc. 2020

Abstract

Here we highlight the importance of considering relative performance and the standardization of measurement in psychological research. In particular, we highlight three key analytic issues. The first is the fact that the popular method of calculating difference scores can be misleading because current approaches rely on absolute differences, neglecting what proportion of baseline performance this change reflects. We propose a simple solution of dividing absolute differences by mean levels of performance to calculate a relative measure, much like a Weber fraction from psychophysics. The second issue we raise is that there is an increasing need to compare the variability of effects across studies. The standard deviation score (SD) represents the average amount by which scores differ from their mean, but is sensitive to units, and to where a distribution lies along a measure even when the units are common. We propose two simple solutions to calculate a truly *standardized* SD (SSD), one for when the range of possible scores is known (e.g., scales, accuracy), and one for when it is unknown (e.g., reaction time). The third and final issue we address is the importance of considering relative performance in applying exclusion criteria to screen overly slow reaction time scores from distributions.

Keywords Measurement · Difference scores · Relative performance · Validity · Reaction time · Methodology · Experimental · Correlational

Imagine a scenario in which an athlete running the 100 m sprint beats a world record run by one second. Now compare this with an athlete competing in a 42 km marathon, who also shaves one second off a world record. They are both phenomenal achievements – but even without doing any math, it is obvious that beating a record by one second is more spectacular in the sprint than the marathon. This is true even though both athletes made the same *absolute* change in performance – one second. How they differ is instead in their *relative* performance (relative to total run time) – one second is a much larger proportion of a ten-second sprint time than of a two-hour run time. Analogously, in psychological science, many of the widely used bread-and-butter measures quantify absolute performance or absolute performance changes, neglecting relative performance, even where this is clearly important and at times even a more appropriate quantification. The purpose of this paper is to highlight three key issues with current analysis

approaches in psychology, and propose simple but effective solutions to them. It is intended for an audience of psychological researchers using these approaches, and no highly specialized or sophisticated knowledge of statistics or math is assumed. The aim is to be as user-friendly as possible.

In the first section, we consider the issue of difference scores. That is, a widely used approach in experimental psychology is to subtract a performance metric (e.g., response time) in one condition from that of another condition, to quantify the change in performance resulting from an experimental manipulation. However, it can be misleading to focus on these absolute changes in performance, without factoring in how these changes relate to mean levels of performance. We compare and contrast the advantages and disadvantages of the relative versus absolute approach to quantifying performance change in different contexts. In the second section, we highlight how the ubiquitous descriptive statistic *standard deviation* is *not* actually standardized such that it can be compared across different samples, studies, or metrics, and how relying on the absolute standard deviation can be misleading with respect to the spread of scores on different measures or even different samples on the same measure. We propose a solution: the *standardized standard deviation*, which can be

✉ Stephanie C. Goodhew
Stephanie.Goodhew@anu.edu.au

¹ Research School of Psychology, The Australian National University, Canberra 2601, Australia

calculated in two different ways depending on whether or not a measure is bounded. Finally, in the third section we address the importance of considering relative performance in a specific data-analytic situation: when applying exclusion criteria to screen outlier reaction time scores from distributions.

Relative versus absolute difference score measures

Difference scores, in which performance in one condition is subtracted from performance in another condition, are ubiquitous in experimental psychology. When response/reaction time (RT) to detect, identify, or otherwise respond to a stimulus is measured, it is very common for the absolute difference in performance between two conditions to be quantified¹. For example, in the classic Stroop effect, RTs to name the color of the ink on trials in which the word and color conflict (e.g., “RED” printed in blue) are compared against RTs on trials for non-word blocks of color (MacLeod, 1991; Stroop, 1935). Specifically, the Stroop effect is quantified as the difference in RT between these two conditions. Similarly, in the fundamental spatial attentional cueing paradigm, RT on valid trials (in which a spatial cue appears in the same location as the subsequent target) is subtracted from performance on invalid trials (in which a spatial cue appears in a different location from that of the subsequent target) in order to yield spatial-cueing magnitudes (Posner, 1980). This basic cueing effect forms the core basis of more recently introduced measures of other attentional phenomena, such as contingent capture (Awh, Belopolsky, & Theeuwes, 2012; Becker, 2010; Becker, Folk, & Remington, 2013; Folk, Remington, & Johnston, 1992), and spatial maps of the distribution of attention created from cueing effects in inhibition-of-return (Bennett & Pratt, 2001; Lawrence, Edwards, & Goodhew, 2018; Wilson, Lowe, Ruppel, Pratt, & Ferber, 2016). Furthermore, the dot-probe task, which is widely used both as a measure (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, & van Ijzendoorn, 2007; Cox, Christensen, & Goodhew, 2017; MacLeod, Mathews, & Tata, 1986) and as the basis of treatment for attentional biases in anxiety (Clarke et al., 2017; MacLeod & Clarke, 2015), is founded on difference scores comparing RTs in two experimental conditions. The importance of difference scores is difficult to overstate, and is not limited to RT measures. For instance, difference

scores are also at the heart of important and widely used accuracy-based measures such as the attentional blink (Dell'Acqua et al., 2015; Dux & Marois, 2009; Raymond, Shapiro, & Arnell, 1992), emotion-induced blindness (Most, Chun, Widders, & Zald, 2005; Most, Smith, Cooter, Levy, & Zald, 2007; Onie & Most, 2017; Proud, Goodhew, & Edwards, 2020), and object-substitution masking (Di Lollo, Enns, & Rensink, 2000; Enns & Di Lollo, 1997; Goodhew, 2017; Guest, Gellatly, & Pilling, 2012; Lleras & Moore, 2003). This illustrates the ubiquity of the difference score methodology in experimental approaches to psychology.

