EFFECT SIZES FOR EQUIVALENCE TESTING:

INCORPORATING THE EQUIVALENCE INTERVAL

NAOMI MARTINEZ GUTIERREZ

A THESIS SUBMITTED TO

THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

GRADUATE PROGRAM IN PSYCHOLOGY

YORK UNIVERSITY

TORONTO, ONTARIO

JUNE 2021

Abstract

Equivalence testing (ET) is a framework to determine if an effect is small enough to be considered meaningless, wherein meaningless is expressed as an equivalence interval (EI). Although traditional effect sizes (ESs) are important accompaniments to ET, these measures exclude information about the EI. Incorporating the EI is valuable for quantifying how far the effect is from the EI bounds. An ES measure we propose is the proportional distance (PD) from an observed effect to the smallest effect that would render it meaningful. We conducted two Monte Carlo simulations to evaluate the PD when applied to (1) mean differences and (2) correlations. The coverage rate and bias of the PD were excellent within the investigated conditions. We also applied the PD to two recent psychological studies. These applied examples revealed the beneficial properties of the PD, namely its ability to supply information above and beyond other statistical tests and ESs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Effect Sizes for Equivalence Testing:

Incorporating the Equivalence Interval

Equivalence testing (ET) is used to determine if an effect is small enough to be considered meaningless, practically insignificant, negligible, etc., where what is considered meaningless, etc. is determined a priori by a researcher via an equivalence interval (EI). For example, a researcher might investigate whether the mean difference in depression scores between healthy controls and those treated with cognitive behavioral therapy is negligible, where negligible might mean a difference less than 2 points on the associated scale. Or a researcher might investigate whether the correlation between depression and achievement is practically insignificant, where practically insignificant might mean a correlation less than .2 in magnitude.

Although the first equivalence tests were proposed decades ago (e.g., Anderson & Hauck, 1983; Schuirmann, 1987; Westlake, 1971), these tests were mostly used in pharmaceutical research for demonstrating the equivalence of brand name and generic drugs (Wellek, 2010). It has not been until recently that ET has garnered attention in the psychological sciences. Although traditional effect size (ES) measures (e.g., *d*, *r*) are meaningful as accompaniments to ET, these measures exclude information about the EI. Incorporating information regarding the EI is extremely valuable for quantifying the magnitude of an effect; more specifically, how far the effect is from the bounds of the EI will supply information above and beyond that provided by other statistical tests and ESs.

Therefore, the primary objective of this paper is to propose a novel ES measure for use in ET that incorporates the bounds of the EI. However, before proposing these new methods, we introduce ET, provide examples of the most common equivalence tests in psychology, and outline traditional ES measures that commonly accompany equivalence tests.

Introduction to Equivalence Testing

The earliest instances of ET can be traced back to Wilfred Westlake in the 1970s (e.g., Westlake, 1971) on the equivalence of chemical reactions. Since then, ET has mostly been used as a statistical tool to demonstrate that the effects of similar drugs did not differ (e.g., the comparison between a highly marketed drug and its generic counterpart). The method was introduced to the social sciences by Rogers et al. (1993). In recent years, it has been gaining traction in the field of psychology. For example, equivalence tests have been used to test for gender similarities in intelligence (Ball et al., 2017), whether there are similarities between the usage of in-lab samples versus online samples (Briones & Benham, 2017), and, to test whether an intervention targeted towards children with autism was ineffective (Silva et al., 2008).

There are numerous instances in which psychological researchers wish to show that an effect is negligible. For example, finding that Method A is similar to Method B can prove to be beneficial if one method is less costly and easier to implement than the other. This is exhibited by Lüdtke et al. (2018), wherein their study tested whether the depression scores of participants in an online intervention condition were equivalent to scores for those in a traditional intervention condition at post-assessment. To address such research questions (e.g., mean group similarity), researchers often opt to use traditional null hypothesis significance testing (NHST) methods (e.g., *t*-test). When adopting a traditional NHST procedure for assessing mean *similarity,* the goal is to NOT reject the null hypothesis (e.g., H_0: $\mu_1 = \mu_2$). However, not rejecting the null hypothesis of traditional tests cannot be interpreted to mean that the effect is null (e.g., the result could be a Type II error). That is, no evidence of a relationship between groups (or any other effect) may be attributed to a lack of power, sampling error, etc. In contrast, with a large sample size (i.e., ample statistical power), evidence for the existence of a

relationship could be found, but the effect be too small to be of any value within the research context. Equivalence testing, on the other hand, seeks to demonstrate that the magnitude of an effect is too small to be considered meaningful.

Although there are several available approaches for testing equivalence (e.g., Hauck & Anderson, 1983; Meyners, 2012), the most common approach to testing for equivalence is the two one-sided tests (TOST) procedure (Schuirmann, 1987), wherein two one-sided hypothesis tests are performed simultaneously to attempt to rule out the presence of an effect in either direction (e.g., mean difference in either direction, correlation in either direction). As introduced above, prior to adopting an ET, a researcher must define an appropriate EI, where the bounds of the equivalence interval represent the minimum meaningful effect size (MMES) (Rogers et al., 1993). For example, when evaluating mean equivalence, an appropriate lower bound (Δ_L) and upper bound (Δ_U) might be EI = $\{\Delta_L = -.2, \Delta_U = .2\}$. Note that equivalence boundaries need not be symmetrical about zero and depend on the scaling of the variables and context of the research.

The null (H_0) and alternate (H_1) hypotheses are established in conjunction with the EI. For example, in a mean equivalence setting the null and alternate hypotheses might be H_0: $\mu_1 - \mu_2 \leq \Delta_L \mid \mu_1 - \mu_2 \geq \Delta_U$ (where | represents the *or* operator) and H_1: $\Delta_L < \mu_1 - \mu_2 < \Delta_U$ (Rogers et al., 1993), whereas in a negligible association setting the null and alternate hypotheses might be H_0: $\rho \leq \Delta_L \mid \rho \geq \Delta_U$ and H_1: $\Delta_L < \rho < \Delta_U$. That is, failure to reject the composite null means that the effect may either be equal to or larger than the upper bound of the EI, *or* the effect is equal to or smaller than the lower bound of the EI. Rejection of the null, on the other hand, suggests that the effect lies between the lower and upper boundaries of the EI and is therefore not meaningful.

Effect of Interest

An observed ES is compared to equivalence bounds to address the question of whether the magnitude of the effect is small enough to be considered meaningless. In a mean difference setting, the effect of interest is the raw mean difference, in a correlation setting the effect of interest is the sample correlation. For example, the observed ES between a group of participants that played a game in a desktop platform and a group of participants that played it on a tablet was 0.42 in raw units in terms of perceived sex-based affinities (Wasserman & Rittenour, 2019). That is, the difference in means between these two groups was 0.42 points.

