Single case neuropsychology: Validity of hypothesis testing in a power-wise pinioned field

B146628[a,b]

[a] *Human Cognitive Neuropsychology, MSc (2020)*
[b] *School of Philosophy, Psychology and Language Sciences, University of Edinburgh*

Abstract

Single case studies where a brain damaged patient is compared to a normative control group have historically been crucial for neuropsychology. Some authors have argued that the methodology is fundamentally flawed, others that it is the only way to draw valid inferences of cognition. Whether the case, the usage of the method has declined. Reasons for this might be several but difficulties in working with a clinical population and publication pressures have been cited as possible culprits. This is while statistical methods for the paradigm have been appropriately refined and developed, mainly by John Crawford and Paul Garthwaite, during the last 22 years. These methods have given a field often viewed as "merely anecdotal" a much needed rigour. The present work first aims to evaluate these tools with regards to statistical power, keeping with the increased focus on power for evaluating validity of scientific fields. In general power was, unexpectedly, found to be low. Low power is known to generate consistent overestimations of effect sizes, so the extent to which such error estimation is affecting the field was quantified and correction methods suggested. Secondly, to encourage the use of the refined tools developed, the present work aims to make them more accessible by i) implementing recommended hypothesis tests along with power calculators in an R package: "`singcar`" ii) providing a basic conceptual intuition behind the need for their development. Lastly, the future of single case neuropsychology is discussed.

Key words: single case neuropsychology; statistical power; Monte Carlo simulation; winner's curse bias correction; singcar

[*] Word count: 11928

1. Introduction

Cognitive neuropsychology is the study of the normal cognitive system and its organisation. Some researchers are interested in the anatomical aspects of it and how this relates to behavioural functionality, but often the goal is solely to map the cognitive architecture of functions. This has to a large extent been done by studying disruption of the system in brain damaged patients. The logic is simple: if specific damage leads to a deficit on some cognitive functionality, it seems reasonable to suggest that this brain area is involved in that function. Similarly, if this damage does not affect some other cognitive function, these two functions might be dissociable. Finding a deficit can have inferential strength in itself, especially if the anatomical structure is of main interest. However, finding dissociable functions is more important for cognitive theory building. This concept is called a "dissociation" and is often cited as the gold standard of neuropsychological evidence.

Due to the rarity of focal damage giving rise to theoretically interesting behavioural patterns and the fact that the variability between patients often is large, neuropsychology has historically relied heavily on single case research designs. The inter-case variability even led some influential neuropsychologists to state that the single case study is the only way to draw valid inferences about cognitive structure (Caramazza, 1986; Caramazza & McCloskey, 1988). One of these authors also formalised assumptions that must be met for inferences from single case neuropsychology to validly inform theories of normal cognition. These assumptions are of: fractionation; modularity; transparency and universality (Caramazza, 1986, 1984). In short: brain damage can have selective effects on cognitive processing because complex functionality can be broken down into smaller separable parts, impairments of which would be the sole modulation compared to a healthy brain and this system is generally organised the same way across individuals. These assumptions were revisited by Coltheart (2017) and argued to still hold. There are, however, those that question Caramazza's strong viewpoints (Caplan, 1988; Kosslyn & Intriligator, 1992; Robertson et al., 1993) and also the usefulness of the paradigm (Goldberg, 1995; Patterson & Plaut, 2009; Seidenberg, 1988).

Not many single case researchers shares the extreme view of Caramazza and even though single case neuropsychology possibly have been under more scrutiny as a method for scientific inquiry than most there are few, if any, that deny its contribution to cognitive theory building. Much seminal work stem from single cases. Textbook examples include Phineas Gage (Damasio et al., 1994), patient H.M (Scoville & Milner, 1957) and Victor Leborgne (Broca, 1861) among others. And many influential theories of e.g. working memory (Baddeley et al., 1988) and visual perception (Goodale et al., 1992) have been heavily influenced by single cases.

Even so, usage of the methodology has declined over the last decades, especially since the entrance of functional brain imaging technologies (Chatterjee, 2005; Fellows et al., 2005; Medina & Fischer-Baum, 2017). The reasons for this are cited as sociological (e.g. publication pressures) and pragmatic in nature rather than being theoretically substantiated. In the referenced reports the authors stress that single case neuropsychology should be seen as complementary rather than competing to imaging studies and that its theoretical value is of importance still. Perhaps even more so when the pressure for new discoveries promotes findings of trivial brain correlates, fueled by an increasing demand of publications with pictures of illuminated brains.

The present work will not be concerned with the legitimacy of the field in terms of its theoretical assumptions. Instead focus will be directed to practical issues of *finding* deficits and dissociations in single cases. In particular the issues of hypothesis testing. For some cases, e.g. patient H.M and his inability to store recent information, one would not need a hypothesis test. We do not need a statistical test to tell us that the earth is round (see Cohen, 1994). However, finding patients with such detrimental and theoretically interesting effects is more exception than the rule and often researchers need a way to estimate how abnormal a case is in comparison to the neurologically intact population.

This is sometimes done by using normed tests from which population data exist and one can estimate abnormality directly. But often a researcher might need to use a test where this is not possible and must hence compare the case to a normative control sample. It is this type of statistical testing, refined and developed during the last 22 years by mainly John Crawford and Paul Garthwaite, that will here be presented and examined.

When estimating abnormality of a case one should preferably compare him/her to a control population of the same premorbid cognitive ability (i.e. education, age etc.), which can be made apparent by observing figure 1. Finding matched controls can be difficult and, hence, control samples in these studies have typically been small (Crawford et al., 2011).

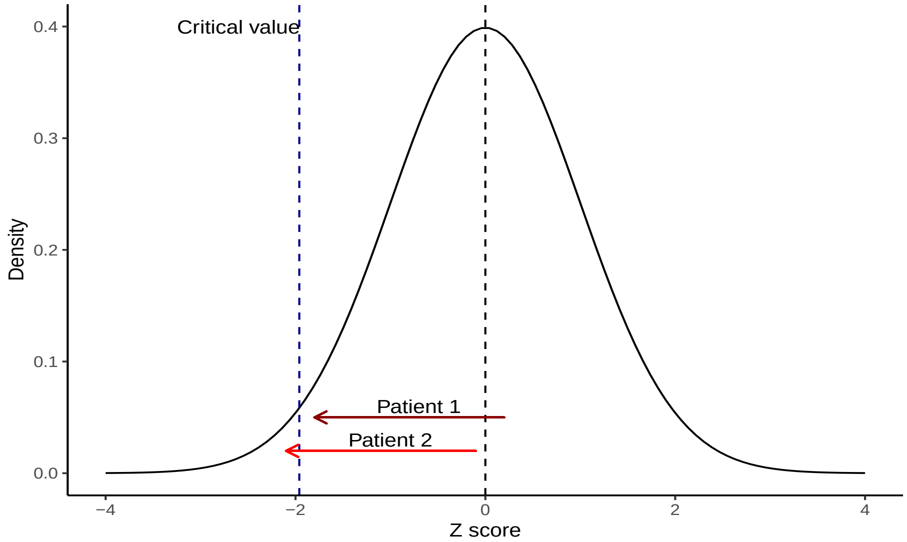

Figure 1: Imagine the bell curve representing the distribution of scores from some task in the normal (unmatched) control population. The red arrows go from the pre- to postmorbid scores for two patients with a deficit of the exact same size of 2 standard deviations (SD) but with a premorbid difference of 0.3 SD. Patient 2 only would be deemed to have a deficit at a significance level of 0.05 (two-sided).

From figure 1 one can deduce that trivially small premorbid task differences between two cases have all the difference in the world for getting diagnosed with a deficit, if using binary cutoff values. The probability of finding this deficit (given that some specific damage always would give rise to the same size of functional impairment) could thus never exceed $\sim 50\%$. That is, only the population that have a premorbid score at or below the mean would be “found”. This probability is called *statistical power* and is discussed further in section 2. The concept of a deficit is easy to grasp, but as noted, dissociations are more sought after for cognitive theory building and we will turn to these next.

1.1. Defining dissociations

Claiming a dissociation is slightly more complex than claiming a deficit. This is because one needs to show how a patient scores on one task *in relation* to another. Typically this has been done by showing that the patient is impaired on task A but performs at premorbid level on task B. This is of course difficult to control without premorbid data so if the post morbid patient performs within the normal range on task B, a deficit on task B is ruled out and a dissociation between A and B said to exist. A number of problems arise with this. First, it includes confirming the null hypothesis for task B (that he/she does not have a deficit), this is not possible since one solely can fail to reject it. Second, with this definition the discrepancy between the tasks might be trivially small if the patient falls just outside the normal range on task A and just within on task B.

These issues were somewhat recognised by Shallice (1988) (p. 228) when he defined two types of dissociations that since then have become influential. He termed these two types *classical* and *strong*. Defining these Shallice (1988) still use the typical criterion that there needs to be a deficit on at least one task, but he also noted that a substantial difference *between* the tasks should be exhibited, hence amending the problem of trivially small differences. However, he does not operationalise what this "substantial difference" means, neither does he operationalise a deficit. The *classical* and the *strong* definitions are illustrated in figure 2.

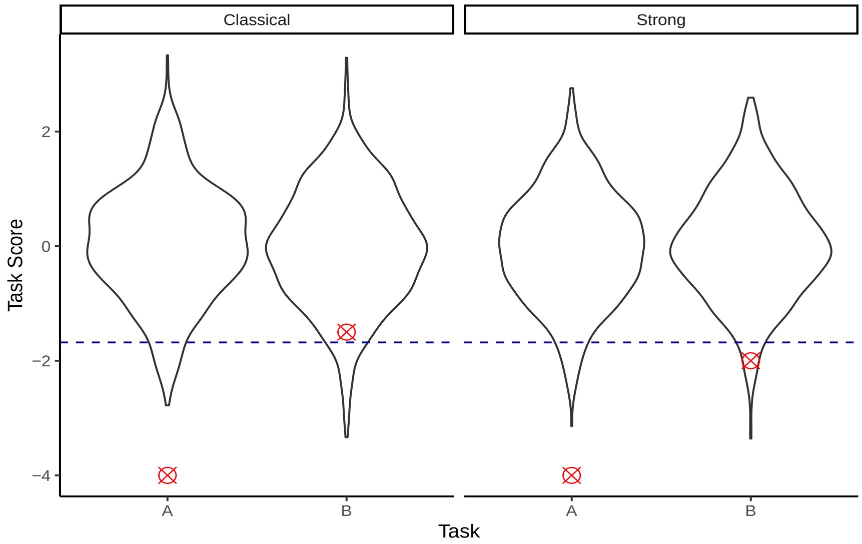

Figure 2: Visualisation of the concepts *classical* and *strong* dissociations introduced by Shallice (1988). The violin plots represent the distribution of the normal population and the crossed circles the case scores. For a dissociation to be termed *classical* the patient is impaired on one task, i.e. well below the normal range, but not on the other. For a dissociation to be termed "strong" the patient is impaired on both tasks but much more so on one of them. The dashed blue line represents a critical value.

To make these definitions more practically useful Crawford et al. (2003) used inferential statistical criteria to define both functional deficits and task discrepancies. The details of the hypothesis tests devised for this purpose are discussed in section 1.2. In short they operationalised Shallice (1988)'s concepts of dissociations by defining a deficit as a case being significantly below the normal population on some task and defining a “substantial difference” between two tasks as a case having a significantly more extreme task discrepancy compared to the normal population.

Although Crawford et al. (2003)’s efforts gave the definitions much needed rigour, why would criteria for a dissociation need to involve deficits at all? Basic probability theory states that the probability for two events to occur never can exceed the probability of one of them to occur. It seems more reasonable that testing for a dissociation need only be contingent on the size of the task discrepancy for the case compared to the distribution of tasks discrepancies within the normative population. This critique was brought up by McIntosh (2018) were it was shown that basing dissociations solely on task discrepancy generated higher statistical power. This should be uncontroversial given the argument above.

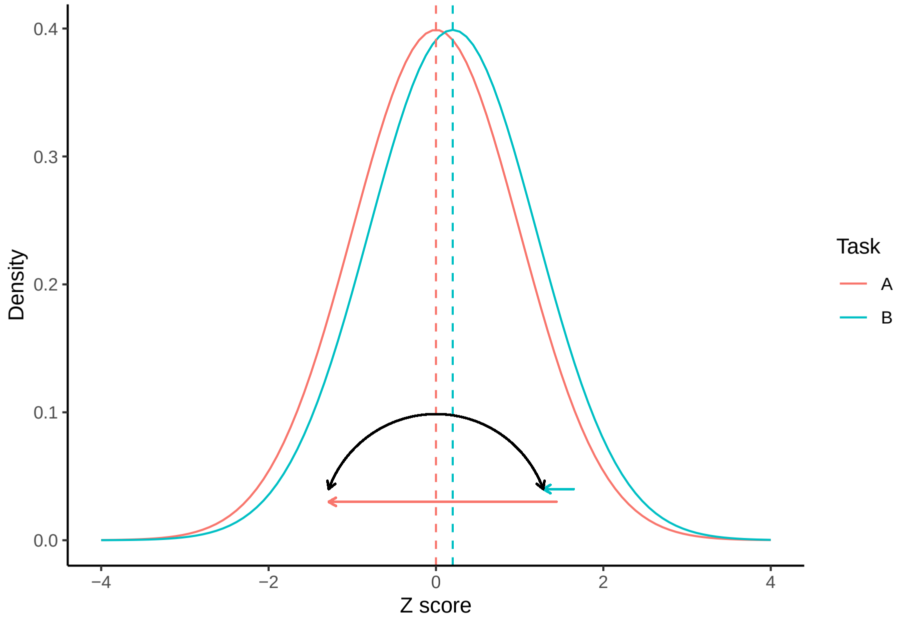

Figure 3: Visualisation of a patient exhibiting a significant discrepancy between two tasks, without having a deficit on either. The single headed arrows go from pre- to postmorbid ability for the case, the double headed arrow represents the task discrepancy for the case and the bell curves represent the distribution of scores for the two tasks within a normative population.

Figure 3 shows the concept of assessing task discrepancy where no deficit can be found. It can be seen that even though a discrepancy between the tasks exists within the normative population, the postmorbid discrepancy for the case is much larger than we would expect if he/she came from this population. However, one should not use these raw discrepancies for significance testing since such comparison would not take the correlation between the tasks into account — something that was overlooked by Shallice (1988) but remedied by Crawford, Howell, et al. (1998).

Hopefully, it should now be clear that these stringent compound criteria of dissociations brought forward by Shallice (1988) and operationalised by Crawford et al. (2003) are somewhat unfounded. Therefore, a dissociation will henceforth be defined by McIntosh (2018)'s simple criteria of task discrepancy only.

1.2. Statistical methods for finding deficits and dissociations

The aim of this section is not to provide an exhaustive mathematical understanding of the formulas used but rather a conceptual understanding of the need for their development.

1.2.1. Frequentist approaches

In the latter part of the 90's Crawford and colleagues started developing statistical tests to use for evaluating single cases that were compared to normative control samples. Typically, the prior methods used treated the distribution estimated by the the control group as if the sample statistics were the population parameters. That is, the estimated distribution was treated as a standard normal distribution from which abnormality of the case score was estimated by:

$$z = \frac{x^* - \overline{x}}{\sqrt{s^2}} \tag{1}$$

This is similar to the familiar z-formula but here x^*, $\overline{x}$ and s^2 is the case score, sample mean and sample variance respectively. $\overline{x}$ and s^2 is plugged in directly as the population parameters μ and σ^2 in the normal z-formula. The p-value obtained from the z-value would then be treated as the estimation of the case's abnormality. This is problematic because the sampling distribution of s^2 is right skewed for small sample sizes. This means that underestimation of s^2 is more probable than overestimation and hence the z-value would often be larger than it should, resulting in an overestimation of the abnormality and inflation of Type I errors (claiming that there is an effect when, in fact, there is not) (Crawford & Howell, 1998).

With a similar logic as in (1), Payne & Gwynne Jones (1957) developed a method for assessing abnormally large discrepancies between two tasks. I.e. a test that estimates the proportion of the control population that would exhibit a greater discrepancy than the case, as seen in (2).

$$z_{disc} = \frac{(z_a^* - z_b^*) - \cancelto{0}{(\overline{z}_a - \overline{z}_b)}}{\sqrt{2 - 2r_{ab}}} \tag{2}$$

Where z_a^* and z_b^* are the standardised case scores on task A and B respectively, $\overline{z}_a$ and $\overline{z}_b$ the means from the sample on the two tasks (which both equates 0 because of standardisation) and r_{ab} the correlation between the two tasks calculated from the sample scores. However, this test suffers from the same problem mentioned above and would overestimate the abnormality of the task discrepancies.

A different approach to comparing a single observation to the mean of a sample was proposed by Sokal & Rohlf (1981) (p. 227) and popularised within neuropsychology by Crawford & Howell (1998). Here the t-distribution (with its fatter tails) is utilised to account for the underestimation of the sample variance. The approach is a modified two samples t-test where the case simply is treated as a sample of size 1. The degrees of freedom for this distribution is $n + 1 - 2 = n - 1$.

$$t_{n-1} = \frac{X^* - \overline{X}}{s\sqrt{\frac{n+1}{n}}} \tag{3}$$

This test of deficit (TD) has been shown to not exceed the specified error rate α unlike other similar tests (Crawford et al., 2004; Crawford, Garthwaite, & Howell, 2009; Crawford & Garthwaite, 2012). Together with its simplicity this makes it a superior choice over many other ways of detecting outliers in small samples. One of its main advantages is that it provides the researcher with an *unbiased* point estimate of the abnormality of the case.