Issues with the reliability of difference scores have been raised (e.g., Edwards, 2001; Hedge, Powell, & Sumner, 2018; Onie & Most, 2017). However, it has been argued elsewhere that some of these issues with reliability may actually stem from them controlling for highly stable individual differences in generic task performance, leaving less variance to be explained by experimental manipulations (Goodhew & Edwards, 2019). From this point of view, the relatively low reliability may actually be meaningful, rather than a bugbear that ought to be eliminated. Regardless of how they are conceptualized, however, these reliability issues have not deterred the widespread use of the difference score in experimental psychology. The issue that we raise here, however, is even more fundamental than reliability, and actually speaks to the validity of how difference scores are currently conceptualized. **That is, Issue #1: at present, difference scores are almost invariably quantified in terms of *absolute* difference, whereas for both experimental and hybrid experimental-correlational approaches to studying psychology, the *relative* difference in performance between conditions can be important.**

To illustrate the importance of relative difference scores, consider an example in which a researcher presents participants with Navon stimuli (Navon, 1977), which are compound stimuli in which a global stimulus (e.g., the letter “H”) is made up of multiple instances of a given local element (e.g., multiple letter “E”s make up the H). In the undirected Navon, participants’ task is to identify a given letter (e.g., identify whether T or H is present in the stimulus) as quickly and accurately as possible. RT for trials when the target appears at the global level is subtracted from RT for which the target appears at the local level in order to yield a measure of global preference, such that a larger positive value indicates a greater tendency toward the global level of processing (Gable & Harmon-Jones, 2008; Goodhew & Plummer, 2019; McKone et al., 2010). Consider a hypothetical scenario in which global preference scores for experimental condition A are 100 ms, whereas for experimental condition B they are 200 ms. According to the standard absolute difference score logic, on this basis one should conclude that there is a greater global preference under experimental condition B than experimental condition A. However, these are absolute difference

¹ We note that RT is almost always measured in conjunction with accuracy, and that RTs are typically “correct RTs”, that is, RTs on trials in which participants respond correctly. We also discuss RTs assuming that this is the primary dependent variable from an experiment, and that accuracy is recorded for the purposes of checking task compliance and assessing for the presence of speed/accuracy trade-offs. Hence, when we talk about RT differences, we assume that accuracy scores for these conditions were either unchanged (e.g., at ceiling), or occurred in the convergent direction with the RT changes (e.g., increase in RT and concomitant decrease in accuracy).

scores. If these absolute difference scores arose from markedly different raw RTs (e.g., say that average RT was 1000 ms in condition A and 2000 ms in condition B), then the absolute difference scores fail to capture an important aspect of measurement – how much the experimental manipulation alters performance relative to an existing baseline. By considering the difference scores in relation to overall RT, we now see that both difference scores actually represent an equal proportion of overall RTs. Specifically, we could calculate *relative* difference score by dividing the absolute difference score by mean RT. That is, the relative difference score for condition A is $100 \text{ ms} / 1000 \text{ ms} = 0.1$ or 10%, and the relative difference score for condition B is $200 / 2000 \text{ ms} = 0.1$ or 10%. In other words, the relative difference scores for the two conditions are actually equivalent. In this scenario, the absolute difference scores would lead to the conclusion that the global preference is stronger in experimental condition B, whereas when accounting for these changes relative to mean performance level, the analysis indicated that global trials were facilitated by equal proportional magnitudes in both conditions. The relative difference score clarified that in relative terms, there was no difference in the effect of the manipulation. In other words, **the solution to Issue #1 is to use relative difference scores, in which absolute difference scores are divided by mean RT.** This process of calculating a relative difference score should be applied to each individual participant's difference score or scores, and then these relative difference scores are collated for statistical analysis.

Absolute difference scores can lead to similar interpretational issues in hybrid experimental-correlational approaches, in which the effect of an experimental manipulation is compared across individuals. Returning to our global preference example, imagine a researcher is interested in comparing whether individuals high in openness to experience personality characteristics (group A) have a larger global preference than individuals low in openness to experience (group B). Typically, here, absolute difference scores would be employed. That is, say the researcher finds group A has a larger absolute global preference effect (100 ms) than group B (50 ms). This would typically facilitate the conclusion that group A is more globally inclined than group B. However, if these groups differed in overall RT, then once again these absolute differences could be misleading. For example, if group A's raw RTs were 1050 and 950 ms for the local and global trials respectively, and group B's raw RTs were 275 and 225 for the local and global trials respectively, then the mean RT for group A is 1000 ms, and mean RT for group B is 250 ms. Therefore, the relative RT difference for group A is $100 / 1000 \text{ ms} = 0.1$ or 10%, and the relative RT difference for group B is $50 / 250 \text{ ms} = 0.2$ or 20%. Here, therefore, the *relative* performance difference between the conditions is actually greater for group B than group A – contrary to the conclusion from the absolute difference scores. Note that we

are not arguing that the absolute difference score does not also provide an important source of information, and in some instances may even be the preferred analytic approach. We will discuss the relative advantages and disadvantages in different contexts at the end of this section. For now, however, it is important to see that absolute versus relative performance changes can provide quite different pieces of information and support divergent conclusions.

To be clear that we are not simply dealing with hypothetical examples, consider the study entitled “Enhanced Facilitation of Spatial Attention in Schizophrenia”, which found larger absolute difference scores between valid and invalid RTs and concluded that individuals with schizophrenia have superior spatial attentional capacities than individuals without schizophrenia (Spencer et al., 2011). If true, this would represent an exciting advance in our understanding of a condition that is often characterized by perceptual and cognitive deficits. However, here the absolute differences in RT between the experimental conditions fail to consider the substantial differences in overall RT between the groups. Spencer et al. (2011) do not report the raw data. However, the means are generally estimable from their graphs. For example, if we examine the median RTs from experiment 1, at the 100 ms SOA, we can see that the absolute cueing magnitude for the schizophrenia group is about 22 ms, compared with about 14 ms for the control group – a considerable difference. However, the schizophrenia group is considerably slower overall. In fact, the absolute difference score derives from two RTs, one approximately 570 and the other 582 (average = 576) for the clinical group, and two RTs, one approximately 410 and the other 424 (average = 417) for the control group. Therefore, the *relative* difference scores are about 3.8% ($22/576$) and about 3.3% ($14/417$) for the clinical and control groups respectively. In other words, the relative difference scores are virtually identical – suggesting that the extent to which spatial attention was oriented *relative to overall performance for that group* was actually equivalent in the two groups, in contrast to the absolute measure, which indicated an advantage for the clinical group. This is an example of how factoring in relative performance can qualitatively alter the conclusions drawn from the same data in a real study.