Equivalence Intervals

The most crucial step in ET is the selection of an appropriate EI, which incorporates the MMES. It must be in the same scale as the units being analyzed and can be expressed as a percentage. As noted by Schuirmann (1987), the EI is determined not by a statistician but by the expert in their respective field. In other words, choosing the EI is a subjective decision that depends on the research context. Of 46 clinical equivalence tests reviewed by Le Henanff et al. (2006), only about 20% provided a justification for the chosen EI. Further, some EIs were too large to be convincing and, in turn, provide misleading results (Gøtzsche, 2006). To avoid potential biases introduced from publishing incentives, researchers should also consider an independent party to specify their EI (Campbell & Gustafson, 2018), a common research practice amongst clinical studies (e.g., Staszewski et al., 2001)

In some instances, it might be easier for researchers to conceptualize the EI in standardized units (e.g., d = -.2 to d = .2). In this case, the researcher simply needs to convert from a standardized scale to raw scale in order to run the equivalence test (e.g., $d * SD_{pooled}$).

It is also important to mention that power is sensitive to changes in the EI. That is, a small decrease in magnitude translates into a larger sample size necessary to achieve the same level of power (Walker & Nowacki, 2011). The reason being that it is harder to establish equivalence when the EI has a smaller critical region. For example, Wasserman and Rittenour (2019) used an EI of ±0.932 on a raw scale to have an 80% probability of obtaining statistical significance using a sample size of n = 34 per group and assuming a true effect size of zero. Were the EI to decrease by ¼ (±0.699), the number of participants required to obtain 80% power would increase by approximately 179% (n = 61 per group).

Generally, it is advised that the EI be specified *a priori* to data collection (Lakens et al., 2018; Wellek, 2010). However, an EI specified *post-hoc* does not render the equivalence test invalid as long as it is independent of the data and its appropriateness is justified (Campbell & Gustafson, 2018).

Independent Groups Equivalence Test

The equivalence test for two independent means, assuming equal variances, is based on Student's t test, wherein H_0 is rejected if $t_1 \leq t_v^{\alpha}$:

$$t_1 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_U}{\sqrt{\frac{(n_1 + n_2)[(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2]}{n_1 n_2 (n_1 + n_2 - 2)}}} \qquad (1)$$

and $t_2 \geq t_v^{1-\alpha}$ where:

$$t_2 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_L}{\sqrt{\frac{(n_1 + n_2)[(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2]}{n_1 n_2 (n_1 + n_2 - 2)}}} \; . \qquad (2)$$

$\bar{X}_1$ and $\bar{X}_2$ are the group means, n_1 and n_2 are the group sample sizes, SD_1 and SD_2 are the group standard deviations, and t_v^α is the lower tail α-level *t*-critical value with degrees of freedom (v) of $n_1 + n_2 - 2$.

Equivalence test for correlations

Because the sampling distribution for Pearson's *r* changes from normal to negatively skewed as the correlation increases, a Fisher's *r* to *z* transformation is necessary to maintain a normal distribution (Lee, 2016). Further, the *z* transformation accounts for the bias introduced if one were to incorporate *r* into a *t*-statistic for equivalence testing (Goertzen & Cribbie, 2010). H_0 is rejected if $Z_1 \leq Z_\alpha$:

$$Z_1 = \frac{\frac{\log\left(\frac{1+r}{1-r}\right)}{2} - \frac{\log\left(\frac{1+\rho_U}{1-\rho_U}\right)}{2}}{\frac{1}{\sqrt{n_{paired}-3}}} \qquad (3)$$

or $Z_2 \geq Z_{1-\alpha}$ where:

$$Z_2 = \frac{\frac{\log\left(\frac{1+r}{1-r}\right)}{2} - \frac{\log\left(\frac{1+\rho_L}{1-\rho_L}\right)}{2}}{\frac{1}{\sqrt{n_{paired}-3}}} \quad , \quad (4)$$

ρ_L and ρ_U are the upper and lower bounds of the EI, Z_α is the lower tail critical value obtained from a normal distribution and n_{paired} are the paired data observations.

TOST-based ETs can also be conducted by using confidence intervals (CIs), where the null is rejected if a 100(1-2α)% CI lies within Δ_L and Δ_U. This is similar to using a 100(1-α)% CI to determine whether the null for a traditional two-sided test was rejected or not (i.e., reject H_0 if the CI excludes the null value). That is, if Δ_L is less than or equal to the lower CI endpoint and Δ_U is greater than or equal to the upper CI endpoint, this would result in statistical equivalence.

On the other hand, if Δ_L is greater than the lower CI endpoint or Δ_U is less than the upper CI endpoint, this would result in statistical non-equivalence (Meyners, 2007).

Imagine a researcher working in a mean difference context with an EI = {-.5, .5}, and a 100(1-2α)% CI of [-0.9, 0.2]. Thus, since the 100(1-2α)% CI does not fall completely within the EI, they would not conclude that the means are equivalent.

Traditional Effect size Measures and their role in ET

ESs provide the primary information regarding the magnitude and direction of a phenomenon that is addressed by a research question (Kelley & Preacher, 2012) and allow researchers to determine the practical significance of sample results (Kirk, 1996). An ES can be reported as unstandardized or standardized. An unstandardized ES is reported in the same units as the data (e.g., difference between two group means, regression coefficient), whereas a standardized ES is measured on a common scale that allows researchers to compare/combine effects across different variables, studies, etc.

Traditionally, standardized ESs can be thought of as belonging to one of two families: The *d* family or the *r* family. Examples of ESs from the *d* family are Cohen's *d*, Hedges *g*, and Glass' Δ, all of which measure the difference between two means in standardized units. Cohen's *d* quantifies the difference between the means of two groups in terms of the number of standard deviations. It can range from -∞ to +∞. Several researchers have proposed recommended cut-offs for interpreting the magnitude of *d*, with Cohen (1988) suggesting *d* values of 0.2, 0.5, and 0.8 as the reference values for small, medium and large effects (respectively). Examples of effect sizes from the *r* family are Pearson's *r* and its derivatives, such as semi-partial *r* (popular in multiple regression), point-biserial *r*, and Spearman's ρ. Other types of effect sizes include risk measures (relative risk, odds ratio and risk difference), all of which measure the probability of an outcome

relative to the predictor (Rosenthal, 2000). Of the ESs in the *r* family, the Pearson product moment correlation coefficient (*r*) is common in the behavioural sciences. For example, Farmus et al. (2020) found that *r* was the most common ES reported in social-personality research.

However, it is important to note that what researchers interpret as meaningful varies greatly (Beribisky et al., 2019) and that Cohen (1988) himself clarified that his published cutoffs for small, medium and large effects should only be applied when no other information regarding the research context is available.

It is important to note that within a traditional NHST context, an ES quantifies the extent to which the sample results deviate from the null hypothesis (e.g., no difference in group means or no association between two variables) (Cohen, 1994; Thompson, 2002). For example, if the null hypothesis is that the population postintervention means of the treatment and control groups are equal, the ES is zero if both samples do not differ (e.g., $M_1 = 0.40$ and $M_2 = 0.40$). On the other hand, the ES would be non-zero if the sample means were to differ (e.g., $M_1 = 0.35$ and $M_2 = 0.40$).