Crawford, Howell, et al. (1998) extended this to Payne & Gwynne Jones (1957) test of task discrepancy or with Crawford, Howell, et al. (1998)'s denotation: 'difference'[1]. I.e. they devised a test that treated sample estimations as statistics

[1] In Crawford, Howell, et al. (1998) they use the term 'difference' instead of discrepancy. This is a somewhat unfortunate usage since a difference can pretty much refer to anything. Hence, 'discrepancy' will be used for the most part when referring to the difference between scores from two tasks

rather than population parameters for dissociations as well, seen in (4).

$$t_{n-1} = \frac{(z^*_a - z^*_b) - \cancelto{0}{(\overline{z}_a - \overline{z}_b)}}{\sqrt{(2 - 2r_{ab})(\frac{n+1}{n})}} \tag{4}$$

Unfortunately the standardised task scores of the case z^*_a and z^*_b suffer from the same problem described for (1) and Type I errors would again be inflated. However, standardisation of the scores is only necessary if the two tasks are measured on different scales. If they are measured on the same the test holds and we have the unstandardised difference test (UDT):

$$t_{UDT_{n-1}} = \frac{(x^*_a - \overline{x}_a) - (x^*_b - \overline{x}_b)}{\sqrt{(s^2_a + s^2_b - 2s_a s_b r_{ab})(\frac{n+1}{n})}} \tag{5}$$

The denominator in the (5) collapse to the denominator in (4) since s^2_a and s^2_b become 1 after standardisation. However, since assessment of task discrepancy between tasks measured on different scales is common, a test that could take standardised scores but still account for the skewness in the sampling distribution of the sample variance was needed. In Garthwaite & Crawford (2004) the authors examined the difference between two correlated, t distributed variables and aimed to derive a quantity with a distribution that would not depend on any population parameters. The math behind this derivation is too technical to be covered here, but in summation they used asymptotic expansion to find a function of the correlation between the variables that when used as a denominator to $(t_1 - t_2)$, where t_1 and t_2 are our t distributed variables, would approximate a t-distribution. The quantity found was:

$$\psi = \frac{\frac{(x^*_a - \overline{x}_a)}{s_a} - \frac{(x^*_b - \overline{x}_b)}{s_b}}{\sqrt{(\frac{n+1}{n})\left((2-2r) + \frac{2(1-r^2)}{n-1} + \frac{(5+y^2)(1-r^2)}{2(n-1)^2} + \frac{r(1+y^2)(1-r^2)}{2(n-1)^2}\right)}} \tag{6}$$

Where r is the correlation between the tasks and y the critical two-tailed t-value with $n-1$ degrees of freedom. Garthwaite & Crawford (2004) demonstrate that $\mathbb{P}[\psi > y] \approx \mathbb{P}[t > y]$, where $\approx$ indicates approximate equivalence. To obtain a precise probability for ψ one solves for $\psi = y$. See Garthwaite & Crawford (2004) and Crawford & Garthwaite (2005) for details. Choosing the positive root of $\psi = y$ yields:

$$
\begin{aligned}
y &= \sqrt{\frac{-b+\sqrt{b^2-4ac}}{2a}}, \text{where} \\
a &= (1+r)(1-r^2), \\
b &= (1-r)[4(n-1)^2+4(1+r)(n-1)+(1+r)(5+r)], \\
c &= -2\left[\frac{X_A^*-\overline{X}_A}{s_A}-\frac{X_B^*-\overline{X}_B}{s_B}\right]^2\left(\frac{n(n-1)^2}{n+1}\right)
\end{aligned}
\tag{7}
$$

Where y is used as a t-statistic. This quantity is referred to as the revised standardised difference test (RSDT). Crawford & Garthwaite (2005) show with Monte Carlo simulations that this test is superior to both their own previous test (Crawford, Howell, et al., 1998) and Payne & Gwynne Jones (1957)'s in controlling Type I errors. Even for very small sample sizes of $n = 5$, RSDT was shown to barely exceed the specified 5% error rate. So three valid frequentist tests remain, for deficits that is TD (3), for dissociations that is UDT (5) and RSDT (7).

For TD and UDT, Crawford & Garthwaite (2002) utilise the non-central t-distribution to set confidence limits on the abnormality of the case, something that grows ever more essential for reminding us that all test results come with uncertainty. However, it did not prove possible to set confidence limits on the estimations from the RSDT since it is only approximately t-distributed. This is one of the reasons why Crawford & Garthwaite (2007) wanted to develop a Bayesian method for estimating abnormality of discrepancy.

The test of deficit is most commonly used one-sided but the dissociation tests should in most cases be used two-sided as the direction of the effect solely depends on the order of the task scores.

1.2.2. Bayesian approaches

There is one main difference between Bayesian and frequentist statistical inference. In the Bayesian framework parameters (like means and standard deviations) are treated as random variables with associated probability distributions and if we believe a parameter has a certain distribution we update that belief as more data is gathered and thus the parameter values can change. Whilst, in the frequentist framework, parameters are treated as fixed attributes of a population, estimations of which in a series or frequency (hence the name) of trials will converge to the *true* value.

An intuitive way of understanding the difference is to note how the two approaches phrase intervals around parameters. The frequentist 95% *confidence interval* would cover the population parameter 95% of the time. That is, if you estimate a population mean from 100 different samples and create a confidence interval around each estimation, 95 of these intervals would include the true population mean. The Bayesian 95% *credible interval*, however, has a more intuitive interpretation. For this interval you say that with a 95% certainty the population mean is covered by the interval. That is because you take the values at the 2.5th and 97.5th percentile of the parameter distribution to form the interval.

To estimate a parameter distribution Bayesians use prior knowledge of that parameter, i.e. they assign probabilities to possible values of the parameter depending on this knowledge — forming what is known as a prior distribution, or simply a prior. If no information exists one often apply a non-informative prior, the most simple of which assigns equal probabilities (uniform) to all possible parameter values. This is then updated when new information is obtained. The distribution formed by the updated prior is called the posterior distribution. The posterior probability of a hypothesis (i.e. a specified value of the parameter) is calculated by using Bayes theorem:

$$\underbrace{\mathbb{P}[H \mid E]}_{\text{Posterior}} = \frac{\overbrace{\mathbb{P}[E \mid H]}^{\text{Likelihood}} \times \overbrace{\mathbb{P}[H]}^{\text{Prior}}}{\underbrace{\mathbb{P}[E]}_{\text{Marginal likelihood}}} \tag{8}$$

The H stands for hypothesis which may be affected by E, evidence, i.e. the data (not used to estimate the prior). $\mathbb{P}[E \mid H]$ is the probability of observing the data, given a hypothesis. $\mathbb{P}[H \mid E]$ is the probability of a hypothesis given that some data have been observed. Say for example that we want to estimate IQ in a population with a sample of $n = 3$. By calculating the mean, standard deviation and standard error from this sample we can create a sampling distribution of the mean. The marginal likelihood, $\mathbb{P}[E]$ will be a constant since it does not contain H, we can therefore disregard that for now. If we assume a uniform prior, i.e. we do not believe that any hypothesis is more likely than another, $\mathbb{P}[H]$ will also be a constant, reducing (8) to:

$$\mathbb{P}[H \mid E] = \mathbb{P}[E \mid H]$$

Say that our sample had IQs of $x_1 = 110,\ x_2 = 115,\ x_3 = 120$. The sample has a mean of 115, a standard deviation of 5 and a standard error of $5/\sqrt{n}$. Our theoretical sampling distribution of the mean (i.e. the distribution of means if we draw a sample of $n = 3$ an infinite number of times) would thus be a normal distribution with mean 115 and standard deviation = standard error = $5/\sqrt{n}$. To get the posterior distribution we now calculate the probability of observing the data given another hypothesised mean. To get the posterior probability for any hypothesis, say $\mu = 118$, we thus calculate:

$$\begin{aligned}\mathbb{P}[H = 118 \mid E = (110, 115, 120)] =& \mathbb{P}[E = (110, 115, 120) | \ H = 118] = \\ & \mathbb{P}[E = 110 \mid H = 118] \times \\ & \mathbb{P}[E = 115 \mid H = 118] \times \\ & \mathbb{P}[E = 120 \mid H = 118]\end{aligned}$$

That is, calculating the joint probability of observing $x_1,\ x_2$ and x_3 given that our observations instead would have come from a distribution with a mean of 118. This is then done for all possible hypotheses, i.e. values of μ, creating a probability distribution of the parameter. The peak of this distribution would in fact be the maximum likelihood estimate of the mean, which is used in more classical estimation methods. Hence, using a non-informative prior often yields estimations with frequentist properties[2]. Figure 4 visualise 3 different hypotheses in relation to the observations.

[2] A Bayesian 95% credible interval can for example be said to have good frequentist properties if it would cover the parameter 95% of the times it is created (as is the case for the frequentist 95% confidence interval)

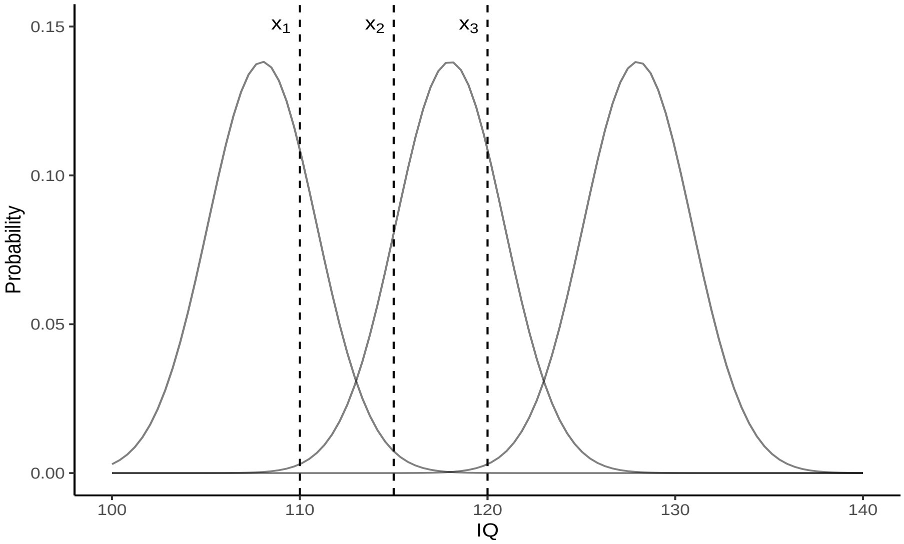

Figure 4: Each curve represents a hypothesis, given which the likelihood of observing the data points are calculated. It would be more likely to observe the values $x_1 = 110,\ x_2 = 115,\ x_3 = 120$ if the middle distribution was true than for either of the other two.

However, if we for example have on good authority that mean IQ in the sampled population is 125 with an SD of 5, we specify our prior as such. This is then weighed in when calculating the posterior. That is, we calculate the joint probability that we have observed our data given a hypothesis (μ) and the probability of observing that μ in the prior distribution, as shown in figure 5. This is done for all possible hypotheses, just as previously shown. Often one wants to use non-informative priors as to not arbitrarily bias the results. For an accessible and more thorough explanation of Bayesian statistics, see e.g. Donovan & Mickey (2019).

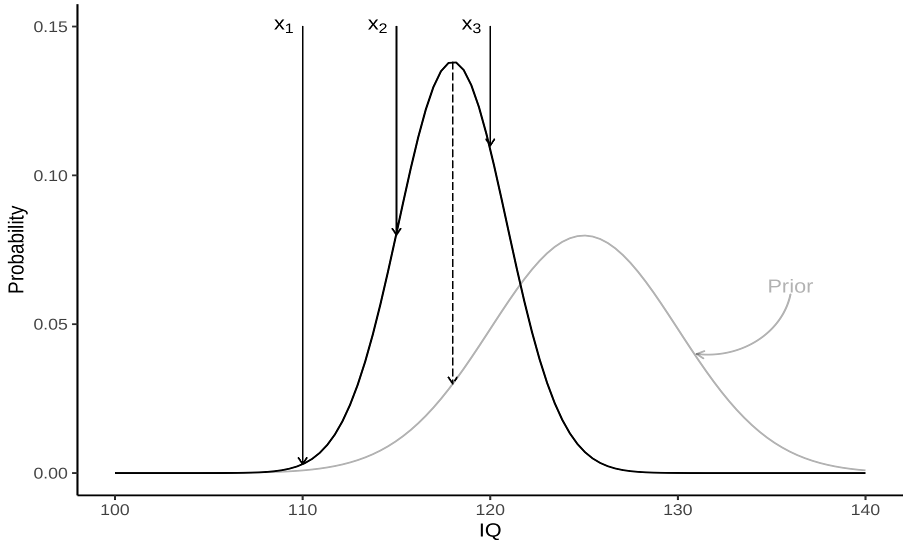

Figure 5: For the hypothesis $\mu = 118$ with a specified prior with $\mu = 125$ and $\sigma = 5$, the probability of this value of μ is found by calculating the joint probability of each arrow. The dashed arrow represents the likelihood of μ being 118 if the prior distribution was true.

The example given here is a special case of Bayesian parameter estimation that can be solved analytically.[3] However, this is in many cases not possible. Instead one utilise Markov Chain Monte Carlo (MCMC) methods. A Markov Chain describe a series of events where the probability for each of them solely depends on the one preceding it. Monte Carlo methods are mathematical algorithms that solves problems by random generation of numbers. The idea behind this in Bayesian statistics is that one estimates the posterior by drawing a large amount of random samples from it. Thus "building" it from scratch. Because the denominator in (8) is a constant and we can ignore it, this becomes possible. Bayes theorem can then be rewritten as:

$$posterior \ \propto \ likelihood \times prior \tag{9}$$

What this is saying is that the posterior density of a hypothesis is proportional ($\propto$) to the likelihood of the data under that hypothesis times the prior density of

[3] The difference between analytic and numerical problem solving is that analytic solutions are exact as well as derived and presented in (for the mathematician) understandable forms. A numerical solution involves "guesswork" and is stopped when a solution is found that satisfies the problem. This method only approximates the "true" solution. Monte Carlo simulations belong to the latter category.

the hypothesis. The methods for drawing these samples differ depending on the type of distribution and problem at hand. But in general they are all building on algorithmic rules ("recipes") of drawing random numbers based on the likelihood and the prior, saving them and after a large number of iterations observing the distribution they form. The average of which often is the parameter of interest.

1.2.2.1. The Bayesian test of deficit. Assume a sample of n controls on which we measure some value x that is normally distributed with mean μ and variance σ^2. Let $\overline{x}$ and s^2 denote the sample mean and sample variance respectively. The case is denoted x^*. The prior used is non-informative, see Crawford & Garthwaite (2007) and DeGroot & Schervish (2012) (p. 495) for the formal specification of the prior. The algorithm developed in Crawford & Garthwaite (2007) for obtaining a point estimate of a case's abnormality p, i.e. a p-value or the proportion of controls that would fall below the case and accompanying intervals is as follows:

1. Let ψ be a random draw from a χ^2-distribution on $n-1$ *df*. Then let $\hat{\sigma}^2 = \frac{(n-1)s^2}{\psi}$ be the estimation of σ^2 for this iteration.
2. Let z be a random draw from a standard normal distribution. Then let $\hat{\mu} = \overline{x} + z\sqrt{(\hat{\sigma}^2/n)}$ be the estimate of μ for this iteration.
3. With estimates of μ and σ, p is calculated conditional on these estimates being the correct μ and σ by calculating $z^* = \frac{x^* - \hat{\mu}}{\sqrt{\hat{\sigma}^2}}$. Let $\hat{p}_i = \mathbb{P}[Z < z^*]$ be the estimate of p for this iteration. That is the probability of drawing a value less than z^* from a standard normal distribution.
4. Repeating these steps a large number of times will yield a distribution of $\hat{p}$, the average of which is the point estimate of p. If repeated e.g. 1000 times, the 25th smallest and 25th largest $\hat{p}_i$ (i for iteration) is the lower and upper boundaries of the 95% Bayesian credible interval for p.

Crawford & Garthwaite (2007) show that this method yields converging results to that of TD (3). As noted, this is often the case when using a non-informative prior, however, not always. For example, RSDT and its Bayesian analogue (BSDT) do not produce identical results. Crawford & Garthwaite (2007) showed that when there was no discrepancy between task A and B, but the case was severely impaired on both of them, RSDT exhibited an error rate much larger than the specified α-level. For example, when the case had a deficit of 8 SD on both task A and B but no discrepancy, RSDT gave a false positive (Type I error) in 34.72% of the simulations ($n = 10$). The BSDT on the other hand gave a false positive under the same circumstances in 8.29% of the simulations. When there was no deficit on either task BSDT had a Type I error rate of 7.32%. compared to RSDT with an error rate of 4.6%. It is also argued that considering the often large effects of brain damage, this is an acceptable tradeoff.

1.2.2.2. The Bayesian standardised difference test. Assume a sample of n controls on which we measure some value x and y from task A and B. Let $\overline{x}$ and $\overline{y}$ denote the sample means and

$$\boldsymbol{A} = \begin{pmatrix} s_{xx} & s_{xy} \\ s_{xy} & s_{yy} \end{pmatrix}$$

the sample covariance matrix (note that s_{xx} is not the same as s_x and denotes variance instead of standard deviation). It is assumed that the observations come from a bivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, note that the parameters are bolded as to represent the vector and matrix:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}$$

Let the case scores be denoted x^* and y^*. Just as for the frequentist dissociation tests we want to estimate the proportion p of the control population that would exhibit a greater difference $x-y$ than the case's x^*-y^*. A non-informative prior was again specified, see Crawford & Garthwaite (2007) and Jeffreys (1998). The algorithm for obtaining $\hat{p}_i$, the ith estimation of p in Crawford & Garthwaite (2007) follows below:

1. Let

   $$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\sigma}_{xx} & \hat{\sigma}_{xy} \\ \hat{\sigma}_{xy} & \hat{\sigma}_{yy} \end{pmatrix}$$

   be a random draw from an inverse-Wishart distribution (a multivariate generalisation of the χ^2-distribution) on n degrees of freedom with scale matrix $\boldsymbol{A}$. And let $\hat{\boldsymbol{\Sigma}}$ be the estimate of $\boldsymbol{\Sigma}$ for this iteration.
2. Let z_1 and z_2 be two random draws from a standard normal distribution. Perform Cholesky decomposition on $\hat{\boldsymbol{\Sigma}}$, that is finding the lower triangular matrix $\boldsymbol{T}$ such that $\boldsymbol{TT'} = \hat{\boldsymbol{\Sigma}}$. Then

   $$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_x \\ \hat{\mu}_y \end{pmatrix} = \begin{pmatrix} \overline{x} \\ \overline{y} \end{pmatrix} + \boldsymbol{T} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \cdot \frac{1}{n}$$

   is the estimation of $\boldsymbol{\mu}$ for this iteration.