We wish to emphasize that we are not the first to raise this issue with absolute difference scores. Indeed, it has been noted elsewhere in individual-difference studies (e.g., Behrmann et al., 2006; Lawrence et al., 2018; Mickleborough, Hayward, Chapman, Chung, & Handy, 2011). However, even in fields involving the comparison of clinical and control groups where these differences are robust and widely observed, there are neither standard techniques to address them, nor even a consensus understanding that there are issues to be addressed. When it comes to experimental rather than correlational studies, the issue is virtually never considered. Therefore, a clear explanation of these issues and a

standardized approach to addressing them is called for. For this purpose, we suggest the relative difference score. While we have focused on RT differences in these examples, the same logic applies for other dependent variables, such as accuracy.

Considering measurement principles from cognate areas reveals that the logic of the relative difference score has already been adopted elsewhere. Readers familiar with Weber fractions from psychophysics will recognize that the approach we propose here applies the same fundamental logic. That is, in psychophysics, it has long been understood that humans' ability to detect differences scales as a function of stimulus magnitude. A Weber fraction, therefore, quantifies a perceptual difference as a function of stimulus magnitude. Put simply, this is a way of quantifying the fact that a 1 cm increase in the size of a stimulus is easy to detect for a stimulus that was originally 1 cm, whereas a 1 cm increase in the size of a stimulus is difficult to detect for a stimulus that was originally 1000 cm (Fechner, 1966 [first published 1860]). Weber fractions have been used for over a century and a half in psychophysics to quantify psychologically-meaningful changes in performance. A relative difference score is essentially a Weber fraction.

Similarly, readers familiar with prospect theory from judgment and decision-making will know that humans recognize that the same monetary difference (e.g. losing \$100) has psychologically very different consequences depending on one's starting point (e.g., whether one started with \$100 or \$1,000,000) (Kahneman & Tversky, 1979). In the same vein, a relative difference score is also sensitive to one's starting point in performance. Instead of merely quantifying an absolute difference in performance, a relative difference score considers this as a function of the overall level of performance.

An astute reader might wonder why we have not recommended adoption of standardized effect sizes, which could to some extent also provide a solution to the issues that arise from absolute difference scores. While we agree with that approach in principle, the relative difference score affords a number of key advantages. First and foremost, a relative difference score is intuitively easy to comprehend – it is straightforward to understand that the difference between two conditions is $X\%$ of overall performance. Standardized effect sizes quantify the magnitude of effects in the units of SDs , which can be more abstract to conceptualize, holding less intuitive appeal. We suspect that this is at least one of the reasons why their uptake has not been as rapid and widespread as proponents might have hoped. (While it is becoming increasingly common to see them reported, especially as a number of outlets stipulate this as a requirement, it is far less common to see them meaningfully *interpreted*.) Second, effect sizes are designed to be calculated on summary data (e.g., on mean differences averaged over all participants). This means that a non-trivial issue still remains – is X standardized effect size

significantly larger than Y standardized effect size? A standardized effect size is still susceptible to sampling and measurement error just like all other behavioral measures, and so it cannot necessarily be assumed that, for example, an effect size of 0.5 is *reliably* larger than one of 0.4. Therefore, statistical testing is required to determine whether or not an observed difference is unlikely to be due to chance. Of course, there are possible statistical solutions to such issues for standardized effect sizes, but this would often take researchers beyond the comfort of statistics that are highly familiar and routinely used, and that enjoy a consensus of being considered normative in their field of research. In contrast, a relative difference score is calculated at the individual participant level. Relative difference scores can then become the primary dependent variable, and therefore the subject of standard statistical testing such as ANOVAs, t -tests, and correlations. For instance, one could use a simple independent-samples t -test to determine whether the sample of relative difference scores for conditions A and B from a hypothetical study 1 are reliably different from those of study 2, a solution that is not readily available when confronted with the single standardized effect size value for study 1 versus study 2.

Finally, we wish to clarify that what we are offering here is intended to broaden and enhance the armory of tools from which psychological scientists can draw in the pursuit of new knowledge about the human mind, brain, and behavior. We are not advocating uncritical adoption of a relative approach over an absolute performance measure in all instances. Instead, a choice of whether a relative or absolute performance measure is more informative is required by the researcher, motivated by judicious consideration of the psychological process being studied, its operationalization, and the research question. What we wish to prevent is the uncritical adoption of an absolute performance measure where a relative one is more appropriate and informative. In the following section, we provide some guidelines regarding when absolute difference scores may be of the greatest utility, and when relative difference scores may be more useful.

In many contexts, both absolute and relative performance and performance changes are important, and therefore absolute and relative difference scores may provide two distinct sources of information. However, there are certain contexts in which one or the other may be the preferred measure. Absolute difference scores may be the preferred measure when the difference between conditions in absolute terms is more important than how this difference relates to overall performance. We can see this as being more likely to be the case the more meaningful that absolute values on the dependent variable are. For example, when the dependent variable is in units such as IQ points, which have a known distribution and scale, then an absolute difference score may be an important consideration. Here, knowing that after an intervention, the treatment group scored five IQ points higher than their

baseline, whereas the control group's IQ score changed by only one IQ point, could be useful information, even when it is not referenced as proportional to the group's initial IQ scores. In essence, this amounts to considering whether or not there is a *functionally significant* change on the measure, which absolute difference scores can convey if the measure has meaningful units. In a similar vein, consider a scenario in which a researcher is testing the efficacy of a psychological intervention for anxiety. Participants are randomly assigned either to receive the intervention or not, and their anxiety is measured via a self-report anxiety questionnaire pre- and post-intervention. Here, an important outcome may be whether an individual moved from a questionnaire score considered diagnostic of generalized anxiety disorder, to a lower score that would not be considered within the clinical range. Simply knowing that a person's score improved by X% following intervention relative to baseline would not provide this insight, whereas their *absolute* score would.

