Does One Effect Size Fit All? The Case Against Default Effect Sizes for Sport and Exercise Science

Aaron R. Caldwell[1,2] **and Andrew D. Vigotsky**[3]

[1]**U.S. Army Research Institute of Environmental Medicine, Thermal and Mountain Medicine Division, Natick, MA**

[2]**Oak Ridge Institute of Science and Education, Oak Ridge, TN**

[3]**Departments of Biomedical Engineering and Statistics, Northwestern University, Evanston, IL**

Corresponding author:
Aaron R. Caldwell[1]

Email address: aaron.r.caldwell2.ctr@mail.mil

Please cite as:
Caldwell A, Vigotsky AD. 2020. A case against default effect sizes in sport and exercise science. *PeerJ* 8:e10314 https://doi.org/10.7717/peerj.10314

ABSTRACT

Recent discussions in the sport and exercise science community have focused on the appropriate use and reporting of effect sizes. Sport and exercise scientists often analyze repeated-measures data, from which mean differences are reported. To aid the interpretation of these data, standardized mean differences (SMD) are commonly reported as description of effect size. In this manuscript, we hope to alleviate some confusion. First, we provide a philosophical framework for conceptualizing SMDs; that is, by dichotomizing them into two groups: magnitude-based and signal-to-noise based SMDs. Second, we describe the statistical properties of SMDs and their implications. Finally, we provide high-level recommendations for how sport and exercise scientists can thoughtfully report raw effect sizes, SMDs, or other effect sizes for their own studies. This conceptual framework provides sport and exercise scientists with the background necessary to make and justify their choice of an SMD. The code to reproduce all analyses and figures within the manuscript can be found at the following link: `https://www.doi.org/10.17605/OSF.IO/FC5XW`.

INTRODUCTION

Effect sizes are a family of descriptive statistics used to communicate the magnitude or strength of a quantitative research finding. Many forms of effect sizes exist, ranging from mean raw values to correlation coefficients. In sport and exercise science, a standardized mean difference (SMD) is commonly reported in studies that observe changes from pre- to post-intervention, and for which units may vary from study-to-study (e.g., muscle thickness vs. cross-sectional area vs. volume). Put simply, an SMD is any mean difference or change score that is divided, hence standardized, by *a* standard deviation or combination of standard deviations. Thus, even among SMDs, there exist multiple calculative approaches (Lakens, 2013; Baguley, 2009). A scientist must therefore decide which SMD is most appropriate to report for their particular study, or if to report one at all. In this manuscript, we will exclusively be focusing on SMD calculations for studies involving repeated-measures since this is a common feature of sport and exercise science studies; other study designs (i.e., between-subjects) have already been extensively covered elsewhere (Baguley, 2009; Kelley and Preacher, 2012; Hedges, 2008).

Different forms of SMDs communicate unique information and have distinct statistical properties. Yet, some authors in sport and exercise science have staunchly advocated for specific SMD calculations, and in doing so, outright rebuke other approaches (Dankel and Loenneke, 2018). While we appreciate that previous discussions of effect sizes have brought this important topic to the forefront, we wish to expand on their work by providing a deeper philosophical and mathematical discussion of SMD choice. In doing so, we suggest that the choice of an SMD should be based on the objective of each study and therefore is likely to vary from study-to-study. Scientists should have the intellectual freedom to choose

All authors have read and approve this version of the manuscript. The article was last modified on Sept. 14th, 2020. The final, peer-reviewed version can be found on PeerJ. Authors can be reached on Twitter at @exphysstudent (ARC) and @avigotsky (ADV)

whatever statistics are needed to appropriately answer their question. Importantly, this freedom should not be encroached on by broad recommendations that ignore the objectives of an individual scientist. To facilitate these reporting decisions, it is imperative to understand what to report and why.

In this paper, we broadly focus on three things to consider when reporting an SMD. First, before choosing an SMD, a scientist must decide if one is necessary. When making this decision, it is prudent to consider arguments for and against reporting SMDs, in addition to *why* one should be reported. Second, we broadly categorize repeated-measures SMDs into two categories: signal-to-noise and magnitude-based SMDs. This dichotomy provides scientists with a philosophical framework for choosing an SMD. Third, we describe the statistical properties of SMDs, which we believe scientists should try to understand if they are to report them. We relate these perspectives to previous discussions of SMDs, make general recommendations, and conclude by urging scientists to think carefully about what effect sizes they are reporting and why.

SHOULD I REPORT A STANDARDIZED MEAN DIFFERENCE?

Before reporting an SMD—or any statistic for that matter—a researcher should first ask themselves whether it is necessary or informative. When answering this, one may wish to consider arguments both for and against SMDs, in addition to field standards. Here, we briefly detail these arguments, in addition to SMD reporting within sport and exercise science.

Proponents and Opponents of Standardized Effect Sizes

Opponents

Although SMDs may be useful in some contexts, they are far from a panacea. Arguments against the use of SMDs, including those by prominent statisticians, are not uncommon. These arguments should be considered when choosing whether or not to report an SMD. In particular, the evidentiary value of reporting an SMD must be considered relative to the strength of the general arguments against SMDs. Below, we have briefly summarized some of the major arguments against the use of SMDs.

As far back as 1969, the use of standardized effect sizes—and by proxy, SMDs—has been heavily criticized. The eminent statistician John Tukey stated that “only bad reasons seem to come to mind” for using correlation coefficients instead of unstandardized regression coefficients to interpret data. To put it simply, scientists should not assume that standardized effect sizes will make comparisons meaningful (Tukey, 1969). This same logic can also be applied to qualitative benchmarks (e.g., Cohen’s $d = 0.2$ is “small”); we believe it is likely that Cohen would also argue against the broad implementation of these arbitrary benchmarks in all areas of research. Similar arguments against the misuse of standardized effect sizes have been echoed elsewhere (Lenth, 2001; Kelley and Preacher, 2012; Baguley, 2009; Robinson et al., 2003).