3. With estimations of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ we can calculate p, given that they are the the correct values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. If an unstandardised test is desirable put:

   $$z^* = \frac{(x^* - \hat{\mu}_x) - (y^* - \hat{\mu}_y)}{\sqrt{\hat{\sigma}_{xx} + \hat{\sigma}_{yy} - 2\hat{\sigma}_{xy}}}$$

   If a standardised test is required, put:

   $$z_x = \frac{(x^* - \hat{\mu}_x)}{\sqrt{\hat{\sigma}_{xx}}},\ z_y = \frac{(x^* - \hat{\mu}_y)}{\sqrt{\hat{\sigma}_{yy}}} \text{ and } \hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\sqrt{\hat{\sigma}_{xx}\hat{\sigma}_{yy}}}$$

   and

   $$z^* = \frac{z_x - z_y}{\sqrt{2 - 2\hat{\rho}_{xy}}}$$

4. Let $\hat{p}_i$ be the tail area of a standard normal distribution less or greater than z^* (depending on alternative hypothesis). $\hat{p}_i$ is then the estimate of p for this iteration. Repeating these steps a large number of times will yield a distribution of $\hat{p}$, the average of which is the point estimate of p. If repeated e.g. 1000 times, the 25th smallest and 25th largest $\hat{p}_i$ is the lower and upper boundaries of the 95% Bayesian credible interval for p.

1.2.2.3. Bayesian tests allowing for covariates. It should now be clear that using a matched control to account for premorbid cognitive ability of the case is important. This has often led to control samples being small. In an attempt to remedy this Crawford et al. (2011) followed up the previously described tests by developing methods that allow for the case to be assessed on abnormality in the presence of covariates. That is, the tests let you compare the case's score on the task of interest conditioned upon the results of the controls having the same score on the covariate(s). If a patient has 15 years of education, his/her score on the task would be compared to the controls with equal length of education. This reduces the need to perfectly match control samples and is thus a major contribution to the field.

The procedural details of these tests are for the present work out of scope, but one thing is important to mention. In Crawford et al. (2011) they change the recommendation of the prior to use, when testing for a dissociation. In Berger & Sun (2008) it is shown that the prior used in Crawford & Garthwaite (2007) has good frequentist properties for estimating σ_x and σ_y but less so for ρ or the discrepancy size. Instead they recommend a calibrated prior. The main difference being that an accept/reject algorithm is applied after the initial draw from the inverse-Wishart distribution. Many Bayesians would argue that good frequentist properties are not necessary when conducting a Bayesian analysis. If so, one can use the "standard theory" Jeffreys (1998)'s prior. In the R package `singcar` (section 3) both types of priors are implemented for both tests. If one wants to compare the results from them, the same prior should be used.

2. The issue of power

Statistical power is the probability of a test to reject the null hypothesis given that the null hypothesis is false. In other words, finding an effect if the effect in fact exists in the population. The power of a test is essentially determined by three parameters: the sample size n, since the standard error of the estimated parameter $\frac{\sigma}{\sqrt{n}} \to 0$ when $n \to \infty$ thereby making it more sensitive to detect smaller effects; the true existing effect size θ, since greater effects are more easily detected; and the chosen significance level α, since adjusting the threshold we set for deciding if an effect has been found makes it more or less probable that the observed data surpass it.

Why does power matter? Do we want to increase it solely to raise our chances of finding an effect and being published? Well, unfortunately the problems of low power extends beyond single studies. Primarily, it gives rise to three issues (Button et al., 2013; Ioannidis, 2005).

1. Low probability of detecting an effect. This is given by the definition of power above. If β is defined as the probability of missing an existing effect (i.e. a Type II error) then power is by definition $1 - \beta$. The risk of committing a Type II error therefore increases as power decreases.
2. As power decreases so does the positive predictive value (PPV), which is especially worrisome if using these tests for clinical diagnoses. PPV is the probability that a patient has a disorder given a positive test result of that disorder (Labarge et al., 2003). The intuition behind this is that given, say 20% decrease in power, 20% more true diagnoses would be missed. But the extent of this 20% depends on the prevalence of the diagnose and the Type I error rate. To understand this better observe table 1.

Table 1: Relations between α, β and power.

		Null hypothesis	
		True	False
Decision	Don't reject	Correct ($\mathbb{P} = 1 - \alpha$)	Type II error ($\mathbb{P} = \beta$)
	Reject	Type I error ($\mathbb{P} = \alpha$)	Correct ($\mathbb{P} = 1 - \beta = \mathit{power}$)

Say that the base rate, that is the prevalence of a diagnosis in a certain population, is 10% and that we have a population of 1000 individuals. We test them all with a diagnostic test with a power of 0.8 (hence $\beta = 0.2$) and a Type I error rate $\alpha = 0.05$. The null-hypothesis in this scenario being that a person does not have a deficit (or disease). Add this to table 1.

Table 2: Example of the relations between α, β and power.

		Null hypothesis		
		True	False	*Total*
Decision	Don't reject	$900 \cdot 0.95 = 855$	$100 \cdot 0.20 = 20$	875
	Reject	$900 \cdot 0.05 = 45$	$100 \cdot 0.80 = 80$	125
	Total	900	100	1000

The PPV is calculated by taking the number of correctly diagnosed patients (bottom right in table 1) and dividing it by the total number of diagnosed patients, i.e. in this case $PPV = \frac{80}{125} = 0.64$. From table 2 one can derive a more easy to use formula (this is in fact based on Bayes' theorem). Note that BR = base rate

$$PPV = \frac{\mathit{power} \cdot BR}{(\mathit{power} \cdot BR) + \alpha(1 - BR)} \tag{10}$$

Consider the same conditions as previous but now the power is 0.21, which was the median statistical power over several subfields in neuroscience found by Button et al. (2013). We would then have $PPV = \frac{0.21 \cdot 0.1}{(0.21 \cdot 0.1) + 0.05(1 - 0.1)} = 0.32$. This drop in power, given a base rate of 10%, would thus cut the probability that a significant test reflects a true effect (diagnose) by half. This is not only a problem in diagnostics but affect basic science as well, where the base rate represents the number of significant studies found in a field divided by the total number of studies conducted. In (Crawford, Garthwaite, & Betkowska, 2009), Crawford and colleagues provides a method for calculating interval estimates for these "post test" probabilities.

3. If a test detects a true effect despite low power and PPV, the estimate of the effect size is often overestimated. This is because of a phenomenon called "the winner's curse" and stems from the inescapable randomness of sampling. Suppose an effect is being investigated in several very low powered studies and suppose that the true underlying effect happens to be small. Due to the randomness of sampling we expect the estimation of the effect to vary from one study to another and if none of the studies are powered enough to detect a small effect, only the studies that happen to draw a sample in which the effect is a little larger would be able to obtain significance. Since scientific journals are biased towards publishing significant results (Duval & Tweedie, 2000), the distribution of effects seen in the literature will be biased and the sizes inflated.

Power in single case neuropsychology has not been thoroughly examined. Most attention has been given to Type I errors (e.g. Crawford & Garthwaite (2005) and Crawford & Garthwaite (2012)). However, in both Crawford & Garthwaite (2006b) and Crawford & Garthwaite (2006a) some power analyses are conducted. In Crawford & Garthwaite (2006b) they assess TD, in relation to another deficits test by Mycroft et al. (2002). This was conducted by Monte Carlo simulations. The simplicity of TD, however, makes it suitable to analytically derive its power function. Crawford & Garthwaite (2006a) assess power mainly between the various criteria for dissociations. But as mentioned, following McIntosh (2018), these criteria seems superfluous and even hindering. Therefore, modelling power for detecting sole task discrepancy would complement this investigation. And since UDT, RSDT and BSDT do not produce equal output, a comparison between them would be useful.

Furthermore, meta-analyses and summarising reviews are becoming ever more necessary for exposing robust findings in basic science. However, to appropriately estimate effects across studies one must take potential biases into account. The extent to which the winner's curse is affecting single case neuropsychology has not yet been quantified. Doing so and developing a correction method would be useful for appropriate post hoc power analyses and meta-studies.

3. R package 'singcar'

All of the tests devised by Crawford and colleagues are accompanied by freely available computer programs, found here[4]. It is a great resource, providing access to the tests which in some cases are cumbersome to calculate by hand. However, these programs are also limited — for example by most often only allowing summary data as input. Further, these programs are fairly rigid and processing larger data-sets or a large number of data-sets cannot be automated. Hence, it was deemed necessary for the present work to implement the tests in a programming language suitable for both analyses and modelling. The tests were implemented in the widely used and accessible statistical software/programming language R (R Core Team, 2020), by developing a package combining Crawford and colleagues' separate programs into one. Even the more advanced Bayesian methods allowing for coviariates, not investigated in the present work, are implemented. Making the tests more accessible should be beneficial for the wider community and hopefully increase usage. The package also comes with power calculators for the majority of the tests. An introduction to the functions and usages of the package `singcar` (single case R) is given in appendix A. The developmental version of the package can be found at https://github.com/AljosjaK/singcar.

4. Power modelling

4.1. Tests of deficit

4.1.1. Frequentist test of deficit

Crawford & Garthwaite (2006b) modeled power for TD using simulations. Reasons for using simulations rather an analytic approach were essentially two: it is easier to model complex problems (such as assessing effects of non-normality) and simulation based approaches can be easier to understand. When using simulations it is very time consuming to to model every possible parameter combination, whereas analytic solutions gives instant results. Therefore an analytic approach to calculating the power of TD gives a more complete picture than the power study conducted by Crawford & Garthwaite (2006b).

Calculating power for a z-test (i.e. using the normal distribution) is fairly easy. The power function for a hypothesis test is defined as follows:

$$power = \mathbb{P}[\text{reject } H_0 \mid H_a \text{ is true}]$$

[4] https://homepages.abdn.ac.uk/j.crawford/pages/dept/psychom.htm

Where $\mathbb{P}$ is probability, $|$ read "given", H_0 the null and H_a the alternative hypotheses. Say that we want to test if a sample of $n = 16$ from a certain population has higher IQ than the mean (μ) of 100 and we know that the SD (σ) is 15. This gives us the hypotheses:

$$H_0 : \ \mu = 100 \ \text{ vs. } \ H_a : \ \mu > 100$$

That is, we have a one-sided one sample z-test (since σ is known). The critical value that the sample mean ($\overline{x}$) has to surpass with an $\alpha = 0.05$ is $\overline{x}_{crit} = \mu + z_{crit} * \frac{\sigma}{\sqrt{n}}$ i.e. $100 + 1.64 * \frac{15}{\sqrt{16}} = 106.15$. If H_a is true and the mean of our population of interest in fact is say $\mu = 105$ then

$$\begin{aligned} power &= \mathbb{P}[\overline{x} > 106.15] \text{ when } \mu = 105 \\ &= \mathbb{P}\left[\frac{\overline{x} - \mu}{\sigma/\sqrt{n}} > \frac{106.15 - 105}{15/\sqrt{16}}\right] \end{aligned}$$

The leftmost quantity is a standard normal random variable, hence

$$power = \mathbb{P}[Z > 0.307]$$

The size of this area under the normal curve as seen (the shaded area in figure 6) is thus the power of the test. Given a true mean of 105, a sample of 16 and $\alpha = 0.05$ we have a power of 37.9%.

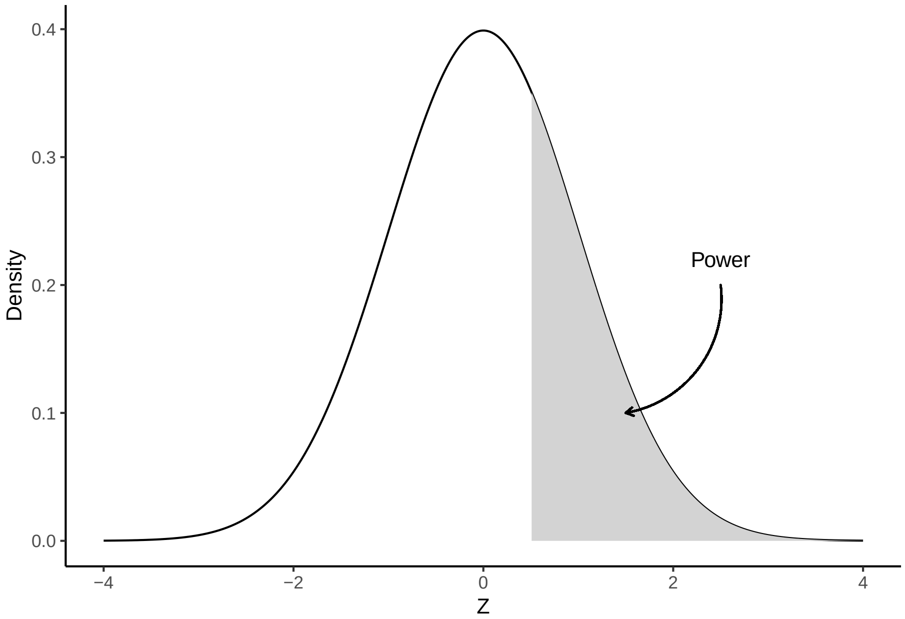

Figure 6: Power of a one-side, one-sample z-test for IQ given a true mean of 105, a sample of 16 and $\alpha = 0.05$.

For large samples one can approximate power calculations for t-tests with the normal distribution as shown above. The normative control sample in single case studies are typically small (Crawford et al., 2011) and thus the calculations are somewhat trickier.

Power calculations for small samples require the use of the non-central t-distribution. The t-distribution is only symmetric when the mean is 0. Given that the alternative hypothesis is true, the distribution representing our population of interest would not have a mean of 0 and would thus be skewed (Harrison & Brady, 2004), figure 7 illustrates this and below follows the derivation of the power function for the test of deficit.

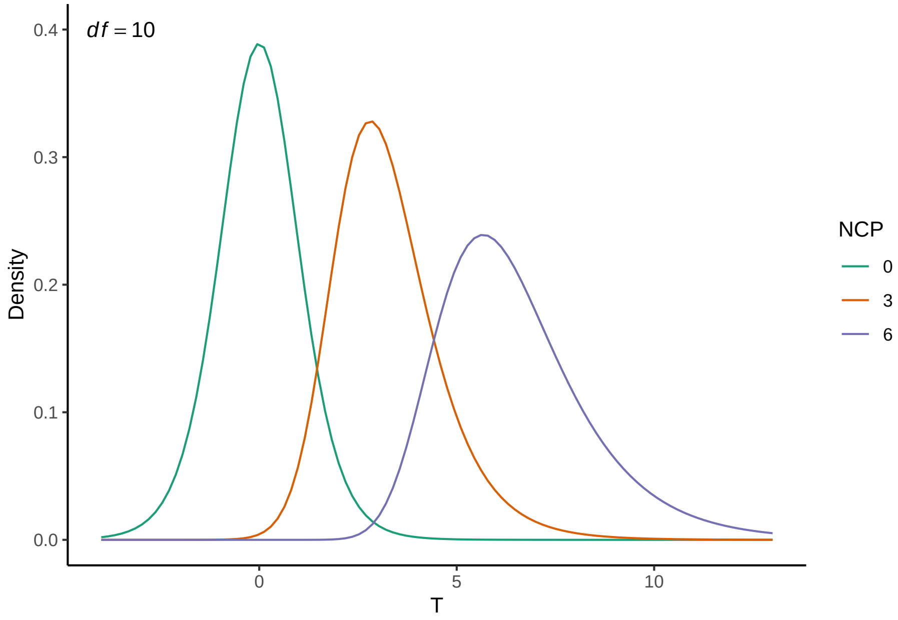

Figure 7: Distribution of T, given three different non-centrality parameters (NCP) on 9 degrees of freedom.

Assume we have a sample $x_i,\ i = 1,\ 2,\ 3\ldots, n$ and the hypotheses:

$$H_0:\ \mu = \mu_0 \ \text{ vs. }\ H_a:\ \mu < \mu_0$$

We test this hypothesis with a one sample t-test:

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Where $\overline{x}$ is the sample mean and s the sample standard deviation. The test statistic T has a central t-distribution on $n-1$ degrees of freedom *given* that the null hypothesis $H_0:\ \mu = \mu_0$ is true. However, when the alternative hypothesis $H_a:\ \mu < \mu_0$ is true T has a *non-central* t-distribution on $n-1$ degrees of freedom with the non-centrality parameter

$$\theta = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

See Harrison & Brady (2004) for more details on different variants of t-tests and their respective non-central distributions. Under the alternative hypothesis,

$H_a: \ \mu < \mu_0$, the power for the test above (one-sided) is given by

$$power = 1 - \beta = T_{n-1}\left(t_{\alpha,\ n-1}\middle|\frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)$$

Where $T_{n-1}(.|\theta)$ is the cumulative distribution function for the non-central t-distribution with $n-1$ degrees of freedom and non-centrality parameter $\theta = \frac{\mu-\mu_0}{\sigma/\sqrt{n}}$. $t_{\alpha,\ n-1}$ is the α quantile of the *central* t-distribution on $n-1$ degrees of freedom.

To adjust this to Crawford & Howell (1998)'s test of deficit one simply put the non-centrality parameter $\theta = \frac{x^*-\overline{x}}{\sigma\sqrt{\frac{n+1}{n}}}$ from equation (3), yielding

$$power = 1 - \beta = T_{n-1}\left(t_{\alpha,\ n-1}\middle|\frac{x^* - \overline{x}}{\sigma\sqrt{\frac{n+1}{n}}}\right)$$

This is most easily calculated using software since the t-distribution is quite messy to integrate by hand. Figure 8 shows the values obtained using the t-distribution functions in R (R Core Team, 2020), where $\overline{x} = 0$ and $\sigma = 1$ as to be able to interpret the deficits in standard deviations.