Recall that in ET, the null hypothesis considers the EI (i.e., H_0: $\mu_1 - \mu_2 \leq \Delta_L \mid \mu_1 - \mu_2 \geq \Delta_U$). Therefore, unlike in traditional NHST, a traditional ES cannot quantify the extent to which the sample results deviate from the null hypothesis because it does not take into account the EI. To solve this, we introduce in the following section an ES that incorporates the EI within its calculation.

Incorporating EIs into ESs: The Proportional Distance

The Proportional Distance (PD) is an ES that incorporates the EI to quantify the extent to which the observed effect deviates from the null hypothesis that the effect is equal to or larger than the upper bound of the EI, *or* that the effect is equal to or smaller than the lower bound of

the EI. In other words, the PD measures the proportional distance from the effect of interest (e.g., mean difference, correlation) to the bound of the equivalence interval that is the same sign as the effect of interest. Its generic formula is:

$$PD = \frac{effect}{|\,\mathrm{EI}_{sign}\,|} \qquad (5)$$

Wherein *effect* is the observed effect of substantive interest (e.g., mean difference, correlation) and the EI_{sign} is the bound of the EI that is the same sign as the observed effect. For example, in a mean difference context, if EI = {Δ_L = -.20, Δ_U = .20} and the effect is .10, then EI_{sign} is 0.20 because the effect is positive. Incorporating these values, PD = .10/|.20| = .50. Keep in mind that equivalence intervals need not be symmetrical about zero, hence the need to specify EI_{sign}.

PD = 0 means that there is a complete absence of association. In a mean difference context, this represents no difference in means between the two groups being compared; $M_2 - M_1 = 0$. In a correlational context, a PD of 0 means that there is zero relationship between the variables being compared ($r = 0$). If PD < 1 and PD > -1 (i.e., -1< PD <1), this means that the observed effect is within the equivalence boundaries. For example, if PD = 0.33, then an effect (i.e., observed mean difference or *r*) lies approximately 1/3 of the distance away from 0 to the upper bound. A PD equal to 1 or -1 means that the effect (i.e., mean difference or *r*) is equal to either the upper or lower equivalence bound, respectively. Lastly, if PD > 1 or PD < -1, then the effect (i.e., mean difference or *r*) lies outside the equivalence bounds. For example, if PD = 1.33, then the effect lies approximately 33% beyond the upper bound; or, 133% of the distance from 0 to the upper equivalence bound.

Assuming we have a specific positive effect (i.e., mean difference > 0 or $r > 0$), the PD decreases exponentially as the magnitude of the upper bound of the EI (more specifically EI_{sign}) increases. This relationship is best represented by the exponential function 1 / $\left|EI_{sign}\right|$. For example, holding the mean difference constant at .01, the PD is equal to 1 when the EI_{sign} is .01; PD = 1/2 when EI_{sign} is .02; PD = 1/3 when EI_{sign} is .03, and so on (see Figure 1). On the other hand, the PD has a period of rapid increase followed by slow increase as the EI_{sign} increases when holding a negative effect constant. For example, holding the mean difference constant at -.01, the PD is equal to -1 when the $\left|EI_{sign}\right|$ is .01; PD = -1/2 when $\left|EI_{sign}\right|$ is .02; PD = -1/3 when $\left|EI_{sign}\right|$ is .03, and so on (see Figure 2).

Holding the EI_{sign} constant, the PD increases by a constant value as the effect (e.g., mean difference, correlation) increases. This directly proportional relationship can best be represented by the linear function $f(x) = x$. For example, holding the EI_{sign} constant at 1, the PD is equal to -.1 when the effect is -.1; PD = -.09 when the effect is -.09; PD = -.08 when the effect is -0.08, and so on.

Advantages

Results of the TOST procedure only inform us regarding the statistical significance of the proposed equivalence test. It is important to recall that *p*-values alone are unreliable measures for drawing conclusions regarding the magnitude of associations of interest (e.g., Cohen 1994; Cumming, 2012). The effect of interest should not be measured in such dichotomous terms, but rather in terms of estimation by providing information regarding the magnitude of the effect and EI.

ESs with integrated EIs allow the reader to shift focus from simple identification of statistical equivalence toward a quantitative description of how far the observed effect is from its

respective EI. In other words, it would allow readers to move beyond the simple question of 'Are they equivalent or not?' to the more insightful, 'How close is the effect to the specified boundaries of the EI?' Consider a clinical researcher interested in whether treatment A and treatment B are equally effective in treating a certain disorder. The researcher chooses raw values of $\Delta_L = -.5$, $\Delta_U = .5$ for their EI in a mean difference context and obtains an observed effect of -.38 and a 100(1-2α)% CI of [-0.48, -0.28]. Because the 100(1-2α)% CI falls completely within the EI, the researcher concludes that the means are equivalent. However, it remains unknown how far the observed effect lies from the lower boundary. In this example, the PD is -0.76 (-.38/.5), which means that the observed effect is fairly close to the lower boundary (i.e., 76% of the distance away from 0 to the lower bound). Had the observed effect been -.10 the effect would remain statistically equivalent; however, the PD would now be -0.2, which means that the observed effect is relatively far from the lower boundary (i.e., 20% of the distance from 0 to the lower bound). Although both examples are statistically equivalent, and both provide evidence that the difference in means is negligible, the latter PD provides even stronger evidence regarding the extent to which the null hypothesis is false.

Another advantage of the PD estimate, like most other ES indices, is that it is not influenced by sample size. That is, ETs with different sample sizes but the same descriptive statistics (e.g., observed effect, EI, and standard deviations) can differ in statistical significance but not in their PD estimates. For example, Wasserman and Rittenour (2019) reported a non-significant independent mean difference ET wherein one group had n = 34 participants and the other had n = 37. Just by adding 7 participants to the former group and 4 participants to the latter so that each group now has n = 41, the ET becomes significant. Yet, the proportional distance of the mean difference to its EI_{sign} never changes.

It is also important to note that the PD that uses raw values and a PD that uses standardized measures will generally be equal in value. The reason being that both the mean difference and the EI need to be divided by a variance measure to be standardized; consequently, this variance measure is cancelled out. For example, the PD for an ET conducted by Weigold, Weigold and Russell (2013) on extraversion scores between participants that completed a paper-and-pencil survey in the lab versus at home was .27. That is, the mean difference was approximately 27% away from 0 to the upper bound of the EI (mean difference =1.87; EI_{sign} = 6.916). To convert these raw values into the standardized measure, Cohen's *d*, divide the effect and EI each by the pooled SD such that mean difference = 1.87/7.07 = .26 and EI = 6.916/7.07 = .978. The standardized PD (.27) will equal to the non-standardized PD.

In conclusion, the PD measures the proportional distance from the effect of interest (e.g., mean difference, correlation) to the bound of the equivalence interval with the same sign as the effect of interest. It has several advantageous properties, with the most important being that it provides additional information not captured by either the observed ES or its associated CI. In the following section, we further investigate the statistical properties of the PD via a Monte Carlo simulation.