Others have outright argued against the use of standardized effect sizes because they oversimplify the analysis of, and distort the conclusions derived from, data. In epidemiology, Greenland et al. (1986) provided a damning indictment of the use of standardized coefficients; namely, because they are largely determined by the variance in the sample, which is heavily influenced by the study design. In psychology, Baguley (2009) offers a similarly bleak view of standardized effect sizes. He argues that the advantages of standardized effect sizes are far outweighed by the difficulties that arise from the standardization process. In particular, scientists tend to ignore the impact of reliability and range restriction on effect size estimates, in turn overestimating the generalizability of standardized effect sizes to wider populations and other study designs (Baguley, 2009).

Proponents

Conversely, prominent statisticians have also argued in favor of standardized effect sizes, especially for facilitating meta-analysis (Hedges, 2008). Cohen (1977) was the first to suggest the use of standardized effect sizes to be useful for power analysis purposes. This is because, unlike the t-statistic, (bias-corrected) standardized effect sizes are not dependent on the sample size. Similarly, while p-values indicate the compatibility of data with some test hypothesis (e.g., the null hypothesis) (Greenland, 2019), SMDs provide information about the ‘effect’ itself (Rhea, 2004; Thomas et al., 1991). Thus, p-values and t-statistics provide information about the estimate of the mean relative to some test hypothesis and thus are sensitive to sample size, while SMDs strictly pertain to the size of the effect and thus are insensitive to sample size. Moreover, any linear transformation of the data will still yield the exact same standardized effect size (Hedges, 1981). The scale invariance property of a standardized effect size theoretically allows them to be compared across studies, various outcomes, and incorporated into a meta-analysis. Therefore, scientists can measure a phenomenon across many different scales or measurement tools and standardized effect sizes should, in theory, be unaffected. Finally, SMDs can provide a simple way to communicate the overlap of two distributions (`https://rpsychologist.com/d3/cohend/`).

Comments on Standardized Mean Differences in Sport and Exercise Science

Sport and exercise scientists have also commented on the use of standardized effect sizes (Dankel et al., 2017; Dankel and Loenneke, 2018; Rhea, 2004; Thomas et al., 1991; Flanagan, 2013). The discussion has focused on the need to report more than just *p*-values, emphasizing that scientists have to discuss the magnitude of their observed effects. Rhea (2004) also provided new benchmarks for SMDs specific to strength and conditioning research, which is certainly an improvement from just using Cohen's benchmarks.

If SMDs are to be reported, they should not be done so in lieu of understanding effects on their natural scales. To this end, we agree with the laments of Tukey (1969): too often, standardized effects sizes, particularly SMDs, are relied upon to provide a crutch for interpreting the meaningfulness of results. Default and arbitrary scales, such as "small" or "large" based on those proposed by Cohen (1977), should generally be avoided. SMDs should be interpreted on a scale calibrated to the outcome of interest. For example, Rhea (2004) or Quintana (2016) have demonstrated how to develop scales of magnitude for a specific area of research. When possible, it is best practice to interpret the meaningfulness of effects in their raw units, and in the context of the population and the research question being asked. For example, a 5 mmHg decrease in systolic blood pressure may be hugely important or trivial, depending on the context—here, the SMD alone cannot communicate clinical relevance.

In our opinion, standardized effect sizes can be useful tools for interpreting data when thoughtfully employed by the scientists reporting them. However, sport and exercise scientists should be careful when selecting the appropriate SMD or effect size, and ensure that their choice effectively communicates the effect of interest (Hanel and Mehler, 2019). Herein, we will discuss things to consider when reporting an SMD, and we will close by providing general recommendations and examples that we believe sport and exercise scientists will find useful.

WHICH STANDARDIZED MEAN DIFFERENCE SHOULD I REPORT?

To facilitate a fruitful discussion of SMDs, here, we categorize them based on the information they convey. We contend that there are two primary categories of SMDs that sport and exercise scientists will encounter in the literature and use for their own analyses. The first helps to communicate the magnitude of an effect (magnitude-based SMD), and the second is more related to the probability that a randomly selected individual experiences a positive or negative effect (signal-to-noise SMD). These categories serve distinct purposes, and they should be used in accordance with the information a scientist is trying to convey to the reader. We will contrast these SMD categories in terms of the information that they communicate and when scientists may wish to choose one over the other. In doing so, we will show that both approaches to calculating the SMD are distinctly valuable. Finally, we demonstrate that, when paired with background information and other statistics—whether they be descriptive or inferential—each SMD can assist in telling a unique, meaningful story about the reported data.

Signal-to-noise Standardized Mean Difference

The first category of SMDs can be considered a signal-to-noise metric: it communicates the average change score in a sample relative to the variability in change scores. This is called **Cohen's** d_z, and it is an entirely appropriate way to describe the change scores in paired data. The Z subscript refers to the difference being compared is no longer between the measurements (X or Y) but the difference ($Z = Y - X$). This SMD is directly estimating the change standardized to the variation in this response, making it a mathematically natural signal-to-noise statistic.

Cohen's d_z can be calculated with the mean change, $\bar{\delta}$, and the standard deviation of the difference, σ_δ,

$$d_z = \frac{\bar{\delta}}{\sigma_\delta}. \quad (1)$$

Alternatively, for convenience, can be calculated from the t-statistic and the number of pairs (n),

$$d_z = \frac{t}{\sqrt{n}}. \tag{2}$$

In Eq. 2, one can see that d_z is closely related to the t-statistic. Specifically, the t-statistic is a signal-to-noise metric for the *mean* (i.e., using its sampling distribution), while d_z is a signal-to-noise metric for the entire sample. This means that the t-statistic will tend to increase with increases in sample size, since the estimate of the mean becomes more precise, while (bias-corrected) Cohen's d_z will not change with sample size.