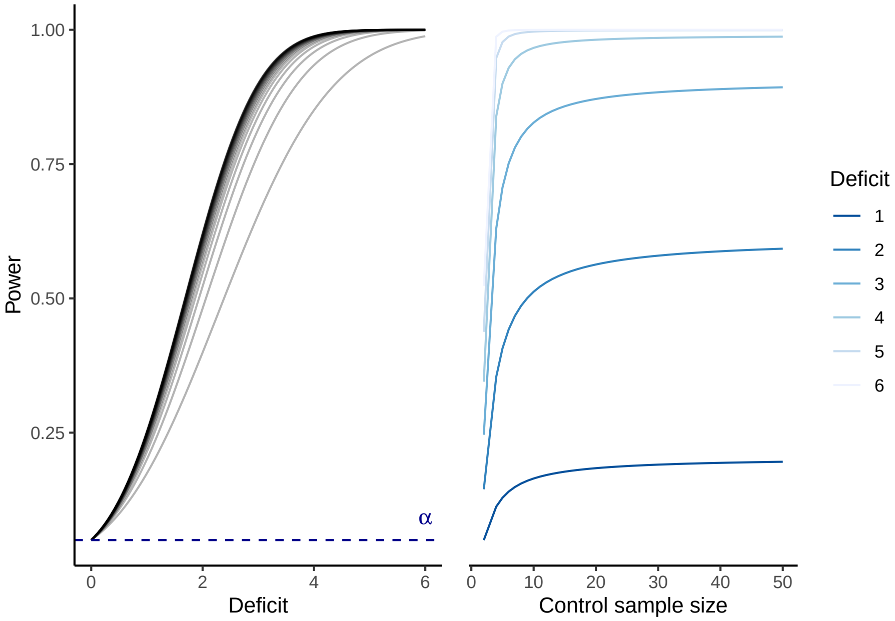

Figure 8: Analytic power curves for TD, n goes from 4 to 50 in steps of 2 (left), each line representing power for a certain sample size. Deficit interpreted in standard deviations.

The most apparent thing to note from figure 8 (right) is the steep increase in power when the control sample size goes from 4 to ~ 16 and then the striking flatness of the curve as it increases more. This means that to increase power for a single case deficit study using TD, a researcher should strive to include at least 16 control participants but any further inclusion would only yield marginal effects on power. For reference, these curves can be compared to the results from the Monte Carlo simulations in Crawford & Garthwaite (2006b). Table 3 illustrates the comparison, where the small differences seen are to be expected from a Monte Carlo approach.

Table 3: Comparisons of power calculations for TD and BTD

		Analytic	Simulation	
Deficit	n	TD	TD	BTD
2	5	44.82	44.83	44.76
	10	54.64	54.63	54.27
	20	59.36	59.46	59.74
	50	62.10	62.06	61.91
	100	62.99	63.00	63.69
3	5	72.67	72.80	72.96
	10	83.92	83.87	83.69
	20	88.02	87.88	88.56
	50	90.05	90.11	90.24
	100	90.65	90.73	90.51

4.1.2. Bayesian test of deficit

Crawford & Garthwaite (2007) note that the point estimates of abnormality for both the Bayesian and frequentist tests of deficit are almost indistinguishable. A small simulation study was conducted to compare the power of the tests and as table 3 shows, they differ only to the extent that would be expected from Monte Carlo variation.

4.2. Tests of dissociation

Due to the unlikely scenario of investigating two negatively correlated tasks, power was modeled for positively correlated variables only.

4.2.1. Unstandardised difference test

For the unstandardised difference test (UDT), where the distribution of the test statistic is exactly t, a similar method as that for TD was used to derive the power function. Since dissociaton tests most often should be used two-sided, the exact power is given by (11) (for one-sided tests the smaller term is ignored).

$$power = 1 - \beta = T_{n-1}\left(t_{\alpha/2,\ n-1} |\ \theta\right) - T_{n-1}\left(-t_{\alpha/2,\ n-1} |\ \theta\right) \tag{11}$$

Where the non-centrality parameter θ is given by equation (5), i.e.

$$\theta = \frac{(x_a^* - \overline{x}_a) - (x_b^* - \overline{x}_b)}{\sqrt{(s_a^2 + s_b^2 - 2s_a s_b r_{ab})(\frac{n+1}{n})}}$$

Following the same logic as for the test of deficit, in figure 9 and 10 power is presented as a function of the task discrepancy and as a function of control sample size respectively. Just as for the test of deficit, UDT shows a steep increase in power for increasing sample sizes up to a certain point, after which any power increase is marginal. This point differs depending on the strength of the correlation.

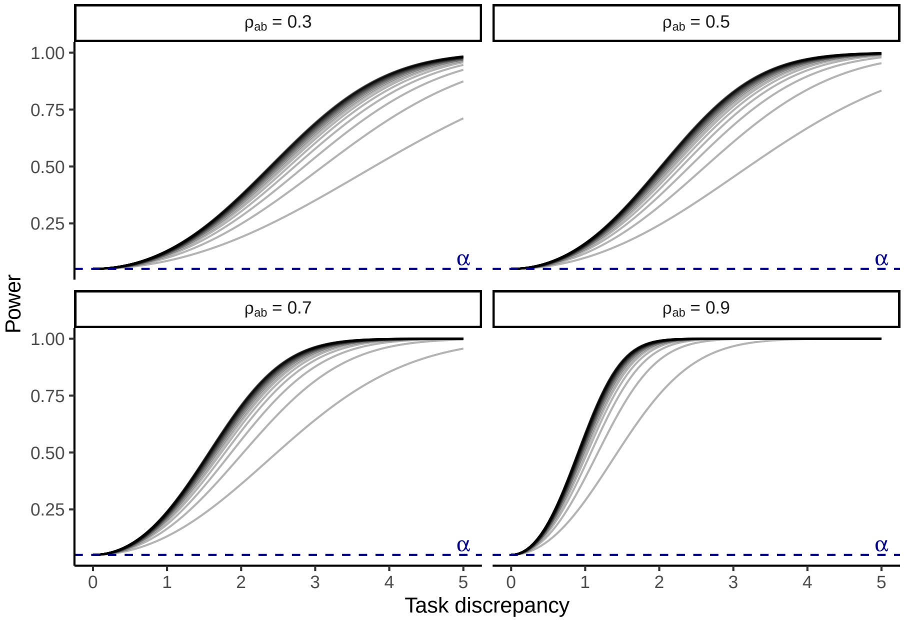

Figure 9: Analytic power curves for UDT as a function of discrepancy size, n goes from 4 to 50 in steps of 2, each line representing power for a certain sample size. Deficit interpreted in standard deviations. ρ_{ab} = correlation between the tasks.

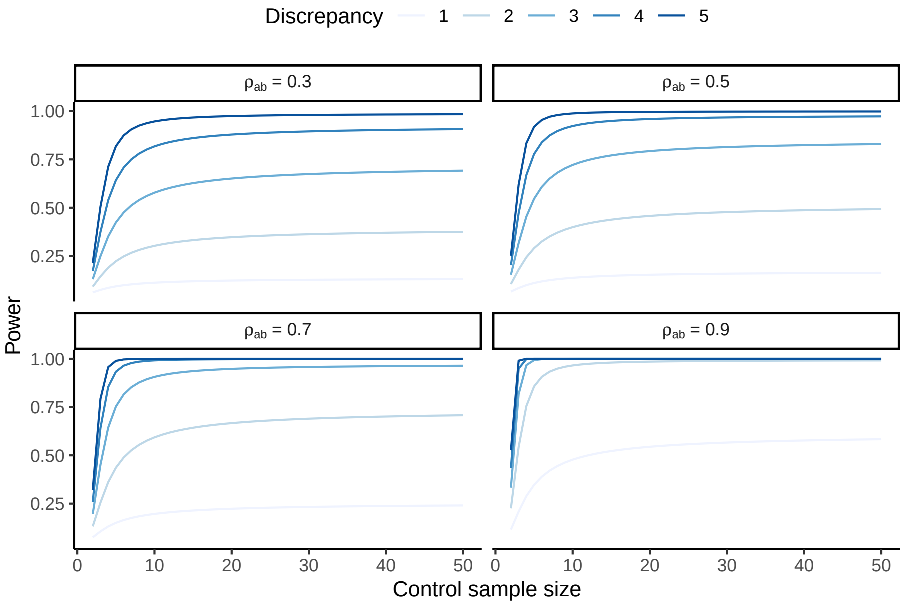

Figure 10: Analytic power curves for UDT as a function of sample size, n goes from 4 to 50 in steps of 2. Discrepancy interpreted in standard deviations. ρ_{ab} = correlation between the tasks.

To verify the appropriateness of this method Monte Carlo simulations were run for each parameter combination presented in table 4, where it can be seen that the two solutions yield almost identical results.

Table 4: Comparison between analytic and Monte Carlo methods for deriving power for UDT.

		Analytic (ρ_{AB})				Simulation (ρ_{AB})			
Disc	n	0.3	0.5	0.7	0.9	0.3	0.5	0.7	0.9
2	5	22.28	29.03	43.60	85.66	22.32	29.01	43.68	85.89
	10	30.25	39.92	59.31	96.54	30.17	39.96	59.14	96.49
	20	34.72	45.75	66.68	98.52	35.03	45.75	67.05	98.51
	50	37.50	49.27	70.75	99.13	37.47	48.93	70.88	99.11
	100	38.44	50.43	72.05	99.28	38.56	50.66	72.02	99.24
3	5	42.38	54.61	75.28	99.28	42.38	54.53	75.28	99.25
	10	57.78	72.14	90.63	99.99	57.84	72.37	90.63	99.98
	20	65.10	79.31	94.77	100.00	65.12	79.22	94.74	100.00
	50	69.18	82.94	96.40	100.00	69.36	83.13	96.39	100.00
	100	70.48	84.04	96.83	100.00	70.53	83.97	96.80	100.00

Note: Disc = task discrepancy. Number of simulations 100 000.

4.2.2. Revised standardised difference test

Given the fact that the test statistic for the RSDT only approximates the t-distribution (Crawford & Garthwaite, 2005; Garthwaite & Crawford, 2004), an analytic approach to deriving the power curve is not appropriate. To derive power through numerical methods, such as Monte Carlo, one simulate testing conditions. Say that you for example want to examine power for TD if a case has a deficit of 2 and is compared to a control sample of 16. Then one would generate 17 random draws from a standard normal distribution. On one of these draws, the case, a deficit would be imposed by subtracting the desired size, i.e. 2. Since numbers are generated from a standard normal distribution, subtracting 2 would mean imposing a deficit of 2 SD. Then the case is compared to the rest of the 16 random draws, i.e. the controls, with the test of interest and the p-value saved. If this is repeated a large number of times power, for these conditions, can be calculated by taking the number of significant tests and divide it by the total number of tests run.

For a dissociation test the random draws are instead generated from a bivariate normal distribution, that is two variables with $\mu = 0$ and $\sigma = 1$ and some specified correlation between them. In mathematical terms, you draw $n+1$ random draws from the distribution:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ where } \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho_{AB} \\ \rho_{AB} & 1 \end{pmatrix} \tag{12}$$

That is, two scores representing task A and B are drawn for each "person". A discrepancy for the case can then be imposed by subtracting the desired size on one of the variables for the $n+1$th draw, i.e. imposing a deficit on task A while leaving task B "unscathed".

For RSDT, Monte Carlo simulations were run with $n = 5,\ 10,\ 20,\ 50,\ 100$, $\rho_{AB} = 0.3,\ 0.5,\ 0.7,\ 0.9$ and task discrepancy ranging from 0 to 5 in steps of 0.2. As well as with discrepancy $= 1,\ 2,\ 3,\ 4,\ 5$, $\rho_{AB} = 0.3,\ 0.5,\ 0.7,\ 0.9$ and sample size ranging from 4 to 50 in steps of 2. Each parameter combination was run 100 000 times, that is 100 000 p-values were saved and the proportion of significant values calculated. Number generation was done using R package MASS (Venables & Ripley, 2002). In figure 11 and 12 the power curves are presented. As can be seen, the pattern is similar to that of both TD and UDT: when sample sizes increase to a certain point, further increase yields little advantage.

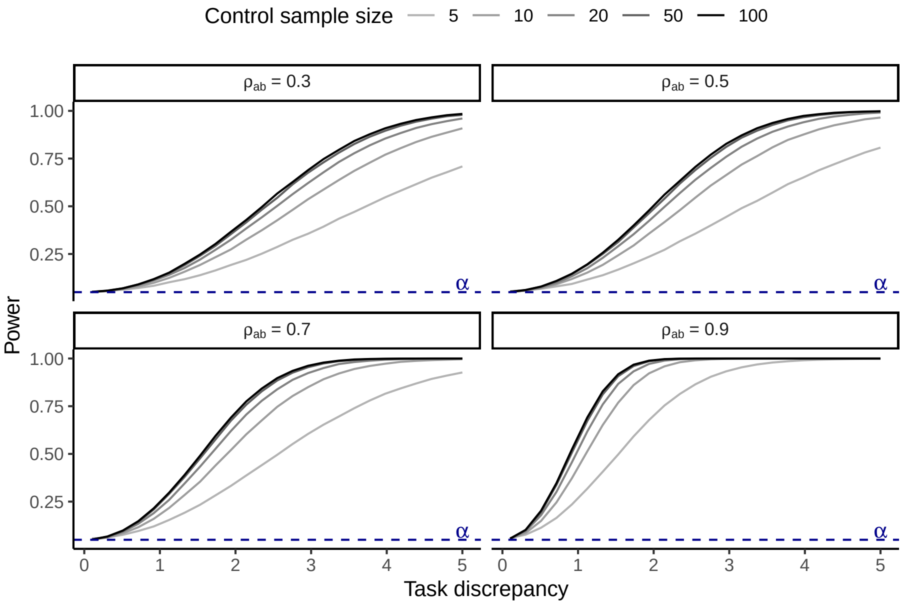

Figure 11: Empirical power curves for RSDT as a function of task discrepancy. Discrepancy interpreted in standard deviations and range in steps of 0.2. ρ_{ab} = correlation between the tasks. Number of simulations = 100 000.

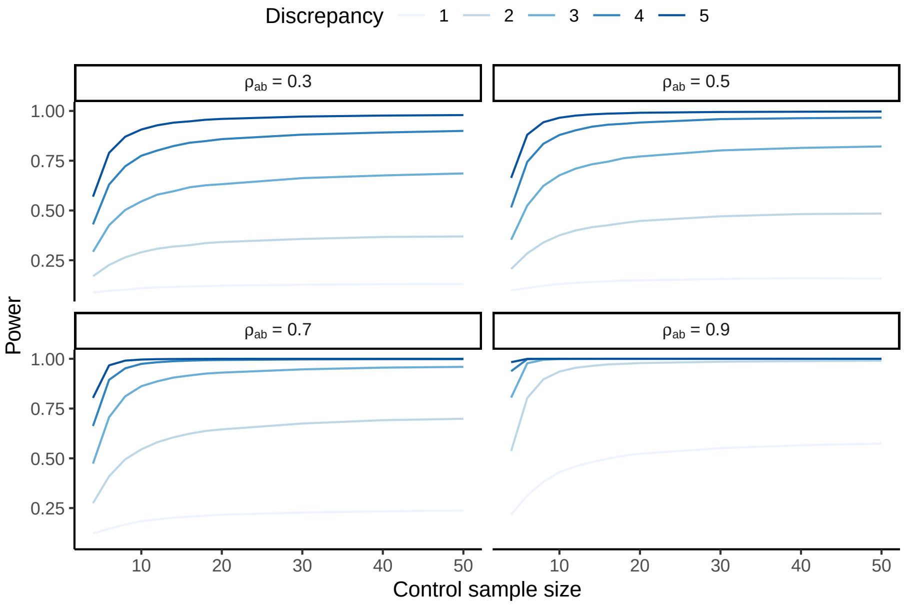

Figure 12: Empirical power curves for RSDT as a function of sample size, n goes from 4 to 50 in steps of 2. Discrepancy interpreted in standard deviations. ρ_{ab} = correlation between the tasks. Number of simulations = 100 000.

4.2.3. Bayesian standardised difference test

For BSDT there is no feasible analytic method. So Monte Carlo simulations for parameter combinations mimicking the ones used for RSDT were run, using BSDT to obtain the p-values. Each parameter combination was run 10 000 due to the heavier computational load of the Bayesian test. In figure 13 and 14 the power curves are presented as a function of discrepancy and sample size respectively. The curves are similar to those of UDT and RSDT. However, when sample sizes are small BSDT lose control of Type I errors. In figure 13 this is seen when the sample of 5 exhibits higher "power" than the specified $\alpha = 0.05$ when the discrepancy is 0. In figure 14 we see this effect as power decreasing by an increase in sample size for a discrepancy of 1 SD.

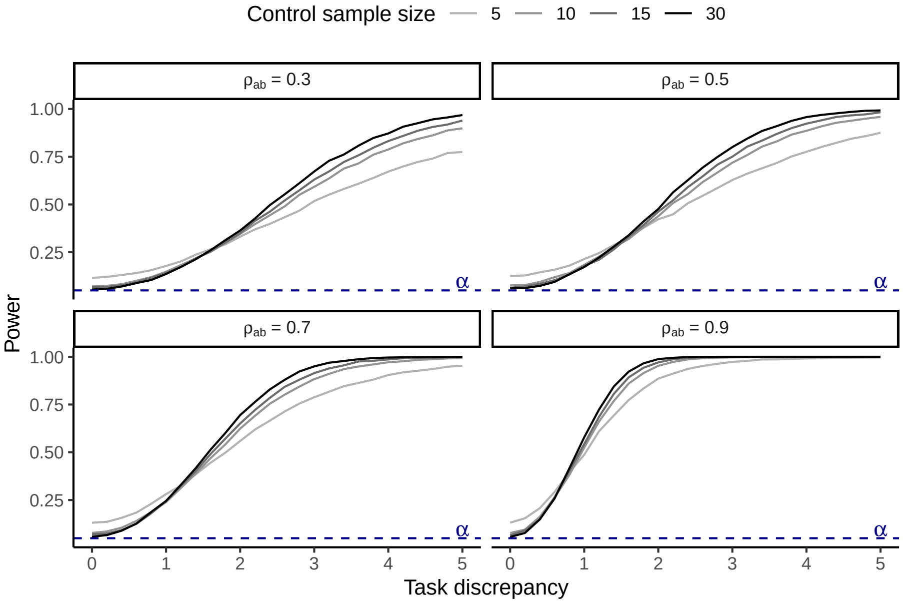

Figure 13: Empirical power curves for BSDT as a function of task discrepancy. Discrepancy interpreted in standard deviations and range in steps of 0.5. ρ_{ab} = correlation between the tasks. Number of simulations = 100 000.