Confidence Interval for the PD

We used bootstrapping procedures to resample the data and determine the bounds of the confidence interval for the PD. The number of bootstrap samples should be between 1000 and 2000 (Efron & Tibshirani, 1993; Davison & Hinkley, 1997); consequently, we chose to bootstrap 2000 samples.

Bootstrapping methods for the creation of CIs fall into three families: pivotal, non-pivotal and test-inversion (Carpenter & Bithell, 2000). The pivotal family includes the most common

methods: basic and studentized bootstrapping. These methods are similar to the classical methods for the construction of CIs (Fisher et al., 2020). The non-pivotal family, on the other hand, uses percentiles of the bootstrapped distribution and includes the percentile, bias-corrected percentile and accelerated method. Lastly, the test-inversion family uses the duality between null hypothesis testing (NHST) and the CI to create CIs (Fisher et al., 2020). This family includes the test-inversion and studentized test-inversion methods. However, considering that these methods cannot be used on non-parametric resampling (Carpenter & Bithell, 2000), we avoided using them.

Of the non-pivotal families, we used the percentile bootstrapping method to generate CIs at the nominal alpha levels of .05 and .10. The percentile interval is defined as:

$$[\, Q_{\frac{\alpha}{2}}, Q_{\frac{1-\alpha}{2}} \,], \quad (6)$$

wherein Q is the quantile of the bootstrap distribution of $\hat{\theta}^*$ at a particular alpha level, α. That is, the percentile method uses the 100 x ($\frac{\alpha}{2}$) and 100 x ($1 - \frac{\alpha}{2}$) percentiles of the bootstrapped sample as the lower and upper CI endpoints, respectively (Efron & Tibshirani, 1993).

Monte Carlo Simulations

Monte Carlo studies (MCS) allow researchers to evaluate the behaviour of a statistic through the empirical process of simulating random samples based on set population parameters (Mooney, 1997; Paxton et al., 2001; Ross, 2013). We conducted two MCS to evaluate the behavior of the PD in the context of (1) mean differences and (2) correlational research. We used the R package `SimDesign` (Chalmers & Adkins, 2020) to conduct our MCS as it avoids using problematic coding strategies (e.g., nested loops; see Sigal & Chalmers, 2016) and implements optimal coding practices including the storage of seeds, etc. (Chalmers & Adkins, 2020).

Mean Difference MCS

We generated a fully-crossed simulation consisting of three population factors: total sample size (*N*), equivalence interval (EI) and population mean difference (MD). There were five conditions for *N*, specifically $N = 30$, 50, 100, 200 and 1000, with group sample sizes being equal (e.g., for $N = 50$, $n_1 = 25$, $n_2 = 25$). We investigated two EI conditions, namely EI = {-.2, .2} and EI = { -.4, .4}. We explored seven MD conditions (MD = -.6, -.4, -.2, 0, .2, .4, or .6), where a positive value indicates that the first mean is larger than the second mean, and vice versa. The EI values were arbitrarily chosen, given that there is no research context, with the mean differences chosen to be appropriate given the EI values. By combining the MD and EI conditions, there were 14 unique PD conditions (see Table 1). For example, when EI = {-.2, .2} and MD = 0, the population PD = 0, and when EI = {-.4, .4} and MD = -.4, the population PD = -1.

Correlational MCS

We also created a fully-crossed simulation for our correlation MCS that included three population factors: *N*, EI, and population correlation (ρ). There were five *N* conditions ($N = 30$, 50, 100, 200 and 1000). There were two EI conditions, EI = {-.1, .1} and EI = {-.2, .2}. In this situation, there is some research to support the choice of EI values. For example, in context-free settings, Cohen (1988) stated that the smallest meaningful correlation magnitude was .2 and Beribisky et al. (2019) found the smallest meaningful correlation magnitude to be around .3. There were seven ρ conditions, namely ρ = -.15, -.1, -.05, 0, .05, .1, .15. Consequently, the PD had 14 levels (see Table 2). For example, when ρ = 0 and EI = {-.1, .1}, the population PD = 0, whereas when ρ = .2 and EI = {-.2, .2}, the population PD = 1.

For each MCS condition, we calculated the raw and standardized bias to judge the proximity of the parameter estimates to the population generating parameters. Specifically, the raw bias is the distance between the population PD and the sample PD. The standardized bias, on the other hand, is the raw bias divided by the standard deviation of the parameter PDs. Good parameters should be close to zero to indicate that the generated estimates are neither systematically too low nor too high (Sigal & Chalmers, 2016).

Confidence Interval Coverage

We also calculated the empirical coverage rate to evaluate how well a nominal 95% and 90% CI captured the intended statistic under repeated sampling conditions. Given minimal bias and skewness within the bootstrapped samples, we expect that the percentile method should provide good coverage rates.

MCS Results

Bias

Regardless of sample size, raw and standardized biases were close to zero for both the correlation and mean-difference-based simulations (see Tables 3 and 4). In other words, the PD estimates did not substantially differ from their respective population values.

Coverage Rates

Mean Differences

In sample sizes of 200 and 1000 an average of 95% and 90% of the population PD values fell within the estimated 95% and 90% percentile CIs, respectively (see Table 5).

In the smaller samples of 30, 50 and 100, an average of 94%, 94% and 95% of the population PDs fell within the estimated 95% CIs, respectively. Likewise, an average of 88%, 89% and 89% of the population PDs fell within the estimated 90% CIs, respectively.

Correlational

Similar to the simulation conditions on mean differences, an average of 95% and 90% of the population PD values fell within the estimated 95% and 90% percentile CIs (respectively) for sample sizes of 200 and 1000 (see Table 6).

Among the samples of 30, 50 and 100, an average of 93%, 93% and 94% of the population PDs fell within the estimated 95% CIs, respectively. An average of 87%, 88% and 89% of the population PDs fell within the estimated 90% CIs, respectively.

Application of the PD

In the following section, we applied the PD to two recent equivalence testing papers published in the psychological literature. One paper explored population mean equivalence, while the other paper explored a negligible association (correlation). In addition to computing the PD, we used the percentile bootstrapped method (2000 resamples) to construct 95% CIs.

Mean Differences

Bonfils and Lysaker (2020) evaluated the equivalence of distress tolerance and emotional regulation in participants with schizophrenia/schizoaffective disorder (n = 55) and borderline personality disorder (n = 32) using the TOST approach. The authors hypothesized that people with schizophrenia-spectrum disorders would not differ from people with borderline personality disorder in terms of self-reported ability to tolerate distress (measured by a total score on the Distress Tolerance Scale) and emotion regulation (measured by a total score and six subscale scores on the Emotion Regulation Scale).

Distress Tolerance

The groups were statistically equivalent on the measure of the ability to tolerate distress (M_{diff} = .03), suggesting that participants in these two groups did not differ on their reported

ability to tolerate distress. Given that EI = {-.43, .43}, the PD for this test is .07(.03/.43) with a 95% CI of [-.40, .53]. This suggests that the observed mean difference of 0.03 is approximately 7% of the distance from 0 to the upper bound (see Figure 3). Considering that the PD is close to 0, we can conclude that the two groups are largely similar on ability to tolerate distress. Furthermore, the CI bounds of the PD provide information regarding the precision of the PD measurement; in this case, we would not expect values for PD outside larger in magnitude than about .5. That is, we would not expect the true mean difference to exceed 53% of the distance away from 0 to the upper bound, or 40% of the distance from 0 to the lower bound.