Although Cohen's d_z may be useful to describe the change in a standardized form, it is typically not reported in meta-analyses since it cannot be used to compare differences across between- and within-subjects designs (see SMDs below). It is difficult to interpret the value of this type of SMD; that is, since the signal-to-noise ratio itself is more related to the consistency of a change, one can wonder how much consistency constitutes a 'large' effect? This is in contrast to other types of SMDs, wherein the statistic conveys information about the distance between two central tendencies (mean) relative to the dispersion of the data (standard deviation). Moreover, it appears that, to sport and exercise scientists, the value of this SMD is measuring the degree of the change in comparison to the variability of the change scores (Dankel and Loenneke, 2018). Therefore, scientists' intent on using d_z should consider reporting the common language effect size (CLES) (McGraw and Wong, 1992), also known as the probability of superiority (Grissom, 1994). In contrast to d_z, CLES communicates the probability of a positive (CLES > 0.5) or negative (CLES < 0.5) change occurring in a randomly sampled individual (see below).

Alternative to the signal-to-noise Standardized Mean Difference

The information gleaned from the signal-to-noise SMD (Cohen's d_z) can also be captured with the CLES (McGraw and Wong, 1992; Grissom, 1994). In paired samples, the CLES conveys the probability of a randomly selected person's change score being greater than zero. The CLES is easy to obtain; it is simply the Cohen's d_z (SMD) converted to a probability ($CLES = \Phi(d_z)$, where Φ is the standard normal cumulative distribution function). Importantly, CLES can be converted back to a Cohen's d_z with the inverse normal cumulative distribution function ($d_z = \Phi^{-1}(\text{CLES})$). CLES is particularly useful because it directly conveys the direction and variability of change scores without suggesting that the mean difference itself is small or large. Further, current evidence would suggest that the CLES is easier for readers to comprehend than a signal-to-noise SMD (Hanel and Mehler, 2019).

Magnitude-based Standardized Mean Difference

The second category of SMDs can be considered a magnitude-based metric: it communicates the size of an observed effect relative to spread of the sample. The simplest and most understood magnitude-based SMD is **Glass's** Δ, which is used to compare two groups, and is standardized to the standard deviation of one of the groups. However, a conceptually similar version of Glass's Δ, which we term Glass's Δ_{pre}, can also be employed for repeated-measures. In Δ_{pre} the mean change change is standardized by the pre-intervention standard deviation.[1] For basic pre-post study designs, Glass's Δ_{pre} is fairly straightforward; mean change is simply standardized to the standard deviation of the pre-test responses. There are other effect sizes for repeated measures designs such as Cohen's d_{av} and d_{rm}, but for brevity's sake these are described in the appendix. Of note, Δ_{pre}, d_{av}, and d_{rm} are identical when pre- and post-intervention variances are the same (see Appendix).

$$\Delta_{\text{pre}} = \frac{\bar{\delta}}{\sigma_{\text{pre}}} \tag{3}$$

Importantly, Δ_{pre} is well-described (Morris and DeShon, 2002; Morris, 2000; Becker, 1988) and can also be generalized to parallel-group designs; in particular, when there are 2 groups, typically a control and treatment group, being compared over repeated-measurements (Morris, 2008). Typically, in these cases, a treatment and control group are being directly compared in a 'pretest-posttest-control design' (PPC). A simple version of the PPC-adapted Δ_{pre} is

$$\Delta_{\text{ppc}} = \Delta_{\text{T}} - \Delta_{\text{C}} \tag{4}$$

[1] Although conceptually similar, Glass's Δ and Δ_{pre} have different distributional properties (Becker, 1988).

where Δ_T and Δ_C are the Δ_{pre} from the treatment and control groups, respectively. There are several other calculative approaches which should be considered for comparing SMDs in a parallel-group designs. We highly encourage further reading on this topic if this type of design is of interest to readers (Morris, 2008; Becker, 1988; Viechtbauer, 2007).

Summary of Standardized Mean Differences

Our distinction between signal-to-noise (namely, Cohen's d_z) and magnitude-based SMDs (including Glass's Δ_{pre}, Cohen's d_{av}, and Cohen's d_{rm}) provides a conceptual dichotomy to assist researchers in picking an SMD (summarized in Table 1). However, along with the conceptual distinctions, researchers should also consider the the properties of these SMDs. In the following section, we briefly go over the math underlying each SMD and its implications. The properties that follow from the math complement the conceptual framework we just presented, in turn providing researchers with a theoretical, mathematical basis for choosing and justifying their choice of an SMD.

Table 1. Types of Standardized Mean Differences for Pre-Post Designs

Magnitude-based	Glass's Δ_{pre}, Cohen's d_{av}, Cohen's d_{rm}
Signal-to-noise	Cohen's d_z

WHAT ARE THE STATISTICAL PROPERTIES OF STANDARDIZED MEAN DIFFERENCES?

An SMD is an estimator. Estimators, including SMDs, have basic statistical properties associated with them that can be derived mathematically. From a high level, grasping how an estimator behaves—what makes it increase or decrease and to what extent—is essential for interpretation. In addition, one should have a general understanding of the statistical properties of an estimator they are using; namely, its bias and variance, which together determine the accuracy of the estimator (mean squared error, $\text{MSE} = \text{Bias}(\hat{\theta}, \theta)^2 + \text{Var}_\theta(\hat{\theta})$, for some true parameter, θ, and its estimate, $\hat{\theta}$). These properties depend on the arguments used in the estimator. As a result, signal-to-noise and magnitude-based SMDs are not only distinct in terms of their interpretation, but also their statistical properties. Although these properties have been derived elsewhere (e.g., Hedges (1981); Morris and DeShon (2002); Morris (2000); Gibbons et al. (1993); Becker (1988)), their implications are worth repeating. In particular, there are several salient distinctions between the properties of each of these metrics, which we will address herein. Although this section is more technical, we will return to a higher-level discussion of SMDs in the next section.

Estimator Components

Before discussing bias and variance, we will briefly discuss the components of the formulae and their implications. Of course, all SMDs contain the mean change score, $\bar{\delta}$, in the numerator, and thus increase linearly with mean change (all else held equal). Since this is common to all SMDs, we will not discuss it further.