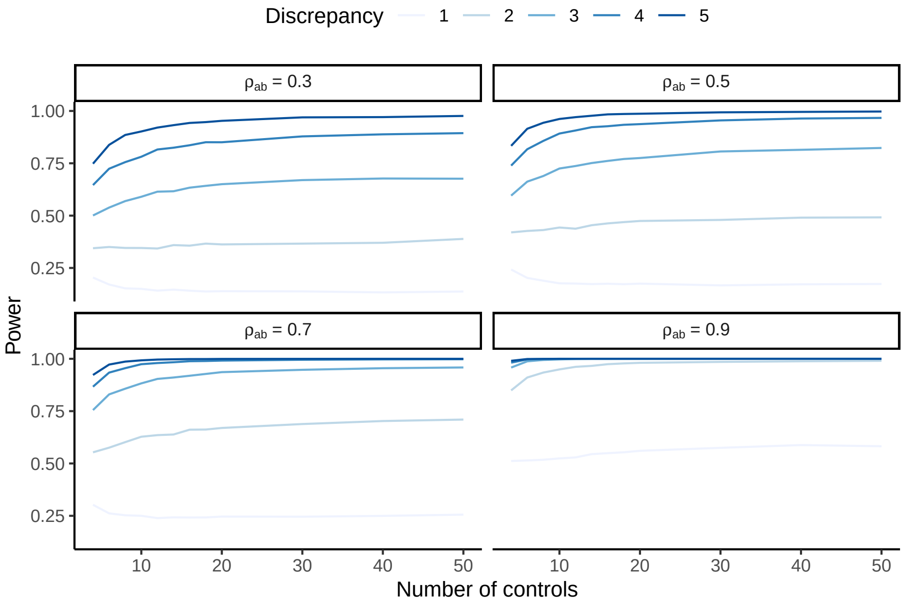

Figure 14: Empirical power curves for BSDT as a function of sample size, n goes from 4 to 50 in steps of 2. Discrepancy interpreted in standard deviations. ρ_{ab} = correlation between the tasks. Number of simulations = 10 000.

4.3. Comparisons between the tests

For comparison between the tests, simulations for the RSDT and BSDT for a sample size of 16, a discrepancy of 2 SD and ρ ranging from 0 to 0.9 in steps of 0.1 were run. Further, simulations for RSDT and BSDT for a deficit on task A ranging from 7 to 4 SD below the mean in steps of 0.2 in addition to a deficit of 4 SD on task B, for a sample of 16 and $\rho_{AB} = 0.3,\ 0.5,\ 0.7,\ 0.9$ were run. The parameter combinations were run 10 000 times each and power calculated as the number of significant tests divided by the total number of tests run. The curves for the UDT were derived analytically. Figure 15 shows that BSDT has a small but relatively consistent power advantage over the other two but that the difference also diminishes between BSDT and UDT as the correlation between task A and task B increase. This must, however, be seen in the light of BSDT failing to control Type I errors.

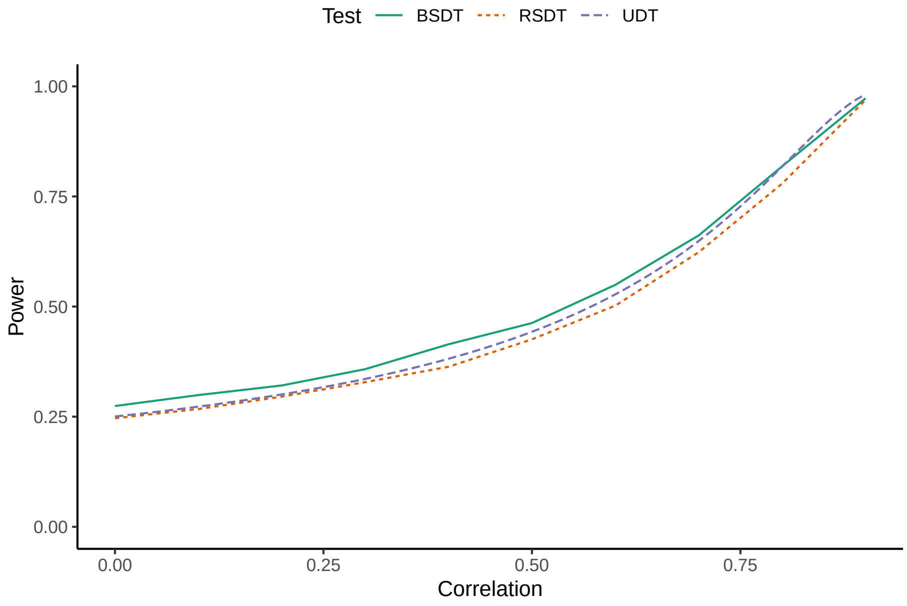

Figure 15: Power for UDT, RSDT and BSDT as a function of task correlation, $n = 16$, deficit$_a = 2$ SD below the mean and deficit$_b = 0$ SD below the mean. Number of simulations = 10 000, UDT derived analytically.

As pointed out by Crawford & Garthwaite (2007) the bigger difference between the tests turns out to be when the case's scores are extreme on both tasks. Since the BSDT is more conservative in these extremes than the RSDT, it is also likely that such circumstances would have a "negative" effect on power, for that test. Figure 16 demonstrates this to be the case. And as was shown by Crawford & Garthwaite (2007), the Type I error rate (i.e. "power" when there is no true discrepancy) is noticeably higher for RSDT which must be taken into account when evaluating performances of the tests.

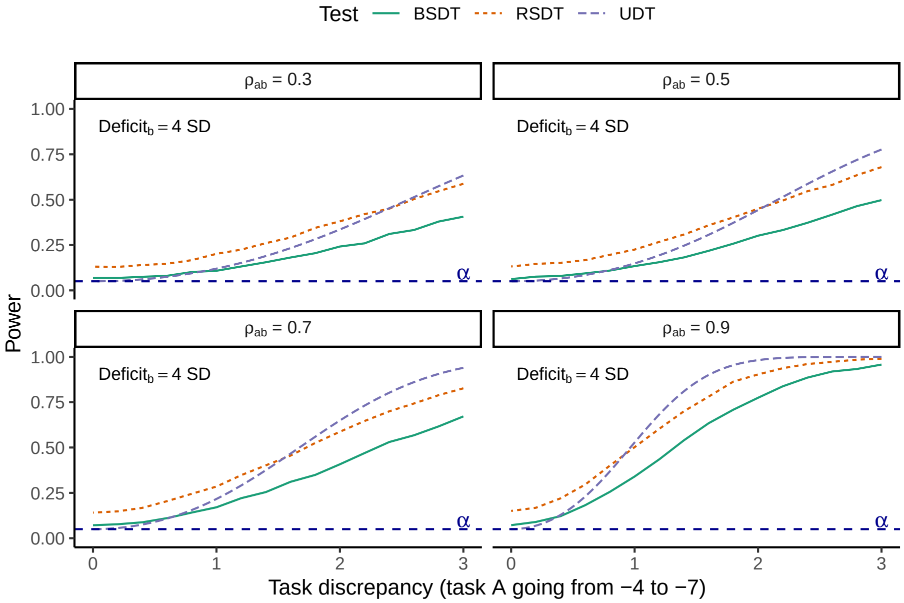

Figure 16: Comparison between UDT, RSDT and BSDT with extreme case scores on both task A and B, $n = 16$, $\rho_{AB} = 0.3,\ 0.5,\ 0.7,\ 0.9$. Number of simulations = 10 000, UDT derived analytically. Note that the difference between the RSDT and BSDT grow somewhat as the discrepancy becomes larger (for modest correlations) which indicates that RSDT becomes more liberal the farther the case is from the mean of the scores.

Furthermore, for UDT it does not matter if both scores are extreme, only discrepancy and correlation are taken into account. Hence, the curve for UDT in figure 16 has the same shape as it would have if the score on task B was near the mean. This reflects the difficulty of estimating the far ends of the score distributions with a sample. The far ends should be reflected in the estimated standard deviation and since the UDT does not depend on these estimates, the test is not affected and hence has a major power advantage without committing more Type I errors than specified.

4.4. Summary

In general power is low in case-control comparison designs and fairly large deficits or discrepancies must be exhibited by a case to have a chance to detect them. At least large in comparison to the effect sizes that are usually found in the social sciences. When testing for a deficit a sample size greater than ~ 16 is unnecessary. When testing for a dissociation one should utilise BSDT rather than RSDT, but if possible UDT, and the recommended sample size depends on the correlation between the tasks.

5. Quantifying and correcting the winner's curse

5.1. Tests of deficit

5.1.1. Frequentist test of deficit

For estimating the winner's curse in deficit studies one simply generate $n + 1$ random draws from a standard normal distribution. Take the $n + 1$th draw and impose the desired deficit by subtracting e.g. 1 for a deficit of 1 SD. Then take the $n+1$th draw and compare it to the nth other draws, with the test of interest. If the test turns out to be significant, then save the observed score. If repeated a large number of times this will yield a distribution of found deficits, which can be seen in figure 17. Taking an average of these values and subtracting the true (i.e. imposed) deficit then yields the average overestimation for the chosen parameter combinations. Because the distribution of found deficits will have a positive skew, taking the median instead of the mean is a more appropriate measure of central tendency. This was the procedure followed.

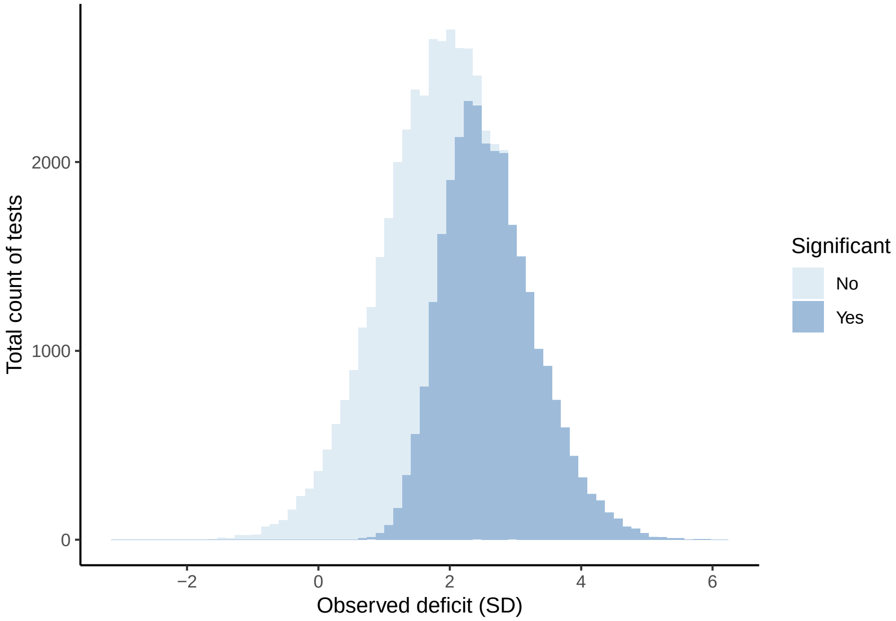

Figure 17: Histogram of all observed case values, i.e. after a deficit was imposed, shaded portion represents the deficits found. True underlying deficit was 2 SD, n = 16.

Since it is clear that power does not increase substantially over $n \sim 16$ and barely increase at all over $n = 30$, investigating the winner's curse with a larger control sample than 30 is not of interest. So a winner's curse simulation was run with $n = 5,\ 10,\ 15,\ 30$ and a deficit ranging from 0 to 5 standard deviations below the mean in steps of 0.2. The alternative hypothesis was one sided. For every parameter combination 1 000 000 simulations were run and the results can be observed in figure 18.

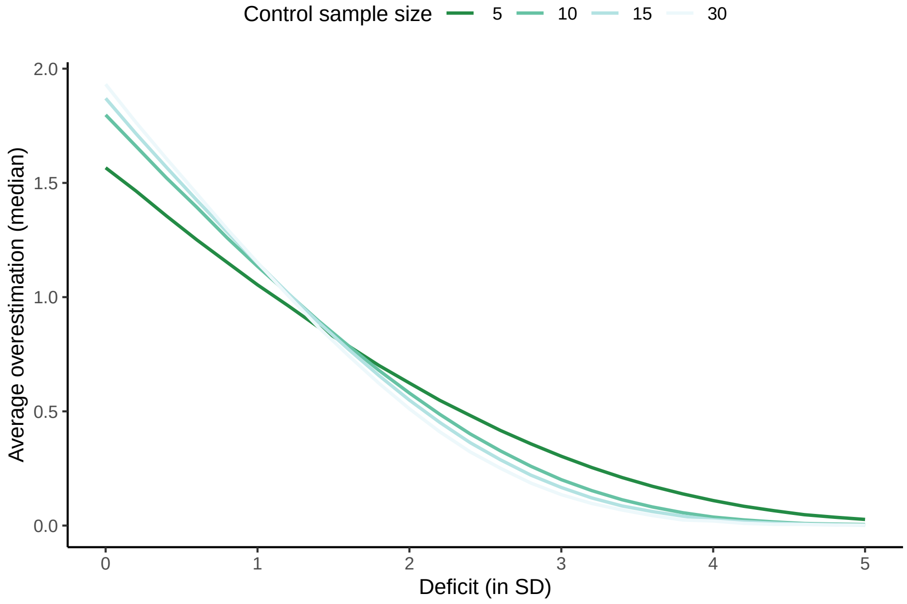

Figure 18: The average overestimation due to publication bias and the winner's curse, for TD, as a function of deficit size. Number of simulations = 1 000 000.

It seems evident that, just as for power, sample size does not have a large effect on the winner's curse. Perhaps the most surprising finding is that the average overestimation is less when the sample size is small and the deficit also small, but as the deficits become larger, increasing sample sizes decrease the error estimation. This seems like unexpected behaviour but when examined closer it is clear that it is a logical necessity.

This is because the sample means from small samples have greater variability than sample means from larger. This greater variability in means is demonstrated in figure 19, where the density of 5 random samples from a population with mean 100 and standard deviation 15 has been sampled with a size of 5 and 30 respectively.

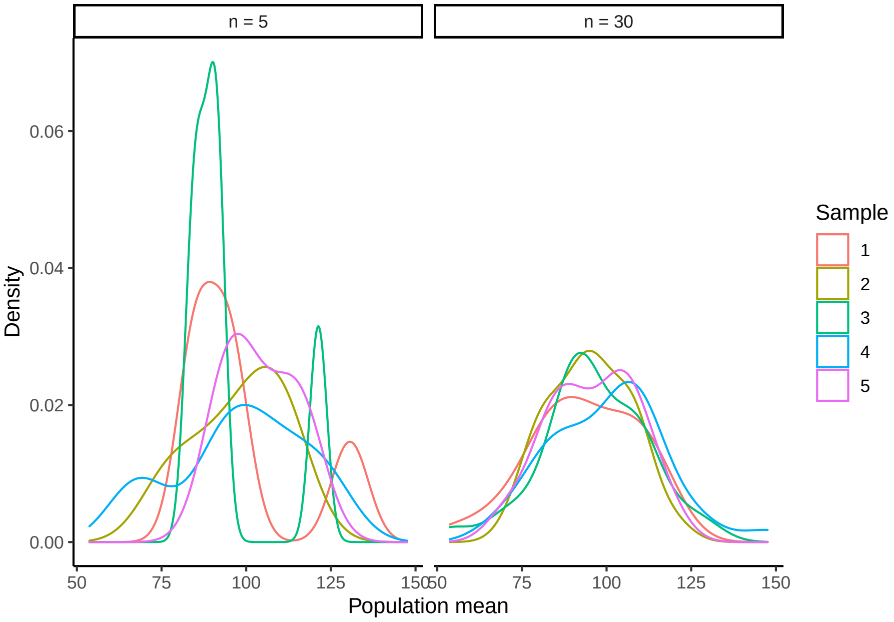

Figure 19: Visualisation of the density from different sizes of samples drawn from the same distribution.

Given this higher sampling variance it is easy to see that when a deficit is small it will only be detected when the random sample by chance happens to have a mean substantially higher than that of the population. Hence, the difference between the empirical and the true deficit does not have to be as large as when the samples better estimates the population mean. The number of *significant* deficits will, however, be a lot fewer.

An example might make this a bit more clear. Say that we have a control sample with 5 individuals and collect their scores on some IQ test. The population mean of IQ scores is 100 and the standard deviation 15. By chance we happen to draw 5 individuals that estimate the population mean to be 115 rather than 100. If a patient with a certain brain damage has a “true” deficit of half a standard deviation below the *population* mean (i.e. a deficit of 7.5). Such a small deviancy would not be detected.

However, to be deemed impaired the patient would only need an IQ below the critical value of a distribution with a mean of 115, rather than one with a mean of 100. Hence, the over estimation of the case does not have to be as large for the deficit to be significant as if the sample had estimated the mean correctly. So, because of the small effect size, a large portion of the significant deficits will be ones that are compared to misrepresenting samples. When the size of the deficit increase, they will more often be found when compared to good sample estimates, giving rise to the more typical winner's curse pattern even for small samples. This is since the proportion of deficits compared to misrepresenting samples grows smaller. The example is visualised in figure 20.