Emotion Regulation

The groups were not statistically equivalent on the total scores for the emotional regulation measure (M_{diff} = -.29). Given that EI = {-.34, .34}, the PD for this test is -.85 (-.29/|-.34|) with a 95% CI of [-1.37, -.30]. This suggests that the observed mean difference of -.29 is at approximately 85% of the distance away from 0 to the lower bound (see Figure 4). Given that the PD is close to -1, we can conclude that, after considering the EI, the two groups' scores show a much large amount of dissimilarity relative to that for distress. Furthermore, we would expect values for the PD to range from -1.37 to -.30. That is, the true mean difference may lie anywhere between 30% and 137% of the distance from 0 to the lower bound; whether the true mean difference lies beyond or within the equivalence bounds is uncertain.

Correlation

Edwards and Schatschneider (2020) conducted several equivalence tests to investigate the relationship between the magnocellular visual system and reading ability to test previous research suggestions that dyslexia may be associated with deficits in the magnocellular pathway, specialized in visual information. Magnocellular performance was measured using flicker

detection (FD) and coherent motion (CM) tasks. Reading ability, on the other hand, was measured via rapid automatic naming (RAN); isolated naming (IN); oral reading fluency (ORF); and silent reading fluency (SRF). Total scores were obtained for magnocellular performance and reading ability.

FD vs SRF

Among a sample of undergraduate students (n = 82), the scores in both tasks were statistically equivalent (r = .12), suggesting that the correlation between magnocellular performance and reading ability was negligible. Given that EI = {-.3, .3}, the PD for this test is .39 (.12/|.3|) with a 95% CI of [-.13, .94]. This suggests that the observed correlation of .12 is at approximately 39% of the distance away from 0 to the lower bound (see Figure 5). Considering that the PD is close to .5, we can conclude that magnocellular performance and reading ability, as measured by FD and SRF (respectively), are somewhat similar. Furthermore, it can be assumed that the true PD does not exceed 94% of the distance away from 0 to the upper bound or 13% of the distance to the lower bound.

Discussion

The primary objective of this paper was to propose a novel ES measure for use in ET that incorporated the width of the EI. The PD fulfilled this objective in measuring the proportional distance from the effect of interest (e.g., mean difference, correlation) to the bound of the equivalence interval with the same sign as the effect of interest. The present study successfully investigated the statistical properties of the PD, including its CI coverage and bias via MCS. We also applied the PD to two recent equivalence testing psychological papers that explored population mean equivalence and negligible association (correlation). A number of important findings emerged.

Firstly, the raw bias and standardized bias were both near 0 in both the mean difference and correlation MCS. This suggests that the average computed PD is very close to the population PD, an attractive property of any estimator. Further, at sample sizes of 200 and above, the coverage rate is extremely accurate (i.e., identical to the nominal rate up to two decimal places). When sample sizes were less than 200, CIs were slightly conservative. The overall simulation results suggest that the PD provides an unbiased estimation on how far the effect is from the EI, with CI coverage close to the targeted nominal alpha levels.

The applied examples revealed the beneficial properties of the PD, namely its ability to supply information above and beyond that of other statistical tests and ESs. For example, researchers can have greater confidence that participants with schizophrenia/schizoaffective disorder and borderline personality disorder did not differ on their reported ability to tolerate distress because the PD was close to zero and the bounds of the bootstrapped confidence interval did not exceed a magnitude of .53. On the other hand, the magnitude of the true mean difference on emotion regulation appears larger, with the bounds of the confidence interval extending as far as 137% of the distance from 0 to the lower bound.

Limitations

A limitation of the Monte Carlo simulations was that we were only able to run a subset of all potential conditions. For example, we limited our EI population factor to EI = {-.2, .2} and EI = { -.4, .4} within the mean difference research context. However, this is an expected consequence of any simulation as it is not feasible to incorporate all possible conditions. Instead, we made an informed decision on which conditions to include according to what is relatively typical in equivalence testing within the psychological literature. Furthermore, there is little

reason to believe that changing the population factors could substantially affect the performance of the proportional distance measure.

An additional limitation was that we only investigated the proportional distance within a mean difference and correlation-based research context. We encourage researchers to apply and explore its statistical properties within numerous frameworks (e.g., multiple regression, etc.).

Conclusion

Similar to traditional ESs under the NHST framework, the PD quantifies the extent to which the sample results deviate from the null hypothesis (i.e., H_0: $\mu_1 - \mu_2 \leq \Delta_L \mid \mu_1 - \mu_2 \geq \Delta_U$). More specifically, the PD provides information regarding how far the observed effect falls from an effect size of 0 to the bounds of the EI. Estimated CIs provide further information regarding the precision associated with the PD measurement.

We encourage researchers to calculate the PD to supplement their tests of statistical equivalence and traditional ESs with information that incorporates the respective EI. We also encourage researchers to estimate and interpret the associated CIs. It is important to highlight that the PD falls under the larger scope of estimation, not NHST. The goal is to quantify the magnitude of an effect; namely, how far the effect is from the bounds of the EI. Therefore, including the PD is consistent with the movement towards estimation and away from solely considering NHST results (e.g., American Psychological Association, 2020; Cumming, 2014). We hope that the results and recommendations of this research will help provide information above and beyond equivalence tests and traditional ESs in the field of psychology.

References

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Washington, DC: Author.

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics - Theory and Methods*, *12*(23), 2663–2692. https://doi.org/10.1080/03610928308828634

Ball, L. C., Cribbie, R. A., & Steele, J. R. (2013). Beyond gender differences: using tests of equivalence to evaluate gender similarities. *Psychology of Women Quarterly*, *37*(2), 147–154. https://doi.org/10.1177/0361684313480483

Beribisky, N., Davidson, H., & Cribbie, R. A. (2019). Exploring perceptions of meaningfulness in visual representations of bivariate relationships. *PeerJ*, *7*, e6853. https://doi.org/10.7717/peerj.6853

Bonfils, K. A., & Lysaker, P. H. (2020). Levels of distress tolerance in schizophrenia appear equivalent to those found in borderline personality disorder. *Journal of Clinical Psychology*, *76*(9), 1668–1676. https://doi.org/10.1002/jclp.22944

Briones, E. M., & Benham, G. (2017). An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples. *Behavior Research Methods*, *49*(1), 320–334. https://doi.org/10.3758/s13428-016-0710-8

Campbell, H., & Gustafson, P. (2018). What to make of non-inferiority and equivalence testing with a post-specified margin? *arXiv preprint arXiv:1807.03413*.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, *19*(9), 1141-1164. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates. http://dx.doi.org/10.4324/9780203771587

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, *60*(1), 1–10. https://doi.org/10.1002/jclp.10217

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, Taylor & Francis Group.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (No. 1). Cambridge university press.