More interestingly, the signal-to-noise and magnitude-based SMDs contain very different denominators. To simplify matters, let us assume the pre- and post-intervention standard deviations are equal ($\sigma_{\text{pre}} = \sigma_{\text{post}} = \sigma$). This assumption is reasonable since pre- and post-intervention standard deviations typically do not substantially differ in sports and exercise science. In this case, the standard deviation of change scores can be found simply:

$$\sigma_\delta = \sqrt{2\sigma^2(1-r)}.$$

With these assumptions, $d_{rm} = d_{av} = \Delta_{\text{pre}}$ for $-1 < r < 1$, where r is the observed pre-post correlation (Appendix). Greater pre-post correlations, r, are indicative of more homogeneous change scores. This makes the behavior of the magnitude-based SMDs fairly straightforward; that is, the estimates themselves will not be affected by the correlation between pre- and post-intervention scores. Their dependence on σ means that the magnitude-based SMD will blow up as $\sigma \to 0$. This is in contrast to d_z, whose denominator contains both σ and r (Figure 1, top), making it blow up if either $\sigma \to 0$ or $r \to 1$.

The parsimonious nature of magnitude-based SMDs arguably makes their interpretation easier; with reasonable assumptions, they only depend on the mean change score and the spread of scores in the

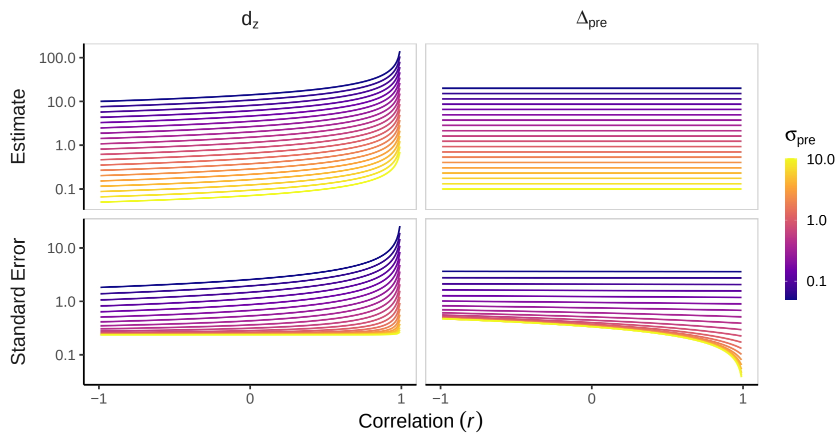

Figure 1. Standardized mean differences for a range of pre-post correlations and pre-intervention standard deviations. Standardized mean differences (SMD) were calculated for a pre-post design study with 20 participants to depict the different properties of the different SMDs. We calculated SMDs for a range of pre-post correlations (r) and pre-intervention standard deviations (σ_{pre}), each with a mean change score of 1. (Top) Magnitude-based SMDs have similar estimates across the range of pre-post correlations and largely only vary as a function of σ_{pre}, whereas signal-to-noise SMDs are a function of both σ_{pre} and r. Note, d_z blows up as $r \to 1$, and all SMDs blow up as $\sigma_{pre} \to 0$. (Bottom) The standard error of each estimator increases as $\sigma_{pre} \to 0$. Importantly, Δ_{pre} has lower or similar standard errors as $r \to 1$, whereas d_z has greater standard errors as $r \to 1$. Additional simulations, including those of other SMDs, can be found in online supplemental material https://www.doi.org/10.17605/OSF.IO/FC5XW

sample. On the other hand, when breaking d_z down into its constituent parts, it depends on the mean change score, the spread of scores in the sample, and the correlation between pre- and post-intervention scores—the latter two will create σ_δ. These sensitivities should be understood before implementing an SMD.

Bias

Bias means that, on average, the estimate of a parameter ($\hat{\theta}$) differs from the "true" parameter being estimated (θ). Most SMDs follow a non-central t-distribution, allowing the bias to be easily assessed and corrected. As shown by Hedges (1981), SMDs are generally biased upwards with small sample sizes; that is, with smaller samples, SMDs are *overestimates* of the true underlying SMD ($\hat{\theta} > \theta$). This bias is a function of both the value of the SMD obtained and the sample size:

$$\mathbb{E}[d] = \hat{d} = \frac{d}{c(n-1)} \tag{5}$$

$$\implies \text{Bias}[\hat{d}, d] = \hat{d} - d = d\left({}^{1}/_{c(n-1)} - 1\right), \tag{6}$$

where d is the "true" parameter being estimated, $\hat{d}$ is its estimate, and $c(m) = 1 - \frac{3}{4(m)-1}$ is Hedges' bias-correction factor (Hedges, 1981) and $m = n - 1$ is the degrees of freedom for a *paired sample*. Please note that this degrees of freedom will differ for different study designs and standard deviations. For example, with two groups and a pooled standard deviation, $m = n_1 + n_2 - 2$. We have noticed the incorrect use of degrees of freedom in some published papers within sport and exercise science, so we urge authors to be cautious.

Because SMDs are biased, especially in small samples, it is advisable to correct for this bias. Thus, when using Cohen's d in small sample settings, most sport and exercise scientists should apply a Hedges' correction to adjust for bias. A bias-corrected $\hat{d}$ is typically referred to as Hedges' g:

$$g = \hat{d} \cdot c(n-1), \tag{7}$$

where $\hat{d}$ can represent any of the SMD estimates outlined above. This correction decreases the SMD by about 10 and 5% with 10 and 15 participants, respectively; corrections are negligible with larger sample sizes. Bias correction can also be applied via bootstrapping (Rousselet and Wilcox, 2019).

More generally, we stress to readers that bias *per se* is not a bad thing or undesirable property. Especially in multidimensional cases, bias can *improve* the accuracy of an estimate by decreasing its variance—this is known as Stein's paradox (Efron and Morris, 1977). Indeed, biased (shrunken) estimators of SMDs have been suggested which may decrease MSE (Hedges and Olkin, 1985). However, these are not commonly employed. Having said this, the upward bias of SMDs is generally a bad thing. As will be discussed in the next subsection, by correcting for the upward bias, we also improve (decrease) the variance of the SMD estimate, in turn decreasing MSE via both bias and variance (Hedges, 1981; Hedges and Olkin, 1985).