In contrast, when the units of the dependent variable have less intrinsic value or widely understood meaning (e.g., points on a questionnaire where the units are arbitrary and not linked to a clinical cutoff), then changes quantified relative to baseline may be more informative. The knowledge that group A changed by three points and group B by six points on a self-reported measure of attentional control whose distribution is not widely known may be less useful information than the percentage change relative to their baseline scores on this measure (e.g., a 5% versus a 10% improvement in attentional control). More broadly, relative scores may be of greatest utility when the research question relates to the *extent to which two different groups of participants are affected by an experimental manipulation, and the groups have large pre-existing differences on the outcome measure*. Here, if the absolute change in the outcome measure does not have intrinsic value or accepted meaning, then the information regarding how the groups fared relative to their baseline can provide important insight into the effect of the manipulation on a group's performance.

Standardized standard deviations (SSDs)

The second key issue in the analysis of experimental data in psychology is the need for a standardized measure of variation in scores. A ubiquitous descriptive statistic in psychological measurement is the standard deviation (SD)², which reflects the average amount by which scores deviate from the mean of a distribution. However, unlike a measure such as a correlation coefficient, whose value can be meaningfully interpreted irrespective of the underlying units (e.g., Pearson's $r = 0.9 =$

strong positive correlation), SDs are not actually truly standardized. That is, SDs are not unit-invariant, and even on the same units, SDs are sensitive to the mean of the distribution (i.e., sensitive to the relative placement of scores along a variable). This is **Issue #2: The value of SDs is not standardized, such that it is not always the case that a measure with a larger SD necessarily has a larger spread of scores than a measure with a smaller SD**. This is illustrated in Table 1. Here, the same distribution of 10 hypothetical values of individuals' heights is shown in centimeters (cm), meters (m), and inches. The means and SDs of these scores are shown at the bottom of the table. Even though these are the identical distribution of scores simply appearing in different units, the SDs change according to the units in which the original scores appear.

SDs are one of the most common summary statistics used to describe a sample in psychological research, a close second perhaps only to measures of central tendency, such as the mean. An understanding of what SDs truly reflect and the limits on their interpretability is therefore of the utmost importance. Furthermore, the need for a standardized measure of the spread of scores is only increasingly in demand. This is because, while there are conflicting criteria for selecting the most suitable tasks for experimental versus correlational studies (Goodhew & Edwards, 2019; Hedge et al., 2018), what they share in common is a need to quantify the spread of scores in a standardized way. That is, as emphasized in recent work (Goodhew & Edwards, 2019; Hedge et al., 2018), when the goal of research is to study the effect of experimental manipulations and their mean effect for a group of participants, the most sensitive paradigms are those with the *lowest* spread when the magnitude of the effect is considered as a

Table 1 The same 10 hypothetical scores displayed in three different units. Despite these scores being identically distributed, and therefore having identical spread, the value of the SDs changes according to the units of the scores. This means that we cannot use SD alone to determine the spread of scores on a measure. Instead, a standardized measure of SD is called for. Here, we propose the standardized standard deviation (SSD)

	Height in cm	Height in m	Height in inches
Score #1	150	1.50	59.06
Score #2	155	1.55	61.02
Score #3	160	1.60	62.99
Score #4	165	1.65	64.96
Score #5	170	1.70	66.93
Score #6	175	1.75	68.90
Score #7	180	1.80	70.87
Score #8	185	1.85	72.83
Score #9	190	1.90	74.80
Score #10	195	1.95	76.77
Mean:	172.50	1.73	67.91
SD:	15.14	0.15	5.96

² We remind the reader that $SD^2 = \text{Variance}$, and therefore variance is susceptible to the units of the distribution of scores in the same way as SD.

distribution. In other words, the experimental manipulation should affect all participants to similar degrees. For example, the Stroop effect fits this category because most individuals experience it, and to similar degrees (Goodhew & Edwards, 2019; Hedge et al., 2018). This fits the mindset of most experimental psychologists in seeking robust manipulations that have strong and consistent magnitude effects.

In contrast, when the goal of the study is to determine how individuals or groups differ (either on an existing measure or in response to an experimental manipulation), the most sensitive variables are those with the *highest* spread of scores (Goodhew & Edwards, 2019; Hedge et al., 2018). Where the variable of interest is the extent to which individuals are influenced by an experimental manipulation, the most sensitive measures will be those that affect individuals to very different degrees (e.g., the attentional blink, since some individuals do not experience the blink at all, whereas others experience a deep blink) (Martens & Wyble, 2010). While the two types of research call for different magnitudes of spread in scores, what they share in common is a need to quantify the spread of scores in a distribution. Specifically, it is the need to compare magnitudes of spread in scores across studies and across different dependent variables in order to determine whether this variation is high or low. In other words, a standardized measure of the spread of scores is required.

On first glance, it might seem logical to consult the SD of an effect to determine whether the spread of scores is high or low. This is because the SD represents how much scores differ on average from the mean score of the distribution. However, as outlined above, it is problematic to compare SDs across studies that employ different units. For example, we could not meaningfully compare the SD of RT dot-probe emotional-bias scores with the SD of accuracy-based emotion-induced blindness scores to determine which paradigm has the largest between-participant variation. Indeed, this is not just a hypothetical conundrum, because the comparison of these two different types of scores and their ability to predict self-reported negative affect has been a focus of recent literature (e.g., Onie & Most, 2017). This means that relying on absolute SD values could be misleading with respect to the true spread of scores across different measures or samples.