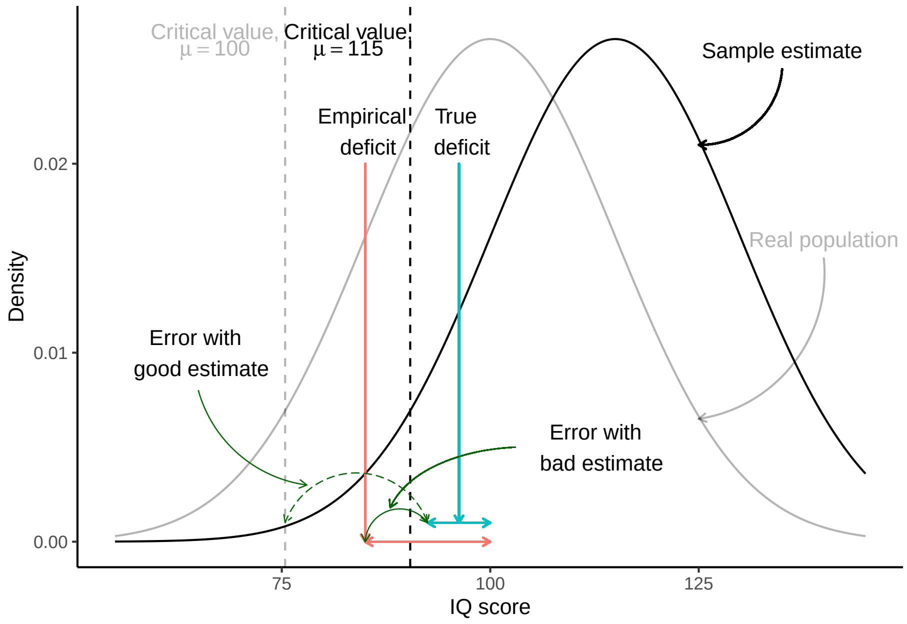

Figure 20: Visualisation of how small samples can yield a lower average overestimation when the true effects are small.

To confirm the limited effect of sample size on the winner's curse a second simulation was run with sample sizes ranging from 4 to 30 and deficit $= 0,\ 1,\ 2,\ 3,\ 4,\ 5$ standard deviations below the population mean. As can be seen in 21 the simulation confirmed the limited reducing effect of increasing sample sizes on the error estimation. Strikingly, the effect described above is even more apparent in figure 18 where the error estimations for a true deficit of 1 and 0 SD below the mean increases with n.

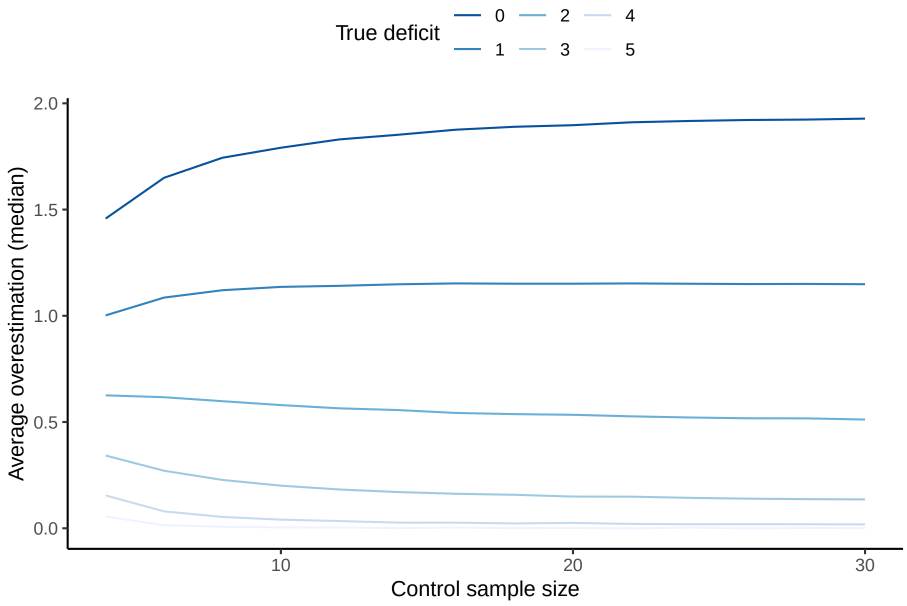

Figure 21: The average overestimation due to publication bias and the winner's curse, for TD, as a function of sample size. Number of simulations = 1 000 000.

5.1.2. Bayesian test of deficit

Given that we did not see a difference between TD and BTD in power, we would not expect to see a difference regarding error estimation. However, for consistency, a simulation study for the Bayesian test of deficit with the same parameters as for the frequentist test was conducted, apart from lowering the number of simulations to 100 000. The same patterns were exhibited, seen in figure 22. No more simulations were run for BTD.

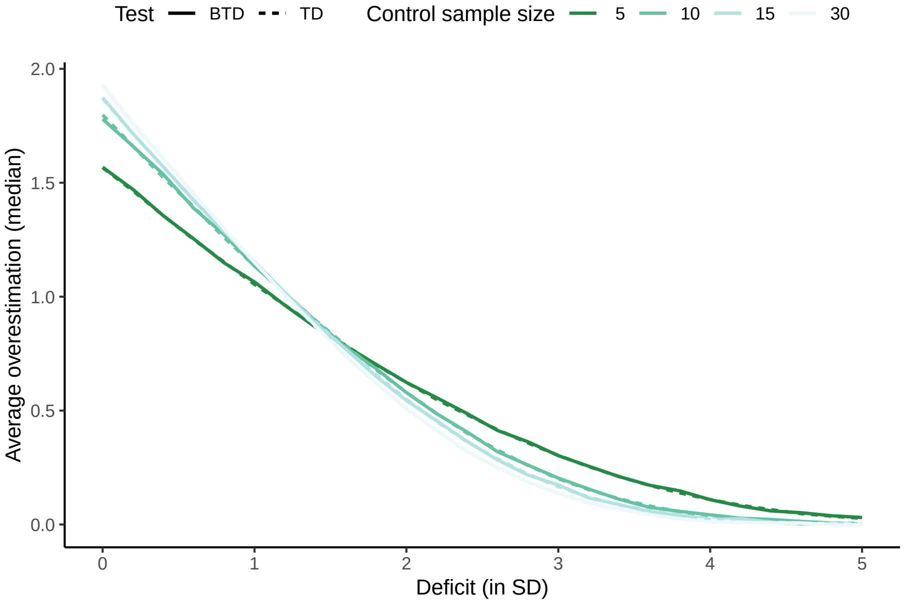

Figure 22: The average overestimation due to publication bias and the winner's curse, for BTD, superimposed on the graph for TD. Number of simulations = 100 000.

5.2. Tests of dissociation

For the tests of dissociation, the method of estimating the winner's curse was similar to that of the tests of deficit. The simulation procedure resembled that of the power simulations described in section 4.2.2, but for each significant test the observed task discrepancy was saved instead of the p-value. For each parameter combination this yielded a distribution of found discrepancies that then was averaged, as for the test of deficit. And similarly, since the distributions are skewed, the median is a more appropriate measure of central tendency.

The dissociation tests are normally run two-sided. However, when estimating found effect sizes, running simulations with two-sided tests when the true task discrepancy is small will yield a bimodal distribution of the found effects. When there is no true discrepancy at all, i.e. all effects found will be a Type I error, the 5% of found effects will be equally distributed at either side of the mean. Hence, the expected value of the error estimation for a two-sided test is not appropriately estimated with neither a median nor a mean. Observe figure 23. The discrepancies being part of the leftmost peak should be treated as Type I errors when the true discrepancy > 0 and should therefore not be included when averaging the error estimation. Because of this the winner's curse for dissociations was quantified with one-sided tests. Effectively ignoring discrepancies that would have been found with a negative sign and thus slightly increasing power. Hence, the winner's curse would have a slightly larger effect on the two-sided versions of the tests.

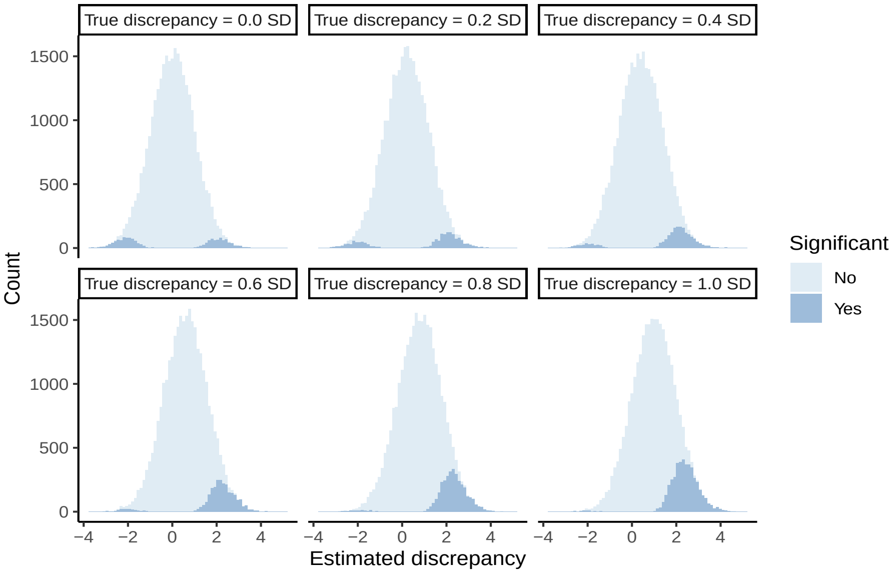

Figure 23: Distributions of discrepancies for a two-sided UDT. The found discrepancies create a bimodal distribution, i.e. a distribution with two peaks, making it difficult to get an unbiased measure of central tendency.

5.2.1. Simulations

The procedure of simulating the winner's curse was approximately the same for all three types of dissociation tests, with $n = 5,\ 10,\ 15,\ 30$ and a deficit on task A ranging from 0 to 5 standard deviations below the mean in steps of 0.2, while leaving functionality on task B unscathed. Correlation between the tasks was set at 0.3, 0.5, 0.7 and 0.9. The alternative hypothesis was one-sided. Each parameter combination was for UDT and RSDT run 100 000 times and for BSDT 10 000, because of the increased computational demand. Average error estimation was calculated as as the median of the found discrepancies subtracted from the true. Results for UDT, RSDT and BSDT are presented in figure 24, 25 and 26 respectively.

Generally we see the same surprising effect (explained previously) of small samples giving rise to more moderate error estimation when the true discrepancy is small as well. The average error estimation does seem to be slightly smaller for

BSDT than for RSDT as well as for RSDT compared to UDT. But no conclusion can be drawn without direct comparison.

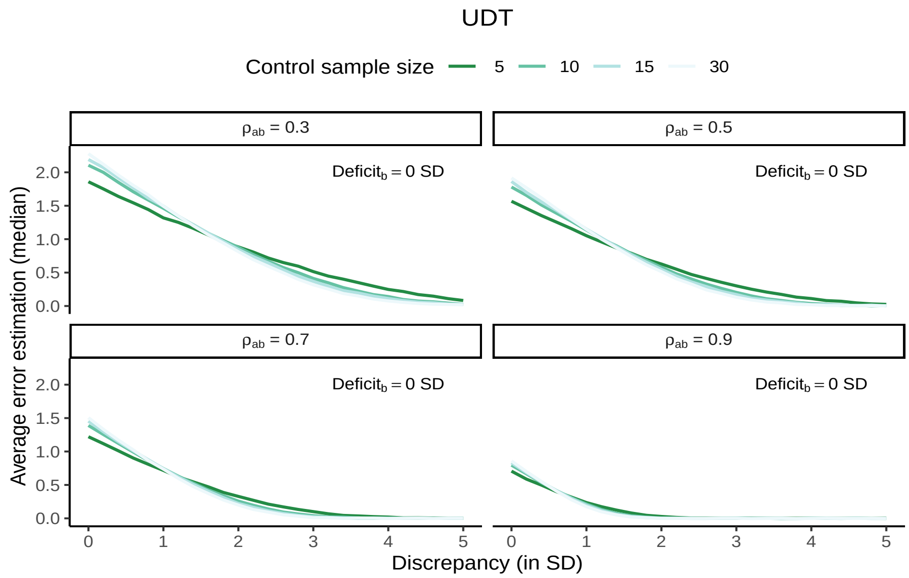

Figure 24: Average overestimation of discrepancy for UDT. Alternative hypothesis is one sided and there is no deficit on task B. ρ_{ab} = correlation between the tasks. Number of simulations = 100 000.

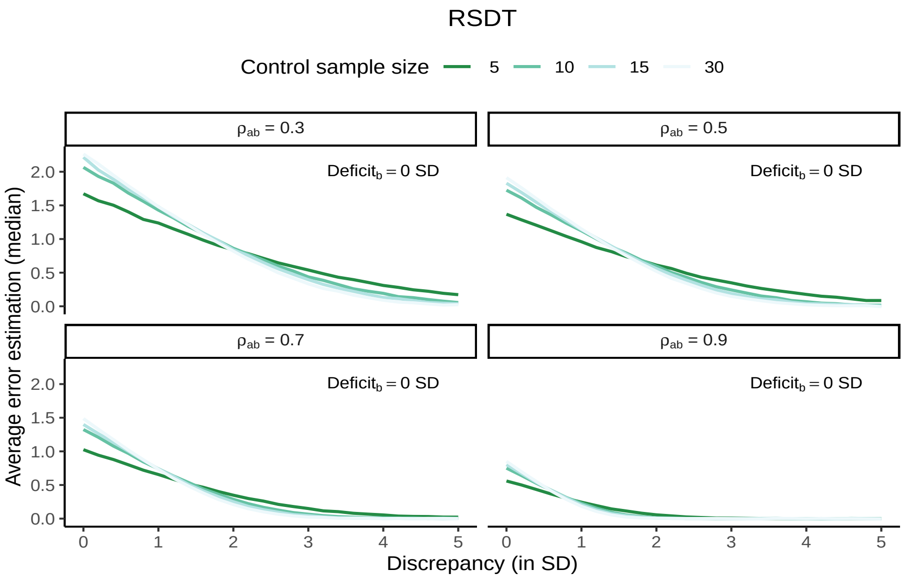

Figure 25: Average overestimation of discrepancy for RSDT. Alternative hypothesis is one sided and there is no deficit on task B. ρ_{ab} = correlation between the tasks. Number of simulations = 100 000.

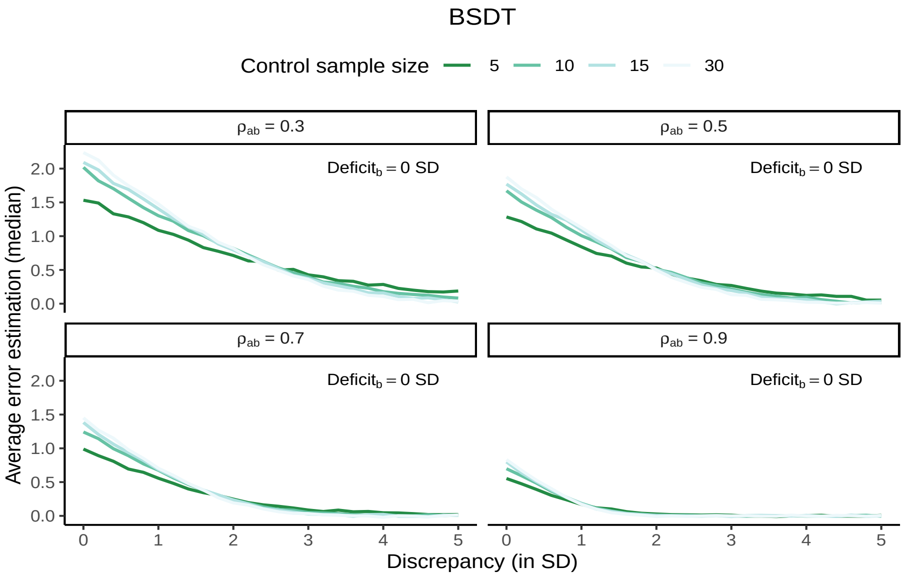

Figure 26: Average overestimation of discrepancy for BSDT. Alternative hypothesis is one sided and there is no deficit on task B. ρ_{ab} = correlation between the tasks. Number of simulations = 10 000. The higher uncertainty in the graph is due to the lower number of simulations.

5.2.2. Comparisons between the tests

Figure 27 demonstrates a comparison between the three dissociation tests, for a control sample size of 16 with a true discrepancy ranging from 0 to 5 standard deviations in steps of 0.2, where a deficit solely was imposed on task A while leaving task B unscathed.

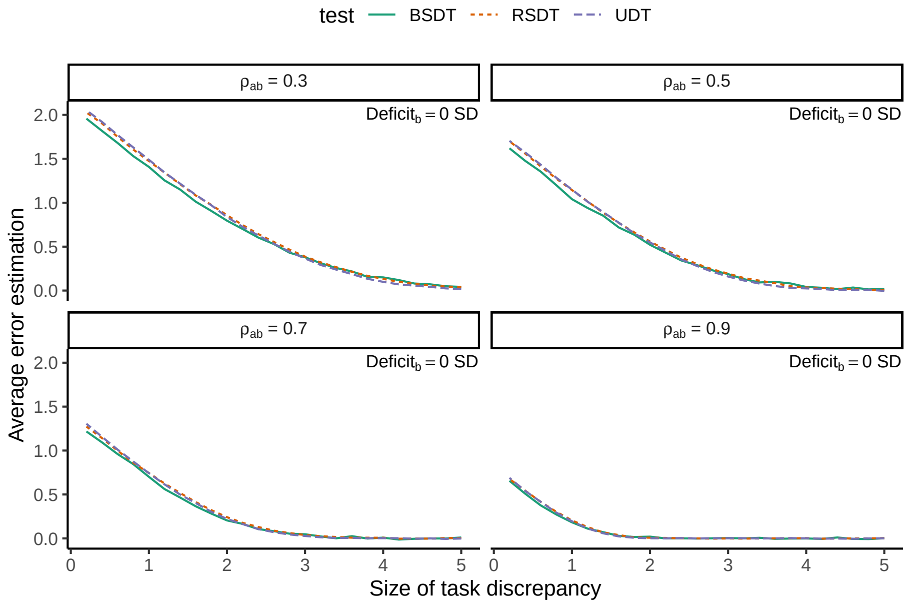

Figure 27: Comparing WC between UDT, RSDT, BDT with no deficit on task B and n = 16. Each parameter combination was run 100 000 times for UDT and RSDT and 10 000 for BSDT.