Variance

While bias tells us about the extent to which an estimator over- or underestimates the value of a true parameter, variance tells us how variable the estimator is. Estimators that are more precise (less variable) will have tighter standard errors and thus confidence intervals, allowing us to make better judgments as to the "true" magnitude of the SMD.

By looking at formulae for variance and its arguments, we can gain a better understanding of what affects its statistical properties. Below are the variance formulae for Cohen's d_z and Glass's Δ_{pre}, which are the two best understood SMDs for paired designs (Becker, 1988; Gibbons et al., 1993; Goulet-Pelletier and Cousineau, 2018; Morris, 2000; Morris and DeShon, 2002).

$$\text{Var}[d_z] = \frac{n-1}{n(n-3)}(1+d_z^2 n) - \frac{d_z^2}{c(n-1)^2} \tag{8}$$

$$\text{Var}[\Delta_{\text{pre}}] = \frac{n-1}{n(n-3)}\left(2(1-r)+\Delta_{\text{pre}}^2 n\right) - \frac{\Delta_{\text{pre}}^2}{c(n-1)^2} \tag{9}$$

Variances for the biased SMDs (above) can be easily converted to variances for the bias-corrected SMDs by multiplying each formula by $c(n-1)^2$, which is guaranteed to decrease variance since $c(\cdot) < 1$ (Hedges, 1981).

Each variance formula contains the SMD itself, meaning that variance will tend to increase with an increasing SMD. This also complicates matters for d_z; since σ_δ can increase from a smaller σ_{pre} or greater r, d_z's variance explodes with homogeneous populations or change scores (Figure 1). Such a quality is not very desirable, as typically, we would like *more* precision as effects become more homogeneous; this property is a further indication that d_z is not a measure of effect magnitude. This is in contrast to the magnitude-based SMDs, which become more precise as the effect becomes more homogeneous (Figure 1). Of note, these differences in variance behaviors do not reflect differences in statistical efficiency; after adjusting for scaling, all are unbiased and equally efficient.

By investigating and understanding the statistical properties of a statistic—here, the SMDs—we can gain a better understanding of what we should and should not expect from an estimate. These properties provide us with an intuitive feel for the implications of the mathematical machinery underlying each SMD, in turn helping us choose and justify an SMD.

Considering Previous Arguments for Signal-to-noise Standardized Mean Differences

There have been arguments against SMDs—at least certain calculative approaches—with one particular article claiming that magnitude-based SMDs are flawed (Dankel and Loenneke, 2018). Specifically, Dankel and Loenneke (2018) profess the superiority of Cohen's d_z over magnitude-based SMDs—specifically, Glass's Δ_{pre}—because of its statistical properties[2] and its relationship with the t-statistic. Regarding the former, Dankel and Loenneke (2018) opine that the magnitude-based SMD is "dependent

[2]Dankel and Loenneke (2018) suggest that, "... normalizing effect size values to the pre-test SD will enable the calculation of a confidence interval before the intervention is even completed ... This again also points to the flaws of normalizing effect sizes to the pretest SD because the magnitude of the effect ... is dependent on the individuals recruited rather than the actual effectiveness of the intervention" (p. 4). This is of course not the case, since the variance of the SMD will depend on, among other things, the change scores themselves. Thus, the confidence interval of the magnitude-based SMD estimate cannot be calculated *a priori*.

on the individuals recruited rather than the actual effectiveness of the intervention." We do not find this to be a compelling argument against magnitude-based SMDs for several reasons. First, it is in no way specific to magnitude-based SMDs; all descriptive statistics are always specific to the sample. Second, if the data are randomly sampled (a necessary condition for valid statistical inference), then the sample should, on average, be representative of the target population. If imbalance in some relevant covariate is a concern, then an analysis of covariance, and the effect size estimate from this statistical model, should be utilized (Riley et al., 2013).

It is certainly the case that Cohen's d_z has a natural relationship with the t-statistic. Stemming from this relationship, Dankel and Loenneke (2018) suggest that it is a more appropriate effect size statistic for repeated-measures designs. Although it is true that d_z is closely related to the t-statistic, this does not imply that d_z is the most appropriate SMD to report. First, the t-statistic and degrees of freedom (which should be reported) together provide the required information to calculate a Cohen's d_z, meaning d_z may contain purely redundant information. Second, although Cohen's d_z has a clear relationship with the statistical power of a paired t-test, we want to emphasize that utilizing an *observed* effect size in power analyses is an inappropriate practice. Performing such power analyses to justify sample sizes of future work implicitly assumes that 1) the observed effect size is the true effect size; 2) follow-up studies will require this observed SMD; and 3) this effect size is what is of interest (rather than one based on theory or practical necessity). In most cases, observed effect sizes do not provide accurate estimates of the population-level SMD, and utilizing the observed SMD from a previous study will likely lead to an underpowered follow-up study (Albers and Lakens, 2018), and moreover, relying on previously reported effect sizes ignores the potential heterogeneity of observed effect sizes between studies (McShane and Böckenholt, 2014). Rather, there exist alternative approaches to justifying sample sizes (Appendix 2).

In general, and in contrast to Dankel and Loenneke (2018), we believe that SMDs can be used for different purposes—whether to communicate the size of an effect, calculate power, or some other purpose—and what is best for one objective is not necessarily what is best for the others. Furthermore, we want to emphasize that these are not arguments against the use of signal-to-noise SMDs, but rather a repudiation of arguments meant to discourage the use of magnitude-based SMD by sport and exercise scientists.

RECOMMENDATIONS FOR REPORTING EFFECT SIZES

In most cases, sport and exercise scientists are strongly encouraged to present and interpret effect sizes in their raw or unstandardized form. As others previously discussed, journals should require authors to report some form of an effect size, along with interpretations of its magnitude, instead of only reporting p-values (Rhea, 2004; Thomas et al., 1991). However, an SMD, along with other standardized effect sizes, do not magically provide meaning to meaningless values. They are simply a convenient tool that can provide some additional information and may sometimes be helpful to those performing meta-analyses or who are unfamiliar with the reported measures. Specifically, there are situations where the outcome measure may be difficult for readers to intuitively grasp (e.g., a psychological survey, arbitrary units from Western Blots, moments of force). In such cases, a magnitude-based SMD—in which the SD of pre- and/or post-intervention measures is used in the denominator—can be used to communicate the size of the effect relative to the heterogeneity of the sample. In other words, a magnitude-based SMD represents the expected number of sample SDs (not the change due to the intervention) by which the participants improve.