SSD for bounded measures

The issue with SDs extends beyond unit specificity. Even if we are comparing two measures or tasks that use the same dependent variable with the same units (e.g., RT, ratings), SDs do not necessarily scale as a proportion of the spread in the distribution of scores. A real example arises from the highly popular self-report measures of depressive symptomology, the Depression scale from the Depression Anxiety Stress Scales (DASS), and the Beck Depression Inventory (BDI). In a widely cited (6200+ at the time of writing) paper

(Lovibond & Lovibond, 1995), the SD for both the BDI and the Depression scale of the DASS was reported to be 6.5. Does this mean the two measures produced the same spread in scores? No, because these SDs arise from scales that have different ranges of possible scores. This means that it is not the absolute magnitude of the SD that is informative, but instead how this absolute SD compares with the *total possible range of scores*. For scales such as these, the total possible range of scores is known – the difference between the minimum and maximum possible score on the measure. Therefore, we propose that to standardize SD for a bounded measure, **standardized standard deviation (SSD) = SD / total possible range of scores**. This means that the SSD is essentially a proportion score (i.e., varies between 0 and 1), which can easily be converted into a percentage (0–100%), providing intuitive interpretational value. Furthermore, conceptually this SSD score makes logical sense. It is essentially a ratio of the observed SD to the total range of scores. Where an absolute SD is truly large, it will represent a large portion of the total possible range of scores, yielding a high SSD value. Where an absolute SD is small such that scores are tightly clustered, absolute SD will represent a small portion of the total range of scores, thereby yielding a low SSD value.

Standardizing SD against the possible range of scores now provides a unit-invariant measure of the variance of the scores. For the BDI, the maximum possible score is 63, and the minimum 0, so the possible range is 63. For the Depression (D) scale from the DASS, the maximum possible score is 42, and the minimum 0, so the possible the range is 42. Therefore, BDI's $SSD = (6.5/63) \times 100 = 10\%$, whereas D's $SSD = (6.5/42) \times 100 = 15\%$. This means that the spread of scores is greater in the D scale from the DASS than the BDI, making it better suited to the goals of individual-differences research (i.e., facilitating the ability to observe correlations with other variables). This highlights the importance of considering standardized standard deviation in measure selection.

This solution to the issue of non-standardized SSDs is simple and feasible when the total range of possible scores is known, such as a questionnaire in which we know the minimum and maximum possible scores that can be obtained (i.e., bounded scores). Accuracy scores also share this property – typically the minimum possible score is 0% and the maximum is 100%. Of course, the likely range of scores is often less. For example, in a two-alternative forced-choice (2AFC) task, a participant could score 50% purely by guessing (i.e., chance-level performance is 50%). This could therefore represent a *likely* minimum score. However, participants can (and do) score below chance-level performance for various reasons, and so the total *possible* is still 0–100%. If, however, all scores that fall below chance are excluded from further analysis (a common and justifiable exclusion), then the effective range would be chance-ceiling. For example, for a 2AFC, it would be $100 - 50 = 50\%$, whereas for a 4AFC it would be $100 - 25 = 75\%$.

SSD for unbounded measures

While using the theoretical possible range in order to obtain an SSD is simple and feasible for distributions of scores where the range is known, in experimental psychology we often employ unbounded measures for which the total possible range is not known, or is theoretically infinite – for example, RT. Consider a scenario in which the SD of a distribution of RTs to *detect* a stimulus is 50 ms, and the SD of a distribution of RTs to *identify* a stimulus is 100 ms. Is the spread of scores larger for the second distribution than the first? Not necessarily. Detection RTs are likely to be quicker, and so conceivably mean RT for the first distribution could be 200 ms, whereas say the RTs from the second distribution arise from a visual search task, then conceivably mean RT could be 1200 ms. While the absolute SD is larger for the second distribution, the spread of scores is actually smaller relative to the mean of the distribution, and therefore the absolute SD could be misleading. Instead, in order to meaningfully quantify SD in a way that we can compare across these different distributions, they need to be *standardized*.

What is the best way to standardize a measure for which the bounds are unknown? One could use the *observed* range of scores. However, these tend to be very sample-specific (i.e., they are very sensitive to single scores), and therefore are not a good candidate for standardization across studies. While the comparison of SDs is problematic when distributions have different means, they are comparable *when the distribution of scores has the same mean* (assuming the distributions have comparable shapes). Since an SD represents the average amount by which scores in the distribution deviate from the mean, a meaningful comparison between them can be made when the scores have the same mean. Where the goal is to compare distributions that do not originally have the same

mean, a transformation is required such that they do. This is illustrated in Table 2.

For unbounded measures, a solution to arrive at an SSD is to multiply scores in one distribution by a given factor so that both distributions have the same mean, and then the SDs are directly comparable. Often, one might have access only to the summary statistics (e.g., mean, SD) rather than the original raw scores for published data. In this instance, can the mean and SD themselves be used to calculate an SSD? The answer is yes – and the scores in Table 2 illustrate this. The second distribution in each highlighted pair represents half the values of the first pair. Even without knowing the original scores, can we simply determine how the two means relate to one another, and standardize the SD accordingly? Yes. Consider distribution 1A and 1B. We can see that the $\text{Mean}_{1A} = 2 \times \text{Mean}_{1B}$. The SDs have this same relationship. That is, the $\text{SD}_{1A} = \text{SD}_{1B} \times 2$. If we wanted to compare distribution 1A and 1B and only had their summary statistics, once we have established what factor the second mean needs to be multiplied by to equal the first mean, this factor can be applied to the SD. In other words, in order to enable a comparison between distributions 1A and 1B, we calculate $\text{SD}_{1B} \times 2 = 316.23$, that is, the same SD as distribution 1A, so we can see that they have the same spread of scores (which makes sense, since they are the same distribution, just on different scales). Does this also work across distinct distributions? Let's say we wanted to compare distribution 1A with distribution 2B. We know that distribution 2B has greater spread. Does SSD reveal this? We can see that $\text{Mean}_{1A} = 2 \times \text{Mean}_{2B}$. That is, our multiplicative factor is 2. Therefore, $\text{SSD}_{2B} = \text{SD}_{2B} \times 2 = 353.56$. So, yes, this shows that the SSD_{2B} , which we can now directly compare in magnitude with SD_{1A} , is greater, as would be expected given that the spread of scores in distribution 2B is greater.

Table 2 shows six distributions of five hypothetical RT scores (in ms), where each highlighted pair (i.e., A and B labels of the same distribution) represents a distribution of scores where their rank ordering and the relative separation between scores is maintained, but they appear on a different scale (i.e., the second distribution = $\frac{1}{2}$ of each of the original scores). As above, it is clear that the SD scales according to the relative placement of the scores, rather than being an invariant measure for the same distribution of scores. Relative to the first pair of distributions, the

second pair has been created to have larger spread, and the third pair of distributions has been created to have a smaller spread (note that all of the scores cluster closer to the mean than the previous distributions). Since distributions 1A, 2A, and 3A have the same mean, their SDs are now comparable, and we can see that SD scales as a function of the true variance of the scores. Similarly, since distributions 1B, 2B, and 3B have the same mean, their SDs are now comparable, and we can see that SD scales as a function of the true variance of the scores.