Edwards, A. A., & Schatschneider, C. (2020). Magnocellular pathway and reading rate: an equivalence test analysis. *Scientific Studies of Reading*, *24*(3), 264–273. https://doi.org/10.1080/10888438.2019.1663856

Efron, B., & Tibshirani, R. (1993). An introduction to the bootstrap. Chapman & Hall.

Farmus, L., Beribisky, N., Gutierrez, N. M., Alter, U., Panzarella, E., & Cribbie, R. (2020, November 9). Effect Size Reporting and Interpretation in Social Personality Research. https://doi.org/10.31234/osf.io/nvczj

Fisher, E., Schweiger, R., & Rosset, S. (2020). Efficient construction of test inversion confidence intervals using quantile regression. *Journal of Computational and Graphical Statistics*, *29*(1), 140–148. https://doi.org/10.1080/10618600.2019.1647215

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: an equivalence testing approach. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 527–537. https://doi.org/10.1348/000711009X475853

Gøtzsche, P. C. (2006). Lessons from and cautions about noninferiority and equivalence randomized trials. *JAMA*, *295*(10), 1172. https://doi.org/10.1001/jama.295.10.1172

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152. https://doi.org/10.1037/a0028086

Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759. https://doi.org/10.1177/0013164496056005002

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: a tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lee, D. K. (2016). Alternatives to P value: Confidence interval and effect size. *Korean Journal of Anesthesiology*, *69*(6), 555. https://doi.org/10.4097/kjae.2016.69.6.555

Le Henanff, A., Giraudeau, B., Baron, G., & Ravaud, P. (2006). Quality of reporting of noninferiority and equivalence randomized trials. *JAMA*, *295*(10), 1147. https://doi.org/10.1001/jama.295.10.1147

Lüdtke, T., Westermann, S., Pult, L. K., Schneider, B. C., Pfuhl, G., & Moritz, S. (2018). Evaluation of a brief unguided psychological online intervention for depression: A controlled trial including

exploratory moderator analyses. *Internet Interventions*, *13*, 73–81. https://doi.org/10.1016/j.invent.2018.06.004

Meyners, M. (2007). Least equivalent allowable differences in equivalence testing. *Food Quality and Preference*, *18*(3), 541–547. https://doi.org/10.1016/j.foodqual.2006.07.005

Meyners, M. (2012). Equivalence tests – a review. *Food Quality and Preference*, *26*(2), 231–245. https://doi.org/10.1016/j.foodqual.2012.05.003

Mooney, C. Z. (1997). *Monte carlo simulation* (No. 116). Sage.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. https://doi.org/10.1037/0033-2909.113.3.553

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *57*(3), 221–237. https://doi.org/10.1037/h0087427

Ross, S. M. (2013). *Simulation* (Fifth edition). Academic Press.

Schuirmann, D. J. (1987). A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. https://doi.org/10.1007/BF01068419

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: teaching statistics with Monte Carlo simulation. *Journal of Statistics Education*, *24*(3), 136–156. https://doi.org/10.1080/10691898.2016.1246953

Silva, L. M. T., Ayres, R., & Schalock, M. (2008). Outcomes of a pilot training program in a qigong massage intervention for young children with autism. *American Journal of Occupational Therapy*, *62*(5), 538–546. https://doi.org/10.5014/ajot.62.5.538

Staszewski, S. (2001). abacavir-lamivudine-zidovudine vs indinavir-lamivudine-zidovudine in Antiretroviral-Naive HIV-Infected adults: a randomized equivalence trial. *JAMA*, *285*(9), 1155. https://doi.org/10.1001/jama.285.9.1155

Thompson, B. (2002). What future quantitative social science research could look like: confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25–32. https://doi.org/10.3102/0013189X031003025

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, *26*(2), 192–196. https://doi.org/10.1007/s11606-010-1513-8

Wasserman, J. A., & Rittenour, C. E. (2019). Who wants to play? Cueing perceived sex-based stereotypes of games. *Computers in Human Behavior*, *91*, 252–262. https://doi.org/10.1016/j.chb.2018.09.003

Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychological Methods*, *18*(1), 53–70. https://doi.org/10.1037/a0031607

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. CRC press.

Westlake, W. J. (1971). Problems associated with analysis of pharmacokinetic models. *Journal of Pharmaceutical Sciences*, *60*(6), 882–885. https://doi.org/10.1002/jps.2600600616

Figure 1

PD values plotted against EI_{sign} values, holding the mean difference constant at .01