Let us consider examples presented previously in the sport and exercise science literature. The examples presented in Figure 1 by Dankel and Loenneke (2018), in which both interventions have a pre-intervention SD_{pre} = 6.05 and undergo a change of Δ= 3.0 ($SMD = \frac{3.0}{6.05} = 0.5$). This can be interpreted simply: the expected change is 0.5 standard deviation units relative to the measure in the sample. Put differently, if the person with the median score (50th percentile) were to improve by the expected change, she would move to the 69th percentile.[3] Like a mean change, this statistic is not intended to provide information about the variability of change scores. The magnitude-based SMD simply provides a unitless, interpretable value that indicates the magnitude of the expected change relative to the between-subject standard deviation. Of course, it can be complemented with a standard error or confidence interval if one is interested in the uncertainty around this estimate.

[3]Assumes a normal distribution. We note that Dankel and Loenneke (2018) data vignettes are approximately uniformly distributed which is an odd assumption to make about theoretical data, but nonetheless, sufficiently conveys the point.

The above can be contrasted with Cohen's d_z, which uses the SD of change scores. Again, using the examples presented in Figure 1 of Dankel and Loenneke (2018), Cohen's d_z of 11.62 and 0.25 are reported for interventions 1 and 2, respectively. If one tries to interpret these SMDs in a way that magnitude-based SMDs are interpreted, he will undoubtedly come to incorrect conclusions. The first would suggest that a person with the median score who experiences the expected change would move to >99.99th percentile, and the second would imply that she moves to the 60th percentile. Clearly, both of these interpretations are wrong. As opposed to a magnitude-based SMD, Cohen's d_z is a signal-to-noise statistic that is related to the probability of a randomly sampled individual experiencing *an* effect rather than its magnitude alone. In our opinion, Cohen's d_z does not provide any more information than that which is communicated by the t-statistic and the associated degrees of freedom (which should be reported regardless of the effect size). Instead, if the signal-to-noise is of interest, a CLES may provide the information a sport and exercise scientist is interested in presenting. Going back to our earlier example (d_z = 11.62 and 0.25 respectively), the CLES would be approximately > 99% and 59.9%, or the probability of a randomly sample individual undergoing an improvement is > 99% or 59.9% for intervention 1 and 2, respectively. As Hanel and Mehler (2019) demonstrated, the CLES may be a more intuitive description of the signal-to-noise SMD. While our personal recommendation leans towards the use of magnitude-based SMDs and CLES, it is up to the individual sport and exercise scientist to decide what effect size they feel is most appropriate for the data they are analyzing and point they are trying to communicate (Hönekopp et al., 2006).

In choosing an SMD, we also sympathize with Lakens (2013), "... to report effect sizes that cannot be calculated from other information in the article, and that are widely used so that most readers should understand them. Because Cohen's d_z can be calculated from the t-value and the n, and is not commonly used, my general recommendation is to report Cohen's d_{av} or Cohen's d_{rm}." Along these same lines, if scientists want to present an SMD, it should not exist in isolation. It is highly unlikely that a single number will represent all data in a meaningful way. We believe that data are often best appreciated when presented in multiple ways. The test and inferential statistics (p-values and t-statistics) should be reported alongside an effect size that provides some type of complementary information. This effect size can be standardized (e.g., Δ_{pre}) or unstandardized (raw), and should be reported with a confidence interval. Confidence intervals (CI) of a magnitude-based SMD will provide readers with information concerning both the magnitude and uncertainty of an effect size; CIs can be calculated using formulae, or perhaps more easily, using the bootstrap. In situations where the measurements are directly interpretable, unstandardized estimates are generally preferable. The CLES can also be reported when the presence of a change or difference between conditions is of interest.

Percent Changes

It is not uncommon for sport and exercise scientists to report their data using percentages (e.g., percent change). While this is fine if it supplements the reporting of their data in raw units, it can be problematic if it is the only way the data are presented or if the statistics are calculated based on the percentages. In the case of SMDs, an SMD calculated using a percent change is not the same as an SMD calculated using raw units. More importantly, the latter—which is often of greater interest to readers or those performing meta-analysis—cannot be back-calculated from the former. It is imperative that authors consider the properties of the values that they report and what readers can glean from them.

Data Sharing

To facilitate meta-analysis, we suggest that authors upload their data to a public repository such as the Open Science Framework, FigShare, or Zenodo (Borg et al., 2020). This ensures that future meta-analysis or systematic reviews efforts have flexibility in calculating effect sizes since there are multitude of possible calculative approaches, designs, and bias corrections (see Baguley (2009)). When data sharing is not possible, we highly encourage sport and exercise scientists to upload extremely detailed descriptive statistics as supplementary material (i.e., sample size per group, means, standard deviations, and correlations), or alternatively, a synthetic dataset that mimics the properties of the original (Quintana, 2020).

Examples

In the examples below, we have simulated data and analyzed it in `R` (see supplementary material) to demonstrate how results from a study in sport and exercise science could be interpreted with the appropriate application of SMDs. For those unfamiliar with `R`, there is an online web application (`https://`

doomlab.shinyapps.io/mote/) and extensive documentation (https://www.aggieerin.com/shiny-server/introduction/) to simplify the process of calculating SMDs for those without R programming experience (Buchanan et al., 2019).