Score #	Distribution 1A	Distribution 1B	Distribution 2A	Distribution 2B	Distribution 3A	Distribution 3B
1	0	0	0	0	200	100
2	200	100	100	50	300	150
3	400	200	400	200	400	200
4	600	300	700	350	500	250
5	800	400	800	400	600	300
Mean:	400	200	400	200	400	200
SD:	316.23	158.11	353.55	176.78	158.11	79.06

Next, we can consider how this SSD solution works for real distributions of scores. Here, we consider the data from a recent study which examined the efficiency of people's ability to dynamically contract their attention focus (i.e., resize from global to local focus), versus their ability to expand their attention focus (i.e., resize from local to global focus) (Goodhew & Plummer, 2019). (Note that the raw data for these experiments are available in the supplementary material for the Goodhew & Plummer paper, for interested readers). Experiment 2 revealed that people were more efficient at expanding their attention (expansion cost RT = RT for infrequent trials requiring a global focus minus RT for the frequent local trials) than contracting it (contraction cost RT = RT for infrequent trials requiring a local focus minus RT for the frequent global trials).

There were also substantive (and reliable) individual differences in participants' dynamics of resizing. Some individuals were almost perfectly efficient at resizing their attention with almost zero cost, whereas others took up to half a second to do so. These are considerable individual differences, given that, by comparison, shifts in the spatial location of attention occur on sub-100 ms time frames. However, was there greater variability in individuals' contraction or expansion efficiency? If we look at the SDs, we see that it is larger for contraction cost (127.43) than it is for expansion cost (110.38). However, these costs also have different means (158.20 for contraction cost versus 121.80 for expansion cost). If we calculate mean contraction cost / mean expansion cost = 1.30, we can then calculate SSD expansion cost = SD expansion cost $\times$ 1.30 = 143.50. Comparing the two, we can see that SSD expansion cost (143.50) is larger than SSD contraction cost (127.43), indicating that there is actually greater variability in expansion than in contraction cost.

Relative exclusion criteria for reaction times

The two issues we have discussed so far illustrate how considering only absolute performance can miss an important source of information, and how this can alter or distort conclusions that are drawn from the same set of data. The final issue we address follows this same theme, although it relates to a more specific context than the previous two. Specifically, the third and final issue we raise here is that, when using mean RTs, it is typically necessary to apply exclusion criteria to filter out (1) very rapid anticipatory responses and (2) very prolonged responses that do not reflect the process of interest, but instead reflect spurious effects such as the participant sneezing, or for more audacious participants, checking their phone. Means are sensitive to outliers, and so the inclusion of extreme values that do not reflect the process of interest can obscure or distort genuine effects. Such cutoffs can be absolute (e.g., 100 ms, 1500 ms), or they can be relative (e.g., ± 3

SDs above/below participant's mean RT). **Issue #3: When is it more appropriate to use absolute RT cutoffs and when is it more appropriate to use relative RT cutoffs? We propose that minimum RT cutoffs be absolute, and maximum RT cutoffs be relative.**

The reason for the need to have an absolute minimum RT cutoff is that RT distributions are typically positively skewed. This means that using a relative cutoff (e.g., removing RTs less than 3 SDs below mean RT) will typically yield a *negative* (i.e., not possible) RT cutoff, meaning that there is effectively no minimum RT criterion. The reason for the need to have a relative maximum RT cutoff (e.g., + 3 SDs above participant's mean RT) rather than an absolute one (e.g., 1500 ms) is that participants differ markedly in their RTs. This means that an absolute cutoff will disproportionately affect slower participants, even potentially cutting the majority of their RTs, which are likely to reflect genuine data points. Furthermore, if experimental conditions do differ in mean RT, imposing such an absolute RT cutoff may introduce a confounder that disproportionately affects the slower condition, and also slower participants. A hypothetical example is shown in Tables 3 and 4.

Table 3 illustrates the point of how an absolute cutoff could introduce a confounder by disproportionately excluding trials from the slower condition 2, and Table 4 illustrates how an absolute cutoff could disproportionately affect trials from a slower participant. A relative cutoff applied to *group* data could produce the same problem. Instead, a relative cutoff should generally be applied at the level of the individual *participant*, excluding only their slowest RTs from analysis. There may be some instances where it is also appropriate to apply a relative upper RT cutoff on the basis of blocks or conditions; the key consideration is whether particular units of analysis (e.g., participants or conditions) are going to be disproportionately affected by the cutoff, in such a way that likely genuine responses that do reflect the process of interest are being unnecessarily excluded, and/or responses are excluded in a way that is biased against particular conditions and/or participants.

Table 3. Hypothetical responses from a single participant with six trials in two different conditions, and how they would be affected by an absolute maximum cutoff (in this case, of 1500 ms)

Trial #	Condition 1	Condition 2
1	1000	1700
2	1100	1600
3	900	1550
4	900	1400
5	700	1450
6	800	1300

Table 4. Hypothetical responses to six trials from two participants in a single condition, and how they would be affected by an absolute maximum cutoff (in this case, of 1500 ms)

Trial #	Participant 1	Participant 2
1	1000	1700
2	1100	1600
3	900	1550
4	900	1400
5	700	1450
6	800	1300

We recognize that researchers are being increasingly called upon to choose their method of analysis in advance (e.g., via preregistration). This includes nominating a prescribed method for handling RT outliers. Researchers can be confronted with a range of possible choices in such circumstances. We hope that the issue and solution proposed here will help researchers in making reasoned and informed choices, irrespective of whether these decisions are made a priori or post hoc. While we have not prescribed a particular SD cutoff, others have suggested that *three* SDs represents a good outlier cutoff (Tabachnick & Fidell, 2013). We suggest that 3 SDs represents a reasonable ballpark. However, we also advocate that the cutoff applied be chosen in consideration of the underlying distribution (e.g., its skew), and therefore of how many scores are ultimately removed – it should be a *small* (but non-zero) proportion (e.g., <5%). Ultimately, scientists ought to be interested in robust effects, and robust effects should not depend on whether 2.5 or 3 SDs is chosen as the upper exclusion criterion. In the interest of establishing this, authors could even check and report that results are not qualitatively altered by such minor adjustments.