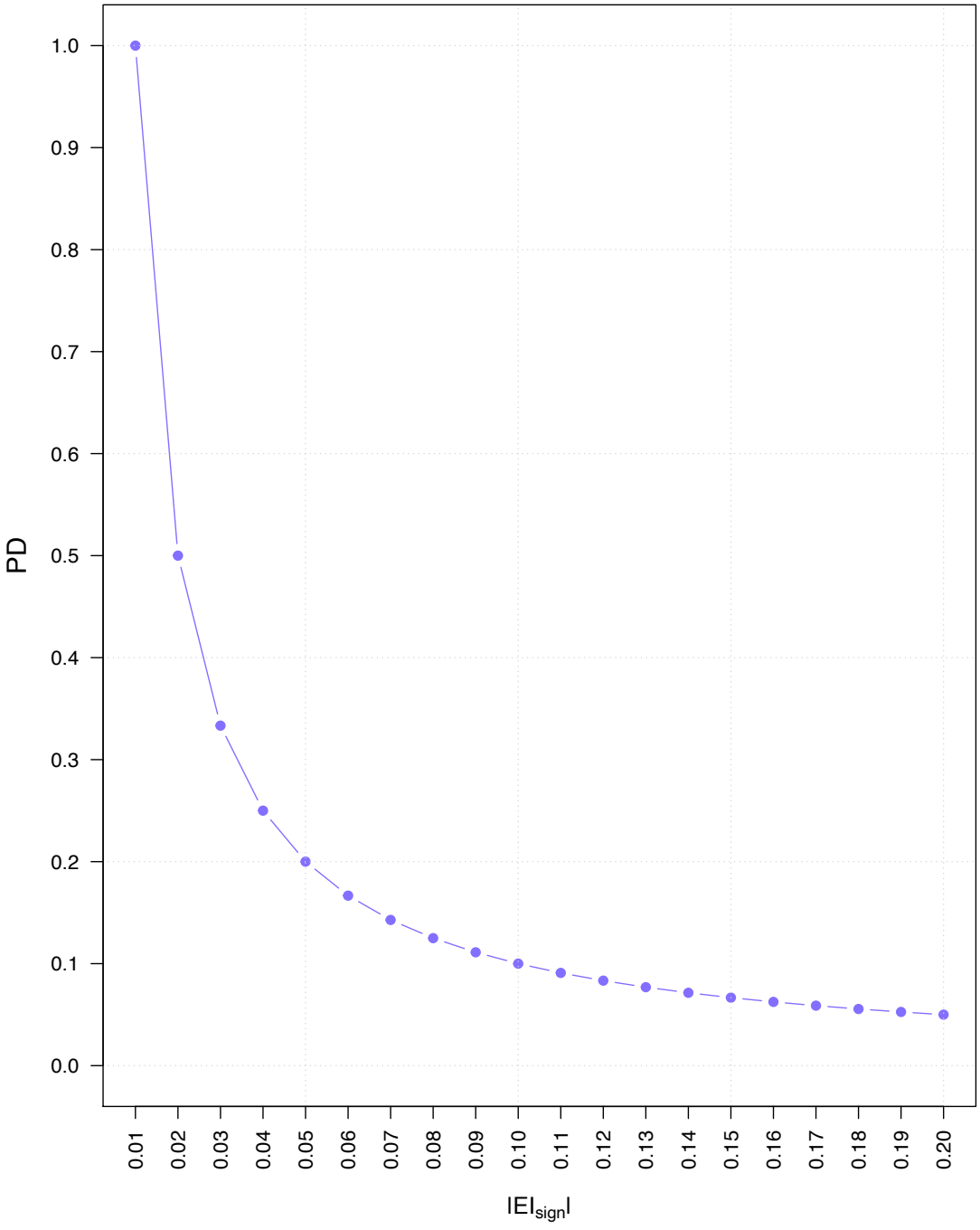

Figure 2

PD values plotted against EI_{sign} values, holding the mean difference constant at -.01