Scenario 1: Interpretable Raw Differences

In the first hypothetical example, let us imagine a study trying to estimate the change in maximal oxygen consumption ($\dot{V}O_2$; L·min^{-1}) in long-distance track athletes before and after a season of training. For this study, maximal $\dot{V}O_2$ was measured during a Bruce protocol with a Parvomedics 2400 TrueOne Metabolic System. The results of this hypothetical outcome could be written up as the following:

> $\dot{V}O_2$ after a season of training with the track team (mean = 4.13 L·min^{-1}, SD = 0.25) increased compared to when they joined the team (M = 3.89 L·min^{-1}, SD = 0.21), $t(7) = 3.54$, $p = 0.009$, $\bar{\delta} = 0.23$ L·min^{-1} 95% C.I. [0.07, 0.38]. The CLES indicates that the probability of a randomly selected individual's $\dot{V}O_2$ increasing after their first season with the team is 89%.

Scenario 2: Uninterpretable Raw Differences

Now, let us imagine a study trying to estimate the effect of cold water immersion on muscle soreness. For this hypothetical study, muscle soreness is measured on a visual analog scale before and after cold water immersion following a muscle damaging exercise. The muscle soreness score would be represented by cm on the scale measured left-to-right. Because sensations tend to be distributed lognormal (Mansfield, 1974)—and are multiplicative rather than additive—it is sensible to work with the logarithm of the reported soreness levels. Since these logged scores are not directly interpretable, it is sensible to use an SMD to help interpret the change scores. The hypothetical study could be written up as follows:

> Muscle soreness was lower after cold water immersion (mean = 27, SD = 7) compared to before (mean = 46, SD = 11) cold water immersion, $t(9) = -6.90$, $p < .001$, Glass's Δ_{pre} = -2.2 95% CI [-3.2, 1.3]. The CLES indicates that the probability of a randomly selected individual experiencing a reduction in muscle soreness after cold water immersion is 99%.

CONCLUSION

We contend that the reporting of effect sizes should be specific to the research question in conjunction with the narrative that a scientist wants to convey. In this context, pooled pre- and/or post-study SDs are viable choices for the SMD denominator. This approach provides insight into the magnitude of a given finding, and thus can have important implications for drawing practical inferences. Moreover, the values of this approach are distinct and, in our professional opinion, potentially more insightful than signal-to-noise SMDs, which essentially provide information that is redundant with the *t*-statistic. At the very least, there is no one-size-fits-all solution to reporting an SMD, or any other statistics for that matter. Despite our personal preference towards other effect sizes, a sport and exercise scientist may prefer a signal-to-noise SMD (d_z) and could reasonably justify this decision. We urge sport and exercise scientists to avoid reporting the same default effect size and interpreting them based on generalized, arbitrary scales. Rather, we strongly encourage sport and exercise scientists justify which SMD is most appropriate and provide qualitative (i.e., small, medium, or large effect) interpretations that are specific to that outcome and study design. Also, sport and exercise scientists should be careful to report the rationale for using an SMD over simply presenting raw mean differences. Lastly, the creation of statistical rituals wherein a single statistic, *by default*, is used to interpret the data is likely to result in poor statistical analyses rather than informative ones (Gigerenzer, 2018). As J.M. Hammersley once warned, "There are no routine statistical questions; only questionable statistical routines" (Sundberg, 1994).

APPENDIX 1: STANDARDIZED MEAN DIFFERENCE CALCULATIVE APPROACHES

Throughout the text, we use Glass's Δ_{pre} as our token magnitude-based SMD. However, there exist other approaches to calculating magnitude-based SMDs. Here, we briefly discuss two other common calculations of magnitude-based SMDs. Of note, these two other calculative approaches may contain some "effects" (variance) from the intervention in the denominator, arguably making Glass's a more "pure" (in the sense that the denominator is uncontaminated by intervention effects) magnitude-based SMD.

Cohen's d_{av}: Some have argued that Cohen's d_z is an overestimate of the SMD, and instead advocate for reporting an SMD very similar to the Cohen's d_s typically utilized for between-subjects (independent samples) designs (Dunlap et al., 1996). The only difference between Cohen's d_{av} and Cohen's d_s is that the *average* standard deviation between the two-samples (e.g., pre- and post-intervention assessments in a repeated-measures design) is used rather than the pooled standard deviation.

$$d_{av} = \frac{\bar{\delta}}{\sigma_{\text{pre}} + \sigma_{\text{post}} / 2} \tag{10}$$

Cohen's d_{rm}: The standardized difference between repeated-measures (hence "rm") is arguably the most conservative SMD among those reported. This approach "corrects" for repeated-measures by taking into account the correlation between the two measurements.

$$d_{rm} = \frac{\bar{\delta}}{\sqrt{\sigma_{\text{pre}}^2 + \sigma_{\text{post}}^2 - 2 \cdot r \cdot \sigma_{\text{pre}} \cdot \sigma_{\text{post}}}} \cdot \sqrt{2 \cdot (1-r)} \tag{11}$$

$$= \frac{\bar{\delta}}{\sigma_{\delta}} \cdot \sqrt{2 \cdot (1-r)} \tag{12}$$

$$= d_z \cdot \sqrt{2 \cdot (1-r)} \tag{13}$$

APPENDIX 2: JUSTIFYING SAMPLE SIZES

There are more appropriate approaches to justifying sample sizes than using previously reported effect sizes. First, if authors have a question that, for some reason, necessitates null hypothesis significance testing, authors should first perform the necessary risk analysis to obtain their desired error rates. Next, authors can specify a smallest effect size of interest (SESOI) or minimal clinically important difference (MCID) (Hislop et al., 2014). Of note, the ontological basis for (or the rationale for the true existence of) such dichotomizations—both in the effect (SESOI, MCID) and p-value domains (α-level)—should be justified. Oftentimes, it is not the researcher, but a reader, clinician, or policymaker who must make a decision; for such decisions, proper, contextual decision analytic frameworks should be employed (Amrhein et al., 2019; Hunink et al., 2014; Vickers and Elkin, 2006). Second, if relying on estimation rather than hypothesis testing, sport and exercise scientists could determine a sample size at which they would have high enough "assurance" that the estimates would be sufficiently accurate (i.e., confidence intervals around the effect size are sufficiently narrow) (Maxwell et al., 2008). Third, authors may simply be working under constraints (e.g., time, money, or other resources) that prohibit them from recruiting more than n participants. We believe such pragmatic constraints are perfectly reasonable and justifiable. No matter the sample justification, it should be thoughtful and reported transparently. Importantly, the utility of an effect size or SMD should not be determined by its ability to be used in sample size justifications or calculations.