To be clear, other authors have conducted in-depth and systematic modeling of different methods for handling reaction time outliers (Ratcliff, 1993). It was not our intention to do so here, but instead to highlight a simple yet important conceptual issue with absolute upper RT exclusion criteria that can be solved by considering relative performance. Interestingly, however, our advice converges with that from Ratcliff, who recommends the application of standard deviation cutoffs when variability among participants is high. RTs are typically characterized by considerable variability across participants (for a review, see Goodhew & Edwards, 2019), and we thus concur with this recommendation from Ratcliff (1993).

Furthermore, a reader might be wondering why we are recommending an SD-based upper criterion, after reviewing the issues of non-standardization of SD scores above. The issue of non-standardization, however, arises when one wants

to *interpret the magnitude* of an SD score – for example, if one wants to compare an SD from one study to that obtained in another. This is not the goal of applying an SD-based upper RT cutoff. In fact, the intrinsic linking between an SD and the mean of a distribution is actually a desirable property in this instance. An SSD would be required, however, if one wanted to compare the spread of RT values in one condition with that in another, or for one participant versus another, where the mean RTs differ.

Conclusions

Here we have highlighted three key issues in the analysis of popular measures in psychology, and we have proposed solutions to each. The first issue was that difference scores are typically reported as absolute differences, whereas there are many instances in which relative changes in performance are equally or even more informative. Making absolute differences proportional to overall RT provides an easy and intuitive relative difference score metric. The second issue was that SDs are not truly standardized units and are therefore *not* readily comparable across different studies. We proposed a simple solution where the absolute bounds of possible scores are known (e.g., accuracy or a questionnaire scale) that involves calculating the proportion of SD relative to possible range, and a different solution for when the bounds are unknown (e.g., RT). That is, where the bounds are unknown, one of the distributions is scaled to have the same mean as the other, making the magnitude of the SDs directly comparable. The third and final issue we addressed was determining the most appropriate types of RT exclusion criteria to adopt. We proposed that an absolute minimum RT in conjunction with a relative (to an individual participant's mean RT for a given condition) maximum RT is most likely to maximize the inclusion of genuine and usable data and not disproportionately affect particular participants or experimental conditions. We anticipate that these should be simple and user-friendly solutions to adopt to address important measurement issues in psychological science. However, our primary aim was to foster greater awareness and consideration of these issues, and to stimulate a dialogue with respect to potential solutions. We welcome other researchers proposing alternative solutions to these issues.

Acknowledgements This work was supported by an Australian Research Council (ARC) Future Fellowship (FT170100021) awarded to SCG.

Open Practices Statement There is no data associated with this manuscript beyond that provided in the tables, and therefore all data is already provided. This manuscript is about analysis approaches, rather than reporting the results of an experiment or experiments, and therefore preregistration is not appropriate or applicable.