Figure 27 confirms the speculation regarding differences between the tests and seems to mirror the power differences from figure 15. I.e. BSDT tends to have the lowest error estimation followed by RSDT closely followed by UDT, for small discrepancies. However, as the true discrepancy increase both RSDT and BSDT start to produce larger errors than UDT. The starting advantage for BSDT in this case must be considered in relation to the slight disadvantage of controlling Type I errors.

As noted in 1.2.2.1, the major difference between the tests is demonstrated when both task scores are extreme. Hence, a simulation mirroring the previous but with a deficit of 4 SD imposed on task B, was run. The discrepancy between the tasks ranged from 0 to 5 standard deviations in steps of 0.2. Results can be observed in figure 28.

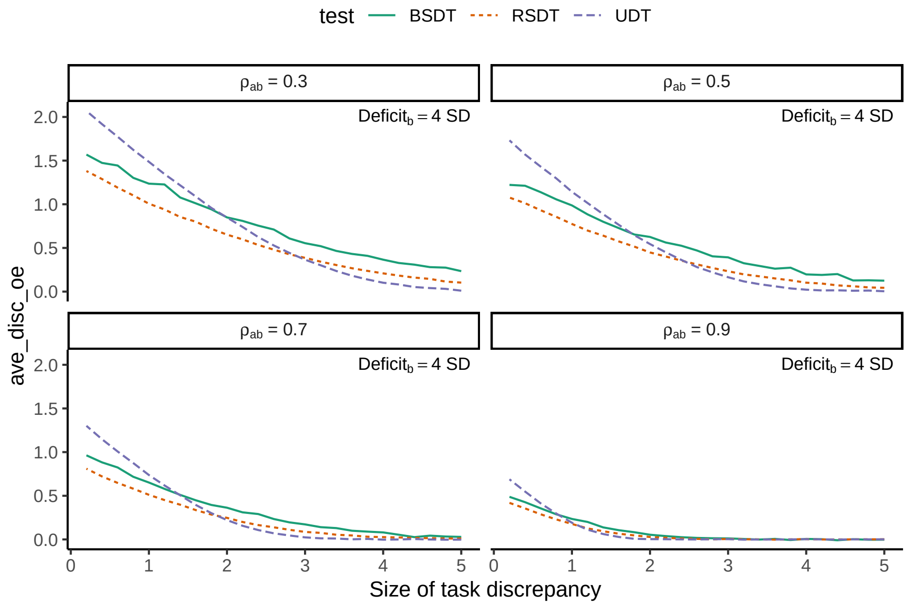

Figure 28: Comparing WC between UDT, RSDT, BSDT with a deficit on task B of 4 SD and n = 16. The deficit on task A ranged from 4 to 9 SD below the mean. ρ_{ab} = correlation between the tasks. Each parameter combination was run 100 000 times for UDT and RSDT and 10 000 for BSDT.

Figure 28 seems to mirror the power differences for extreme task scores. BSDT shows a consistent disadvantage to RSDT in the size of the error estimation, like it showed consistent disadvantage in power (figure 16). Likewise, UDT shows a larger error estimation than RSDT and BSDT when the true discrepancy, and hence power, is low. However, when the true discrepancy increases the error estimation for UDT becomes less than for both the other two, just as it showed a power advantage as effect sizes increased. The advantage of RSDT and BSDT compared to UDT must be seen in the light of their increased Type I error rate under these contingencies.

5.3. Correcting for winner's curse

When reading the single case neuropsychology literature through the lens of the publications bias it is clear that the winner's curse will affect the estimates of deficits or dissociations. It is desirable to adjust for this when conducting meta-analyses or for using unbiased estimates in power calculations. One way of doing so is to fit a linear model where the found (overestimated) deficit predicts the true underlying deficit.

5.3.1. Correcting TD

For the correction of the winner's curse a third degree polynomial regression with interaction was fitted to predict the true underlying effect from the significant median estimates and the number of controls. Further winner's curse simulations, with n ranging from 4 to 30 in steps of 2, were run to get larger variation in sample size to base the model on. In figure 29 the median estimates of found deficits are shown as a function of both true deficit and sample size. Without any bias at all we would expect the plane to be a completely flat and straight diagonal, as is shown in figure 30.

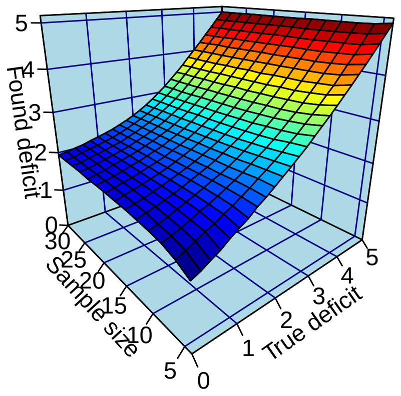

Figure 29: Averaged found deficits as a function of the true underlying effect and the sample size.

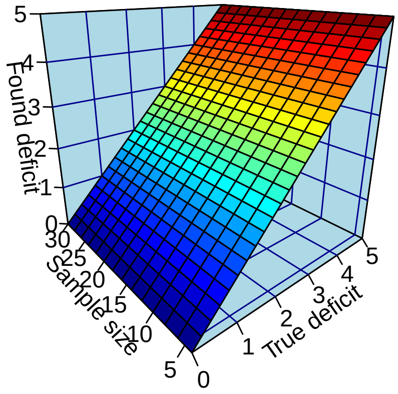

Figure 30: The relationship between sample size and, found and true deficits if we could estimate the underlyning effect perfectly.

The fitted model can be seen in table 5. Even though this might look excessively complicated there is no real risk of over-fitting. The model is based on simulations of the error estimation when the true deficit range from 0 to 5 SD in size and the control sample size range from 4 to 30. As seen in figure 18, there is almost no risk for a deficit to be over estimated if it is found to be > 5 so no correction would be needed, and a control sample larger than 30 will seldom be found in the literature. Therefore, it is not needed to predict out of sample data. Figure 31 shows the plane in figure 29 after having been corrected. To see the results of the method, a winner's curse simulation was run mimicking that of TD (figure 18), but each found deficit was corrected and then averaged. The results are shown in figure 32. It seems like the method works best for small to moderate samples but generally well for a size of 30 too. It is in the extremes the model have difficulties predicting. In particular we see errors when the sample size and size of the deficit is large, where the error estimation without correction was next to nothing.

Table 5

	Dependent variable:
	True deficit
Found deficit^1	-2.418^{**} (0.777)
Found deficit^2	1.042^{***} (0.247)
Found deficit^3	-0.095^{***} (0.025)
Sample size^1	-1.776^{***} (0.201)
Sample size^2	0.068^{***} (0.014)
Sample size^3	-0.001^{**} (0.0003)
Found deficit^1:Sample size^1	1.453^{***} (0.199)
Found deficit^2:Sample size^1	-0.374^{***} (0.062)
Found deficit^3:Sample size^1	0.031^{***} (0.006)
Found deficit^1:Sample size^2	-0.053^{***} (0.014)
Found deficit^2:Sample size^2	0.013^{**} (0.004)
Found deficit^3:Sample size^2	-0.001^{*} (0.0004)
Found deficit^1:Sample size^3	0.001^{**} (0.0003)
Found deficit^2:Sample size^3	-0.0002^{*} (0.0001)
Found deficit^3:Sample size^3	0.00001 (0.00001)
Constant	2.789^{***} (0.772)
Observations	364
R^2	0.999
Adjusted R^2	0.999
Residual Std. Error	0.047 (df = 348)
F Statistic	$25{,}079.810^{***}$ (df = 15; 348)
Note:	*p<0.05; **p<0.01; ***p<0.001
	Standard errors in parentheses
	Interactions denoted with ":"

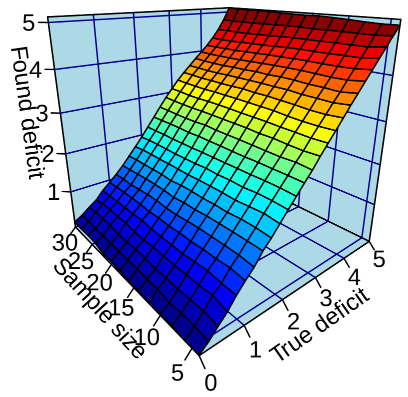

Figure 31: The averaged found deficits shown in figure 29 after having been corrected.

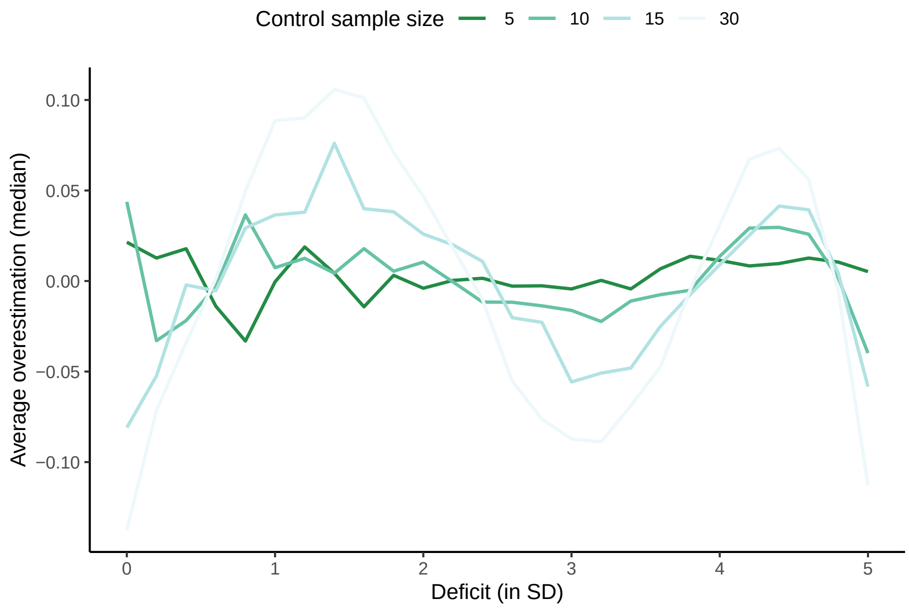

Figure 32: The average error estimation for TD after the correction has been applied to each found deficit. Number of simulations = 100 000.

5.3.2. Correcting UDT

Due to the fact that there are more factors at play for the dissociation tests a correction method was only devised for UDT. This is since its power curve does not change depending on the extremity of the task scores, so to capture the full range of error estimation we solely need to look at task discrepancy.

To correct for the winner's curse a third degree polynomial regression with a three-way interaction between the the average (over)estimated discrepancy, the number of controls and the correlation between the tasks was run to predict the true underlying task discrepancy. Since this is a model with 63 independent variables the specification of it can be observed in appendix B, it should be noted that even though fairly few variables obtain a high degree of significance we are more interested in how well it corrects the found discrepancies and this specification performed better than both higher and lower degree polynomials with/without interaction. However, although a radical improvement, as can be seen in figure 33 it does not perform as well as the correction method for TD. This is unsurprising considering the more complex task of estimating a hyper-plane instead of a plane. As can be seen, the extremes are most difficult to predict for this model as well, especially when both sample size and correlation is high. It is important to note that the "wave" patterns seen in both figure 32 and 33 hint that other model specifications could have captured the variation better, however, such models were not found.

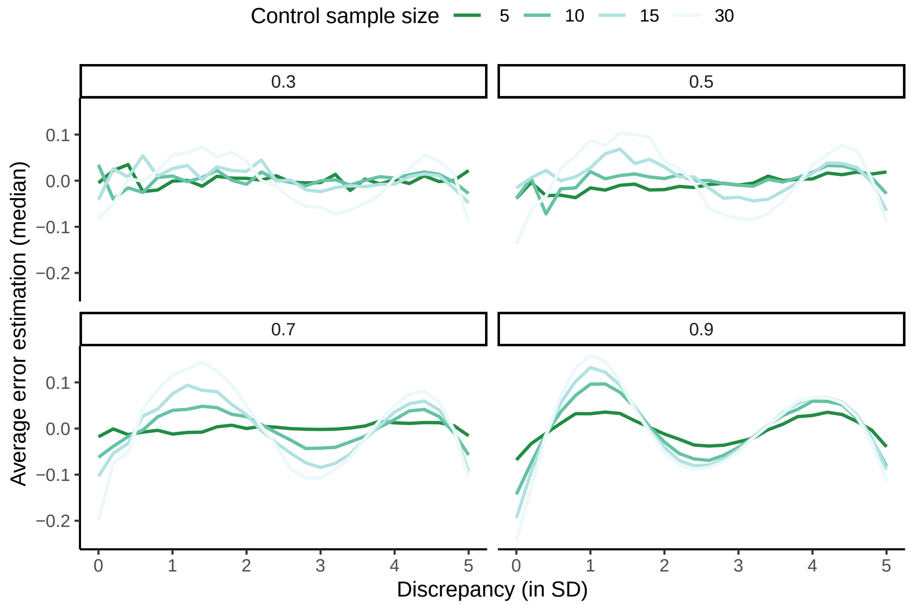

Figure 33: Results from a winner's curse simulation, mimicking that for UDT. But for each found discrepancy the correction method was applied. That is, a simulation with $n = 5,\ 10,\ 15,\ 30$ and a deficit on task A ranging from 0 to 5 standard deviations below the mean in steps of 0.2, while leaving functionality on task B unscathed. ρ_{ab} = correlation between the tasks. Number of simulations = 100 000.

5.4. Summary

This section has demonstrated the extent to which the winner's curse affects research where a case is compared to a control sample. This methodology is inherently low powered and as such, a notable average error estimation is to be expected. It was found that increasing sample sizes yielded limited reduction of the error. A surprising finding was that for smaller true effect sizes (<1.4 SD for deficits), small sample sizes tended to yield lower error estimations. Correction methods were developed for TD and UDT, which perform reasonably well. Implementing them in `singcar` would provide single case meta-analysts with a much needed tool.

6. Discussion

It is clear that single case neuropsychology fundamentally suffers for its low power. The winner's curse has here been shown to have a severe effect on studies with cases having (in neuropsychological terms) moderate "true" effect sizes.

Since brain damage can have detrimental effects on cognition, in many cases there is no need for concern. Deficits or discrepancies of 5 SD (and much lower for highly correlated constructs) gives rise to a power of 100% which also reduces the error estimation to 0.

However, in these cases you would need a hypothesis test because the deviation from normal functioning is so large. It is in the borderline cases the big problems arise. Perhaps it is just assumed that brain damage most often have detrimental effects on cognition because more subtle deficits/dissociations have been hidden in the normal ranges. It is also possible that, if found, they would help us paint a more nuanced picture of the human mind.

Could we make use of these cases by remedying the issues the field faces? Unfortunately, the root of the evil stems from the nature of null hypothesis significance testing (NHST). So as long as we want to use NHST, the issues will be dragged with it. This leaves us with three perhaps radical choices: i) give up single case neuropsychology; ii) stop using NHST; iii) continue with NHST but stop treating results as confirmatory.

The first alternative is undesirable since single cases still has great theoretical value for informing cognitive theories and cannot, in some instances, be replaced by group studies. The second alternative is perhaps a viable option. But giving up on the most widely used inferential method in science requires a feasible replacement. In the Bayesian framework of inference definite cutoff values should not be used since one is trying to quantify a *degree* of belief and the probability of a hypothesis can be updated as more information is gathered. Considering that the BTD produce identical results as the TD and that the BSDT is recommended over RSDT a switch of framework would not be practically difficult. Researchers would simply need to stop treating the estimates of abnormality from the tests as p-values in the frequentist sense. Because statistical power only exists if dichotomous cutoff values do, this change would also rid the field of all the issues that comes with it. However, it also comes with the cost of having to decide for yourself whether the estimates obtained are strong enough to support your hypothesis.

Lastly, if binary decision criteria are deemed important enough to continue with NHST, we might need to face the fact that we cannot view results from single case neuropsychology as confirmatory. That is until “enough” cases have been observed as to be able to use methods such as voxel based lesion-function mapping, for example. This simple change of perspective would not be practically difficult either and would allow for adjustments that would increase power, such as allowing a higher number of Type I errors by increasing α. It has for example been shown that $\alpha = 0.20$ is a better cutoff than 0.05 for inclusion of covariates in regression models (Lee, 2015). However, increasing Type I errors this much would, unsurprisingly, cause a large drop in the positive predictive value. Given a base rate of 10% and a power of 21%, as for the example given in section 2, the PPV would drop by almost two thirds, using equation (10). And as can be seen in figure 34, the increase in power yielded would at most result in a reduction of the error estimation by ~ 0.6 SD, for the test of deficit given a sample size of 16. Would it be worth it?

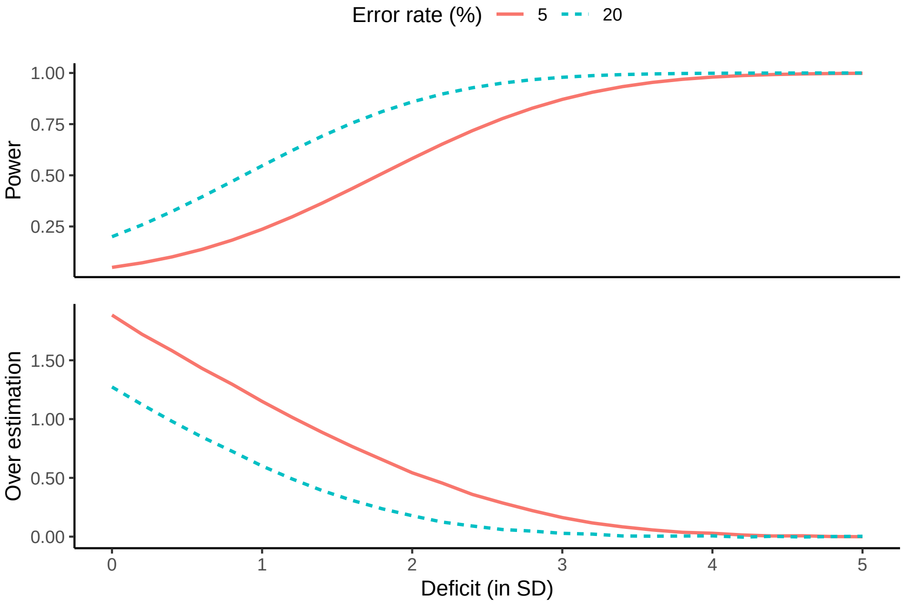

Figure 34: Power and overestimation for a Type I error rate of 5 and 20%.