ACKNOWLEDGMENTS

We would like to thank Brad Schoenfeld for his feedback during the early drafting of this manuscript; our writing of this paper would not have occurred without his encouragement. In addition, we would like to thank Thom Baguley, Israel Halperin, Keith Lohse, Scott Morris, and Kristin Sainani for their thoughtful feedback.

The opinions or assertions contained herein are the private views of the author(s) and are not to be construed as official or reflecting the views of the Army or the Department of Defense. Any citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement of approval of the products or services of these organizations. No authors have any conflicts of interest to disclose. Approved for public release; distribution is unlimited.

REFERENCES

Albers, C. and Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, 74:187–195.

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, 100(3):603–617.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, 41(2):257–278.

Borg, D. N., Bon, J. J., Sainani, K., Baguley, B. J., Tierney, N. J., and Drovandi, C. (2020). Sharing data and code: A comment on the call for the adoption of more transparent research practices in sport and exercise science.

Buchanan, E. M., Gillenwaters, A., Scofield, J. E., and Valentine, K. (2019). *MOTE: Measure of the Effect: Package to assist in effect size calculations and their confidence intervals*. R package version 1.0.2.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York, 2 edition.

Dankel, S. J. and Loenneke, J. P. (2018). Effect sizes for paired data should use the change score variability rather than the pre-test variability. *Journal of Strength and Conditioning Research*, page 1.

Dankel, S. J., Mouser, J. G., Mattocks, K. T., Counts, B. R., Jessee, M. B., Buckner, S. L., Loprinzi, P. D., and Loenneke, J. P. (2017). The widespread misuse of effect sizes. *Journal of Science and Medicine in Sport*, 20(5):446–450.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., and Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, 1(2):170–177.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236(5):119–127.

Flanagan, E. P. (2013). The effect size statistic—applications for the strength and conditioning coach. *Strength and Conditioning Journal*, 35(5):37–40.

Gibbons, R. D., Hedeker, D. R., and Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics*, 18(3):271–279.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218.

Goulet-Pelletier, J.-C. and Cousineau, D. (2018). A review of effect sizes and their confidence intervals, part i: The cohen's d family. *The Quantitative Methods for Psychology*, 14(4):242–265.

Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, 73(sup1):106–114.

Greenland, S., Maclure, M., Schlesselman, M., Poole, C., and Morgenstrern, H. (1986). Standardized regression coefficients: A further critique and review of some alternatives. *Epidemiology*, 2(5):387–92.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, 79(2):314–316.

Hanel, P. H. and Mehler, D. M. (2019). Beyond reporting statistical significance: Identifying informative effect sizes to improve scientific communication. *Public Understanding of Science*, 28(4):468–485.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128.

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, 2(3):167–171.

Hedges, L. V. and Olkin, I. (1985). Chapter 5 - estimation of a single effect size: Parametric and nonparametric methods. In Hedges, L. V. and Olkin, I., editors, *Statistical Methods for Meta-Analysis*, pages 75 – 106. Academic Press, San Diego.

Hislop, J., Adewuyi, T. E., Vale, L. D., Harrild, K., Fraser, C., Gurung, T., Altman, D. G., Briggs, A. H., Fayers, P., Ramsay, C. R., and et al. (2014). Methods for specifying the target difference in a randomised controlled trial: The difference elicitation in trials (delta) systematic review. *PLoS Medicine*, 11(5):e1001645.

Hunink, M. M., Weinstein, M. C., Wittenberg, E., Drummond, M. F., Pliskin, J. S., Wong, J. B., and Glasziou, P. P. (2014). *Decision making in health and medicine: integrating evidence and values*. Cambridge University Press.

Hönekopp, J., Becker, B. J., and Oswald, F. L. (2006). The meaning and suitability of various effect sizes for structured rater × ratee designs. *Psychological Methods*, 11(1):72–86.

Kelley, K. and Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17(2):137–152.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, 4:863.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3):187–193.

Mansfield, R. J. (1974). Measurement, invariance, and psychophysics. In *Sensation and measurement*, pages 113–128. Springer.

Maxwell, S. E., Kelley, K., and Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59(1):537–563.

McGraw, K. O. and Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111(2):361–365.

McShane, B. B. and Böckenholt, U. (2014). You cannot step into the same river twice. *Perspectives on Psychological Science*, 9(6):612–625.

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, 53(1):17–29.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational research methods*, 11(2):364–386.

Morris, S. B. and DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7(1):105–125.

Quintana, D. S. (2016). Statistical considerations for reporting and planning heart rate variability case-control studies. *Psychophysiology*, 54(3):344–349.

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife*.

Rhea, M. R. (2004). Determining the magnitude of treatment effects in strength training research through the use of the effect size. *The Journal of Strength and Conditioning Research*, 18(4):918.

Riley, R. D., Kauser, I., Bland, M., Thijs, L., Staessen, J. A., Wang, J., Gueyffier, F., and Deeks, J. J. (2013). Meta-analysis of randomised trials with a continuous outcome according to baseline imbalance and availability of individual participant data. *Statistics in Medicine*, 32(16):2747–2766.

Robinson, D. H., Whittaker, T. A., Williams, N. J., and Beretvas, S. N. (2003). It's not effect sizes so much as comments about their magnitude that mislead readers. *The Journal of Experimental Education*, 72(1):51–64.

Rousselet, G. A. and Wilcox, R. R. (2019). Reaction times and other skewed distributions: problems with the mean and the median. *PsyArXiv*.

Sundberg, R. (1994). Interpretation of unreplicated two-level factorial experiments, by examples. *Chemometrics and Intelligent Laboratory Systems*, 24(1):1–17.

Thomas, J. R., Salazar, W., and Landers, D. M. (1991). What is missing in $p < .05$? effect size. *Research Quarterly for Exercise and Sport*, 62(3):344–348.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, 24(2):83–91.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.

Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics*, 32(1):39–60.