References

- Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, 16(8), 437–443. <https://doi.org/10.1016/j.tics.2012.06.010>
- Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin*, 133(1), 1–24. <https://doi.org/10.1037/0033-2909.133.1.1>
- Becker, S. I. (2010). The role of target-distractor relationships in guiding attention and the eyes in visual search. *Journal of Experimental Psychology: General*, 139(2), 247–265. <https://doi.org/10.1037/a0018808>
- Becker, S. I., Folk, C. L., & Remington, R. W. (2013). Attentional capture does not depend on feature similarity, but on target-nontarget relations. *Psychological Science*, 24(5), 634–647. <https://doi.org/10.1177/0956797612458528>
- Behrmann, M., Avidan, G., Leonard, G. L., Kimchi, R., Luna, B., Humphreys, K., & Minshew, N. (2006). Configural processing in autism and its relationship to face processing. *Neuropsychologia*, 44(1), 110–129. <https://doi.org/10.1016/j.neuropsychologia.2005.04.002>
- Bennett, P. J., & Pratt, J. (2001). The spatial distribution of inhibition of return. *Psychological Science*, 12(1), 76–80. <https://doi.org/10.1111/1467-9280.00313>
- Clarke, P. J. F., Branson, S., Chen, N. T. M., Van Bockstaele, B., Salemink, E., MacLeod, C., & Notebaert, L. (2017). Attention bias modification training under working memory load increases the magnitude of change in attentional bias. *Journal of Behavior Therapy and Experimental Psychiatry*, 57, 25–31. <https://doi.org/10.1016/j.jbtep.2017.02.003>
- Cox, J. A., Christensen, B. K., & Goodhew, S. C. (2017). Temporal dynamics of anxiety-related attentional bias: is affective context a missing piece of the puzzle? *Cognition & emotion*, 32(6), 1329–1338. <https://doi.org/10.1080/02699931.2017.1386619>
- Dell'Acqua, R., Dux, P. E., Wyble, B., Doro, M., Sessa, P., Meconi, F., & Jolicoeur, P. (2015). The attentional blink impairs detection and delays encoding of visual information: evidence from human electrophysiology. *Journal of Cognitive Neuroscience*, 27(4), 720–735. https://doi.org/10.1162/jocn_a_00752
- Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General*, 129(4), 481–507. <https://doi.org/10.1037/0096-3445.129.4.481>
- Dux, P. E., & Marois, R. (2009). How humans search for targets through time: A review of data and theory from the attentional blink. *Attention, Perception, & Psychophysics*, 71(8), 1683–1700. <https://doi.org/10.3758/APP.71.8.1683>
- Edwards, J. R. (2001). Ten Difference Score Myths. *Organizational Research Methods*, 4(3), 265–287. <https://doi.org/10.1177/109442810143005>
- Enns, J. T., & Di Lollo, V. (1997). Object substitution: A new form of masking in unattended visual locations. *Psychological Science*, 8(2), 135–139. <https://doi.org/10.1111/j.1467-9280.1997.tb00696.x>
- Fechner, G. T. (1966 [First published 1860]). *Elements of psychophysics [Elemente der Psychophysik]* (H. E. Adler, Trans. Vol. 1). United States of America: Holt, Rinehart and Winston.
- Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception & Performance*, 18(4), 1030–1044. <https://doi.org/10.1037/0096-1523.18.4.1030>
- Gable, P. A., & Harmon-Jones, E. (2008). Approach-motivated positive affect reduces breadth of attention. *Psychological Science*, 19(5), 476–482. <https://doi.org/10.1111/j.1467-9280.2008.02112.x>
- Goodhew, S. C. (2017). What have we learned from two decades of object-substitution masking? Time to update: Object individuation prevails over substitution. *Journal of Experimental Psychology: Human Perception and Performance*, 43(6), 1249–1262. <https://doi.org/10.1037/xhp0000395>
- Goodhew, S. C., & Edwards, M. (2019). Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Consciousness and Cognition*, 69, 14–25. <https://doi.org/10.1016/j.concog.2019.01.008>
- Goodhew, S. C., & Plummer, A. S. (2019). Flexibility in resizing attentional breadth: Asymmetrical versus symmetrical attentional contraction and expansion costs depends on context. *Quarterly Journal of Experimental Psychology*, 72(10), 2527–2540. <https://doi.org/10.1177/1747021819846831>
- Guest, D., Gellatly, A., & Pilling, M. (2012). Reduced OSM for long duration targets: Individuation or items loaded into VSTM? *Journal of Experimental Psychology: Human Perception and Performance*, 38(6), 1541–1553. <https://doi.org/10.1037/a0027031>
- Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 50(3), 1166–1186. <https://doi.org/10.3758/s13428-017-0935-1>
- Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263–291. <https://doi.org/10.2307/1914185>
- Lawrence, R. K., Edwards, M., & Goodhew, S. C. (2018). Changes in the spatial spread of attention with ageing. *Acta Psychologica*, 188, 188–199. <https://doi.org/10.1016/j.actpsy.2018.06.009>
- Lleras, A., & Moore, C. M. (2003). When the target becomes the mask: Using apparent motion to isolate the object-level component of object substitution masking. *Journal of Experimental Psychology: Human Perception and Performance*, 29(1), 106–120. <https://doi.org/10.1037/0096-1523.29.1.106>
- Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behav Res Ther*, 33(3), 335–343. [https://doi.org/10.1016/0005-7967\(94\)00075-U](https://doi.org/10.1016/0005-7967(94)00075-U)
- MacLeod, C. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109(2), 163–203. <https://doi.org/10.1037/0033-2909.109.2.163>
- MacLeod, C., & Clarke, P. J. F. (2015). The Attentional Bias Modification Approach to Anxiety Intervention. *Clinical Psychological Science*, 3(1), 58–78. <https://doi.org/10.1177/2167702614560749>
- MacLeod, C., Mathews, A., & Tata, P. (1986). Attentional bias in emotional disorders. *Journal of Abnormal Psychology*, 95(1), 15–20. <https://doi.org/10.1037//0021-843x.95.1.15>
- Martens, S., & Wyble, B. (2010). The attentional blink: Past, present, and future of a blind spot in perceptual awareness. *Neuroscience & Biobehavioral Reviews*, 34(6), 947–957. <https://doi.org/10.1016/j.neubiorev.2009.12.005>
- McKone, E., Davies, A. A., Fernando, D., Aalders, R., Leung, H., Wickramariyaratne, T., & Platow, M. J. (2010). Asia has the global advantage: Race and visual attention. *Vision Research*, 50(16), 1540–1549. <https://doi.org/10.1016/j.visres.2010.05.010>
- Mickleborough, M. J., Hayward, J., Chapman, C., Chung, J., & Handy, T. C. (2011). Reflexive attentional orienting in migraineurs: The behavioral implications of hyperexcitable visual cortex. *Cephalalgia*, 31(16), 1642–1651. <https://doi.org/10.1177/0333102411425864>
- Most, S. B., Chun, M. M., Widders, D. M., & Zald, D. H. (2005). Attentional rubbernecking: Cognitive control and personality in

- emotion-induced blindness. *Psychonomic Bulletin & Review*, 12(4), 654–661. <https://doi.org/10.3758/BF03196754>
- Most, S. B., Smith, S. D., Cooter, A. B., Levy, B. N., & Zald, D. H. (2007). The naked truth: Positive, arousing distractors impair rapid target perception. *Cognition and Emotion*, 21, 964–981. <https://doi.org/10.1080/02699930600959340>
- Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3), 353–383. <https://doi.org/10.1016/0010-0285%2877%2990012-3>
- Onie, S., & Most, S. B. (2017). Two roads diverged: Distinct mechanisms of attentional bias differentially predict negative affect and persistent negative thought. *Emotion*, 17(5), 884–894. <https://doi.org/10.1037/emo0000280>
- Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, 32(1), 3–25. <https://doi.org/10.1080/00335558008248231>
- Proud, M., Goodhew, S. C., & Edwards, M. (2020). A vigilance avoidance account of spatial selectivity in dual-stream emotion induced blindness. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-019-01690-x>
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114(3), 510–532.
- Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, 18(3), 849–860. <https://doi.org/10.1037/0096-1523.18.3.849>
- Spencer, K. M., Nestor, P. G., Valdman, O., Niznikiewicz, M. A., Shenton, M. E., & McCarley, R. W. (2011). Enhanced facilitation of spatial attention in schizophrenia. *Neuropsychology*, 25(1), 76–85. <https://doi.org/10.1037/a0020779>
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643–662. <https://doi.org/10.1037/h0054651>
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics*, 6th Edition: Pearson.
- Wilson, K. E., Lowe, M. X., Ruppel, J., Pratt, J., & Ferber, S. (2016). The scope of no return: Openness predicts the spatial distribution of Inhibition of Return. *Attention, Perception & Psychophysics*, 78, 209–217. <https://doi.org/10.3758/s13414-015-0991-5>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.