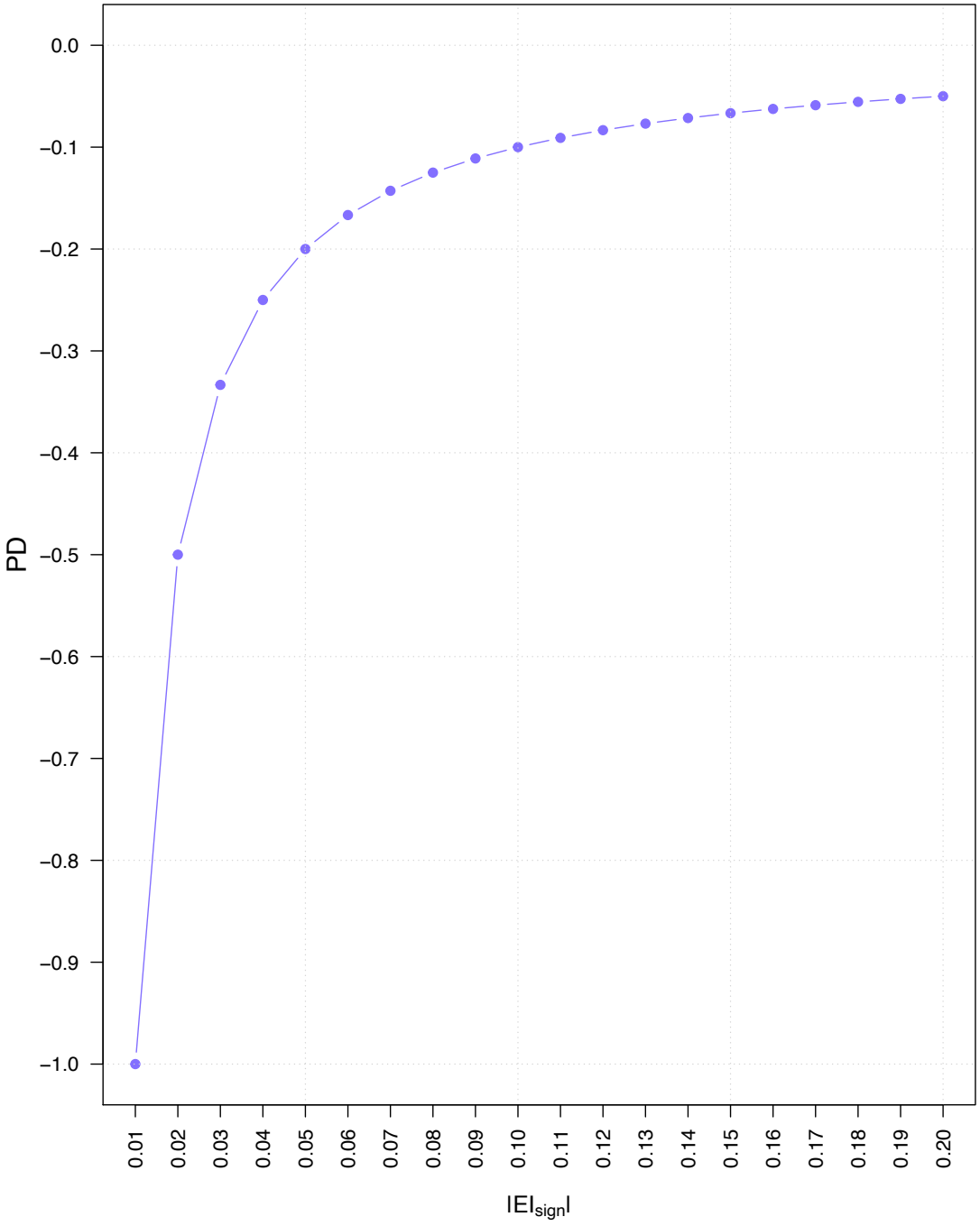

Figure 3

The proportional distance of the mean difference in distress tolerance

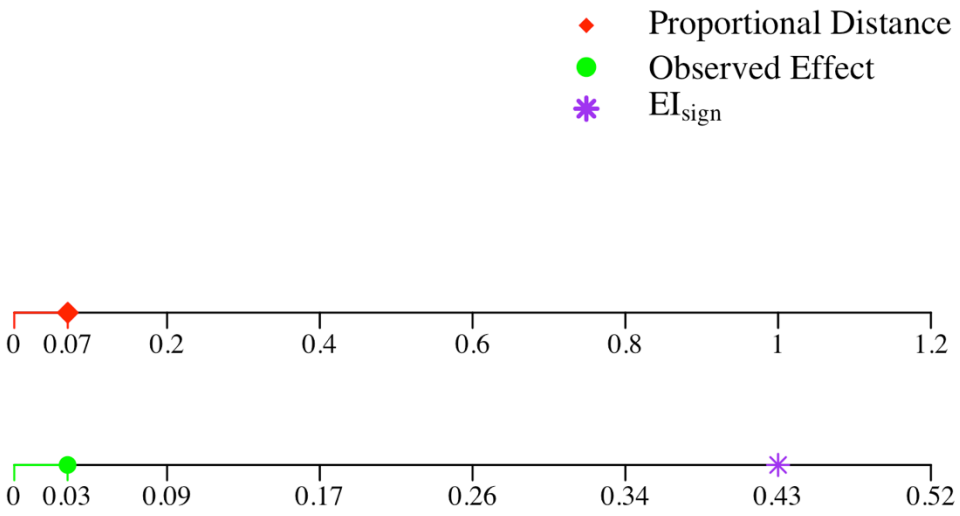

Figure 4

The proportional distance of the mean difference in emotion regulation

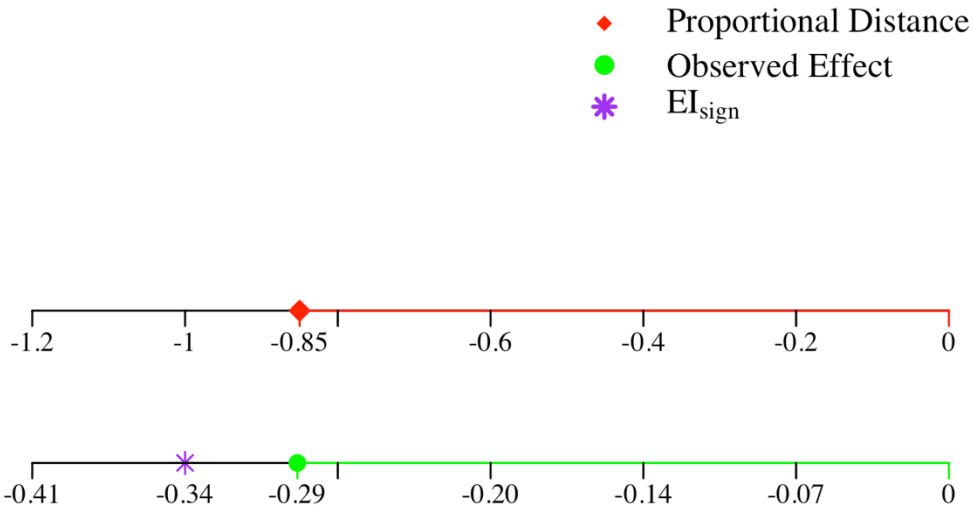

Figure 5

The proportional distance of the association between magnocellular performance and reading ability

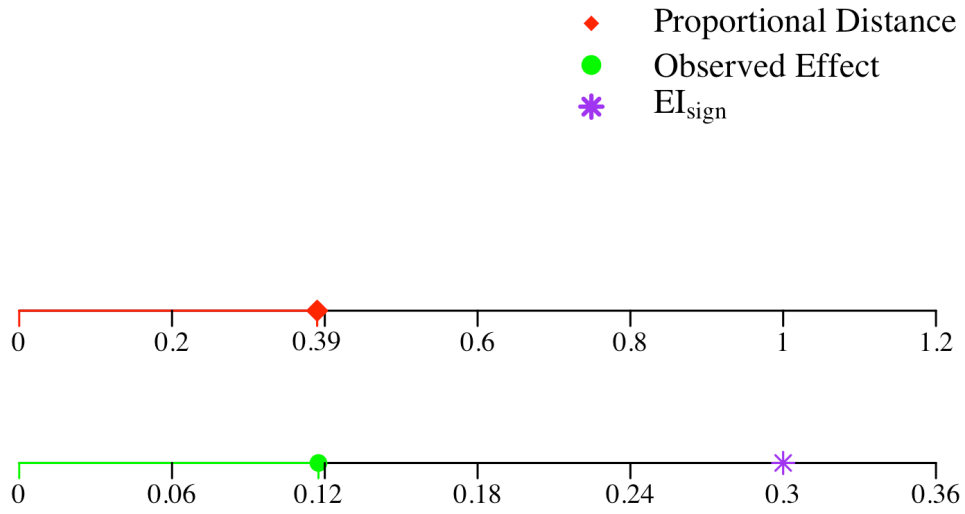

Table 1

The PD associated with each EI and MD condition for the Mean Difference Monte Carlo Study

EI	MD	PD
0.2	-0.6	-3
	-0.4	-2
	-0.2	-1
	0	0
	0.2	1
	0.4	2
	0.6	3
0.4	-0.6	-1.5
	-0.4	-1
	-0.2	-0.5
	0	0
	0.2	0.5
	0.4	1
	0.6	1.5

Note. EI = Equivalence Interval; MD = Mean Difference; PD = Proportional Difference.

Table 2

The PD associated with each EI and ρ condition for the Correlation-based Monte Carlo Study

EI	ρ	PD
0.1	-0.15	-1.5
	-0.10	-1
	-0.05	-0.5
	0	0
	0.05	0.5
	0.10	1
	0.15	1.5
0.2	-0.15	-0.75
	-0.10	-0.5
	-0.05	-0.25
	0	0
	0.05	0.25
	0.10	5
	0.15	0.75

Note. EI = Equivalence Interval; *ρ* = Population Correlation; PD = Proportional Difference.

Table 3

Descriptive Statistics for Bias and Standardized Bias by Sample Size for the Mean Difference Monte Carlo Study

N	Bias				Standardized Bias			
	Mean	SD	Min	Max	Mean	SD	Min	Max
30	0.00	0.03	-0.04	0.05	0.00	0.02	-0.02	0.03
50	0.00	0.01	-0.02	0.02	0.00	0.01	-0.02	0.03
100	0.00	0.01	-0.02	0.02	0.00	0.01	-0.02	0.03
200	0.00	0.01	-0.02	0.01	-0.01	0.02	-0.03	0.02
1000	0.00	0.00	-0.01	0	-0.01	0.02	-0.02	0.01

Table 4

Descriptive Statistics for Bias and Standardized Bias by Sample Size for the Correlation-based Monte Carlo Study

N	Bias				Standardized Bias			
	Mean	SD	Min	Max	Mean	SD	Min	Max
30	0.00	0.03	-0.05	0.04	0.00	0.02	-0.03	0.05
50	-0.01	0.01	-0.03	0.01	-0.01	0.01	-0.03	0.01
100	0.00	0.01	-0.03	0.01	0.00	0.02	-0.03	0.01
200	0.00	0.01	-0.02	0.01	0.00	0.02	-0.03	0.03
1000	0.00	0.00	0.00	0.01	0.00	0.01	-0.02	0.02

Table 5

Descriptive Statistics for 95% and 90% CI Coverage Rate by Sample Size for the Mean Difference Monte Carlo Study

N	95% Percentile CIs				90% Percentile CIs			
	Mean	SD	Min	Max	Mean	SD	Min	Max
30	0.94	0.00	0.93	0.94	0.88	0.01	0.87	0.89
50	0.94	0.00	0.93	0.95	0.89	0.00	0.88	0.90
100	0.95	0.00	0.94	0.95	0.89	0.00	0.89	0.90
200	0.95	0.00	0.94	0.95	0.90	0.00	0.89	0.90
1000	0.95	0.00	0.95	0.96	0.90	0.01	0.89	0.91

Table 6

Descriptive Statistics for 95% and 90% CI Coverage Rate by Sample Size for the Correlation-based Monte Carlo Study

N	95% Percentile CIs				90% Percentile CIs			
	Mean	SD	Min	Max	Mean	SD	Min	Max
30	0.93	0.00	0.92	0.93	0.87	0.00	0.87	0.88
50	0.93	0.00	0.92	0.94	0.88	0.00	0.87	0.89
100	0.94	0.00	0.94	0.95	0.89	0.00	0.88	0.90
200	0.95	0.00	0.94	0.95	0.90	0.00	0.89	0.90
1000	0.95	0.00	0.94	0.95	0.90	0.00	0.89	0.91