If reduction of bias is the main goal it seems obvious that a shift towards Bayesian inference should be recommended. Given that the chance of all single case researchers being open to change inferential orientation seems slim, the third alternative might be the most feasible. If so, correction of the winner's curse would still be needed when reading the literature. This since significant studies will have a greater chance of being published, even though they are solely exploratory. Hence, the correction methods developed in 5.3 (or similar) could prove useful. If nothing else, they are useful when reading the literature hitherto published.

In group studies one can always try to increase power either by amping up the sample size or look for larger effects. Single case neuropsychology is inherently limited in these regards, since one cannot cause further neurological damage to increase an effect nor does the size of the sample resolve the issue above $n \sim 16$. However, carefully matching the control sample to the case or including covariates in the analysis are both ways to methodologically increase power, which is why the methods developed by Crawford et al. (2011) is a substantial contribution to the field and are strongly recommended.

Most importantly, if this route is taken, researchers must be aware of the fact that they are not conducting confirmatory experiments and should form their research questions and hypotheses accordingly. Creating patient profiles with multiple task measures should be encouraged, both for guiding future research and for diagnostics. When it comes to diagnostics it is even more important to be aware of the power issues. Because what would be most detrimental for an individual patient: claiming a deficit even if it is not there or failing to claim a deficit if it is?

Hopefully, the implementation of the Bayesian tests allowing for covariates in the R package `singcar` will allow for more widespread usage. This is true for TD, BTD, UDT, RSDT and BSDT as well, all recommended above other methods presented. The power calculators in the package gives researchers conducting meta-analyses tools for assessing the validity of studies. For future development further single case-comparison tools should be implemented, such as methods for comparing two cases to infer *double* dissociations (Crawford et al., 2010) or calculating interval estimates of positive predictive values (Crawford, Garthwaite, & Betkowska, 2009). All in all, gathering case-comparison research tools in a single software package with accompanying documentation should make them more accessible and in the long run benefit the field, whether the choice is made to switch inferential orientation, to go from confirmatory to strictly exploratory or to do nothing at all. It is my hope however that the present work has demonstrated that if nothing is done, inferences drawn from hypothesis tests in single case neuropsychology will lose their validity and the usage of the paradigm will decline even more.

7. Acknowledgments

I want to thank my supervisor Prof. Robert McIntosh for thoughtful guidance and encouragement throughout this project. The dissertation was written in

R Markdown (Allaire et al., 2020) using `knitr` (Xie, 2014) and `bookdown` (Xie, 2016). Plots and tables were created with `ggplot2` (Wickham, 2016), Latex, `kableExtra` (Zhu, 2019) and `stargazer` (Hlavac, 2018). Data was manipulated with `dplyr` (Wickham et al., 2020).

8. References

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2020). *Rmarkdown: Dynamic Documents for R*. https://github.com/rstudio/rmarkdown

Baddeley, A., Papagno, C., & Vallar, G. (1988). When long-term learning depends on short-term storage. *Journal of Memory and Language*, *27*(5), 586–595.

Berger, J. O., & Sun, D. (2008). Objective Priors for the Bivariate Normal Model. *The Annals of Statistics*, *36*(2), 963–982. https://www.jstor.org/stable/25464652

Broca, P. (1861). Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletin et Memoires de La Societe Anatomique de Paris*, *6*, 330–357.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Caplan, D. (1988). On the role of group studies in neuropsychological and pathopsychological research. *Cognitive Neuropsychology*, *5*(5), 535–547. https://doi.org/10.1080/02643298808253273

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, *5*(1), 41–66. https://doi.org/10.1016/0278-2626(86)90061-8

Caramazza, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and Language*, *21*(1), 9–20.

Caramazza, A., & McCloskey, M. (1988). The case for single-patient studies. *Cognitive Neuropsychology*, *5*(5), 517–527. https://doi.org/10.1080/02643298808253271

Chatterjee, A. (2005). A Madness to the Methods in Cognitive Neuroscience? *Journal of Cognitive Neuroscience*, *17*(6), 847–849. https://doi.org/10.

1162/0898929054021085

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997.

Coltheart, M. (2017). The assumptions of cognitive neuropsychology: Reflections on Caramazza (1984, 1986). *Cognitive Neuropsychology*, *34*(7-8), 397–402. https://doi.org/10.1080/02643294.2017.1324950

Crawford, J., & Garthwaite, P. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia*, *40*(8), 1196–1208. https://doi.org/10.1016/S0028-3932(01)00224-X

Crawford, J., & Garthwaite, P. (2006a). Detecting dissociations in single-case studies: Type I errors, statistical power and the classical versus strong distinction. *Neuropsychologia*, *44*(12), 2249–2258. https://doi.org/10.1016/j.neuropsychologia.2006.05.019

Crawford, J., & Garthwaite, P. (2007). Comparison of a single case to a control or normative sample in neuropsychology: Development of a Bayesian approach. *Cognitive Neuropsychology*, *24*(4), 343–372. https://doi.org/10.1080/02643290701290146

Crawford, J., & Garthwaite, P. (2006b). Methods of testing for a deficit in single-case studies: Evaluation of statistical power by Monte Carlo simulation. *Cognitive Neuropsychology*, *23*(6), 877–904. https://doi.org/10.1080/02643290500538372

Crawford, J., & Garthwaite, P. (2012). Single-case research in neuropsychology: A comparison of five forms of t-test for comparing a case to controls. *Cortex*, *48*(8), 1009–1016. https://doi.org/10.1016/j.cortex.2011.06.021

Crawford, J., & Garthwaite, P. (2005). Testing for Suspected Impairments and Dissociations in Single-Case Studies in Neuropsychology: Evaluation of Alternatives Using Monte Carlo Simulations and Revised Tests for Dissociations. *Neuropsychology*, *19*(3), 318–331. https://doi.org/10.1037/0894-4105.19.3.318

Crawford, J., Garthwaite, P., & Betkowska, K. (2009). Bayes' theorem and diagnostic tests in neuropsychology: Interval estimates for post-test probabilities. *The Clinical Neuropsychologist*, *23*(4), 624–644. https://doi.org/10.1080/13854040802524229

Crawford, J., Garthwaite, P., & Gray, C. (2003). Wanted: Fully Operational Definitions of Dissociations in Single-Case Studies. *Cortex*, *39*(2), 357–370. https://doi.org/10.1016/S0010-9452(08)70117-5

Crawford, J., Garthwaite, P., & Howell, D. (2009). On comparing a single case with a control sample: An alternative perspective. *Neuropsychologia*, *47*(13), 2690–2695. https://doi.org/10.1016/j.neuropsychologia.2009.04.011

Crawford, J., Garthwaite, P., Howell, D., & Gray, C. (2004). Inferential methods for comparing a single case with a control sample: Modified t-tests versus mycroft et al.'s (2002) modified anova. *Cognitive Neuropsychology*, *21*(7), 750–755. https://doi.org/10.1080/02643290342000276

Crawford, J., Garthwaite, P., & Ryan, K. (2011). Comparing a single case to a control sample: Testing for neuropsychological deficits and dissociations in the presence of covariates. *Cortex*, *47*(10), 1166–1178. https://doi.org/10.1016/j.cortex.2011.02.017

Crawford, J., Garthwaite, P., & Wood, L. (2010). Inferential methods for comparing two single cases. *Cognitive Neuropsychology*, *27*(5), 377–400. https://doi.org/10.1080/02643294.2011.559158

Crawford, J., & Howell, D. (1998). Comparing an Individual's Test Score Against Norms Derived from Small Samples. *The Clinical Neuropsychologist*, *12*(4), 482–486. https://doi.org/10.1076/clin.12.4.482.7241

Crawford, J., Howell, D., & Garthwaite, P. (1998). Payne and Jones Revisited: Estimating the Abnormality of Test Score Differences Using a Modified Paired Samples t Test. *Journal of Clinical and Experimental Neuropsychology*, *20*(6), 898–905. https://doi.org/10.1076/jcen.20.6.898.1112

Cumming, G., & Finch, S. (2001). A Primer on the Understanding, Use, and Calculation of Confidence Intervals that are Based on Central and Noncentral Distributions. *Educational and Psychological Measurement*, *61*(4), 532–574. https://doi.org/10.1177/0013164401614002

Damasio, H., Grabowski, T., Frank, R., Galaburda, A., & Damasio, A. (1994). The return of Phineas Gage: Clues about the brain from the skull of a famous patient. *Science*, *264*(5162), 1102–1105. https://doi.org/10.1126/science.8178168

DeGroot, M. H., & Schervish, M. J. (2012). *Probability and statistics* (4th ed). Addison-Wesley.

Donovan, T., & Mickey, R. M. (2019). *Bayesian Statistics for Beginners: A step-by-step approach* (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198841296.001.0001

Duval, S., & Tweedie, R. (2000). Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis.

Biometrics, *56*(2), 455–463. https://www.jstor.org/stable/2676988

Fellows, L. K., Heberlein, A. S., Morales, D. A., Shivde, G., Waller, S., & Wu, D. H. (2005). Method Matters: An Empirical Study of Impact in Cognitive Neuroscience. *Journal of Cognitive Neuroscience*, *17*(6), 850–858. https://doi.org/10.1162/0898929054021139

Garthwaite, P., & Crawford, J. (2004). The distribution of the difference between two t -variates. *Biometrika*, *91*(4), 987–994.

Goldberg, E. (1995). Rise and fall of modular orthodoxy. *Journal of Clinical and Experimental Neuropsychology*, *17*(2), 193–208. https://doi.org/10.1080/01688639508405118

Goodale, M. A., Milner, A. D., & others. (1992). *Separate visual pathways for perception and action.*

Harrison, D. A., & Brady, A. R. (2004). Sample Size and Power Calculations using the Noncentral t-distribution. *The Stata Journal*, *4*(2), 142–153. https://doi.org/10.1177/1536867X0400400205

Hlavac, M. (2018). *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), 6.

Jeffreys, H. (1998). *Theory of probability* (3rd ed). Clarendon Press ; Oxford University Press.

Kosslyn, S. M., & Intriligator, J. M. (1992). Is Cognitive Neuropsychology Plausible? The Perils of Sitting on a One-Legged Stool. *Journal of Cognitive Neuroscience*, *4*(1), 96–105. https://doi.org/10.1162/jocn.1992.4.1.96

Labarge, A. S., Mccaffrey, R. J., & Brown, T. A. (2003). Neuropsychologists' abilities to determine the predictive value of diagnostic tests. *Archives of Clinical Neuropsychology*, *18*(2), 165–175.

Lee, P. H. (2015). Should we adjust for a confounder if empirical and theoretical criteria yield contradictory results? A simulation study. *Scientific Reports*, *4*(1), 6085. https://doi.org/10.1038/srep06085

McIntosh, R. D. (2018). Simple dissociations for a higher-powered neuropsychology. *Cortex*, *103*, 256–265. https://doi.org/10.1016/j.cortex.2018.03.015

Medina, J., & Fischer-Baum, S. (2017). Single-Case Cognitive Neuropsychology in the Age of Big Data. *Cognitive Neuropsychology*, *34*(7-8), 440–448. https://doi.org/10.1080/02643294.2017.1321537

Mycroft, R. H., Mitchell, D. C., & Kay, J. (2002). An evaluation of statistical procedures for comparing an individual's performance with that of a group of controls. *Cognitive Neuropsychology*, *19*(4), 291–299. https://doi.org/10.1080/02643290143000150

Patterson, K., & Plaut, D. C. (2009). "Shallow Draughts Intoxicate the Brain": Lessons from Cognitive Science for Cognitive Neuropsychology. *Topics in Cognitive Science*, *1*(1), 39–58. https://doi.org/10.1111/j.1756-8765.2008.01012.x

Payne, R. W., & Gwynne Jones, H. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology*, *13*(2), 115–121.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Robertson, L. C., Knight, R. T., Rafal, R., & Shimamura, A. P. (1993). *Cognitive Neuropsychology Is More Than Single-Case Studies*. 8.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11–21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC497229/

Seidenberg, M. S. (1988). Cognitive neuropsychology and language: The state of the art. *Cognitive Neuropsychology*, *5*(4), 403–426. https://doi.org/10.1080/02643298808253267

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.

Sokal, R. R., & Rohlf, F. J. (1981). *Biometry: The Principles and Practice of Statistics in Biological Research*. W. H. Freeman. https://books.google.co.uk/books?id=C-OTQgAACAAJ

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. http://www.stats.ox.ac.uk/pub/MASS4

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr

Xie, Y. (2014). Knitr: A Comprehensive Tool for Reproducible Research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing Reproducible Computational Research*. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595

Xie, Y. (2016). *Bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman; Hall/CRC. https://github.com/rstudio/bookdown

Zhu, H. (2019). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra

9. Appendix A

9.1. R package `singcar`

9.1.1. Functions

The main functions of `singcar` are hitherto:

Frequentist tests:

- `TD()`: The test of deficits (equation (3)).
- `UDT()`: The unstandardised difference test (equation (4)).
- `RSDT()`: The revised standardised difference test (equation (7)).

Bayesian tests:

- `BTD()`: Bayesian test of deficit.
- `BSDT()`: Bayesian standardised difference test.
- `BTD_cov()`: Bayesian test of deficit with covariates.
- `BSDT_cov()`: Bayesian standardised difference test with covariates.

Power calculators:

- `TD_power()`: Calculates exact power given sample size or necessary sample size for desired power using analytical methods for the test of deficit.
- `BTD_power()`: Calculates approximate power given sample size using Monte Carlo simulations for the Bayesian test of deficit.
- `UDT_power()`: Calculates exact power given sample size or necessary sample size for desired power using analytical methods for the unstandardised difference test.
- `RSDT_power()`: Calculates approximate power given sample size using Monte Carlo simulations for the revised standardised difference test.
- `BSDT_power()`: Calculates approximate power given sample size using Monte Carlo simulations for the Bayesian standardised difference test.

9.1.1.1. Function `TD()`. Crawford & Howell (1998)'s modified t-test, testing for deficits. Takes a single observation and compares it to a distribution estimated by a control sample. Calculates standardised difference between a case and the mean of the controls and proportions falling above or below the case score (p-value), as well as associated confidence intervals. The function calculates the point estimate of the standardised difference (an effect size labeled zcc) between the case score and the mean of the controls and the point estimate of the p-value (i.e. the percentage of the population that would be expected to obtain a lower or higher score, depending on the alternative hypothesis).

Usage:

```
TD(case, controls, controls.sd = NULL, sample_size = NULL,
alternative = c("less", "greater", "two.sided"), conf_int
= TRUE, conf_level = 0.95, conf_int_spec = 0.01, na.rm = FALSE)
```

Argument descriptions:

- `case`: Single value from case observation on some task.
- `controls`: Numeric vector of observations from the control sample. If single value, treated as mean.
- `controls.sd`: If input of controls is single value (and thus treated as mean), the standard deviation of the sample must be given as well.
- `sample_size`: If input of controls is single value (and thus treated as mean), the size of the sample must be given as well.
- `alternative`: A character string specifying the alternative hypothesis, must be one of "less" (default), "greater" or "two.sided". It is possible to specify using only the initial letter.
- `conf_int`: If set to `TRUE` (default) initiates a search algorithm to find the confidence intervals of the effect size and the p-value. The algorithm uses the non-central t-distribution, searching for the non-central distributions having $zcc\sqrt{n}$ as its 2.5th and 97.5th quantile, for a 95% CI (default). Setting the non-centrality parameters of these distributions divided by the square root of n as the upper and lower interval limits, respectively. More details of the procedure can be found in Cumming & Finch (2001) and Crawford & Garthwaite (2002). Set to `FALSE` for faster calculation. Note though that calculating the confidence intervals depends on the degrees of freedom, the confidence level and the effect size. For some extreme distributions exact accuracy from the `stats::qt()` function (the quantile

function for the t-distribuiton) used can not be guaranteed. However, the approximations should be good enough for most cases.
- `conf_level`: Sets the level of confidence for the intervals (e.g. 95%, 99% etc.), defaults to 95%.
- `conf_int_spec`: Specifies the size of iterative steps for calculating confidence intervals. Smaller values gives more precise intervals but takes longer to calculate. Defaults to 0.01.
- `na.rm`: If set to `TRUE` removes NA from `controls`.

9.1.1.2. Function `UDT()`. `UDT()` is the modified paired samples t-test for testing dissociations that most resembles the original test by Payne & Gwynne Jones (1957), i.e. the unstandardised version of the test developed in Crawford, Howell, et al. (1998). It should solely be used when the two tasks under investigation are measured on the the *same* scale. Takes two (case) observations and compares the size of the difference between them to the distribution of differences estimated by a control sample. The function calculates the point estimate of the difference as well as the proportion of the population that would be expected to exhibit a more extreme task discrepancy.

Usage:

```
UDT(case_a, case_b, controls_a, controls_b, sd_a = NULL, sd_b
= NULL, sample_size = NULL, r_ab = NULL, alternative = c("two.sided",
"greater", "less"), na.rm = FALSE)
```

Argument descriptions:

- `case_a`: Single value from case observation on task A.
- `case_b`: Single value from case observation on task B.
- `controls_a`: Controls' scores on task A. Takes either a vector of observations or a single value interpreted as mean. Note: a vector can be supplied as input for task A while mean and SD for task B.
- `controls_b`: Controls' scores on task B. Takes either a vector of observations or a single value interpreted as mean. Note: a vector can be supplied as input for task B while mean and SD for task A.
- `sd_a`: If single value for task A is given as input the standard deviation of task A must be supplied.
- `sd_b`: If single value for task B is given as input the standard deviation of task B must be supplied.

- `sample_size`: If A or B is given as mean and SD the sample size must be supplied. If controls_a is given as vector and controls_b as mean and SD, sample_size must equal the number of observations in controls_a.
- `r_ab`: If A or B is given as mean and SD, the correlation between the tasks must be supplied.
- `alternative`: A character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter. Since the direction of the expected effect depends on which task is set as A and which is set as B, be very careful if changing this parameter.
- `na.rm`: If set to `TRUE` removes NA from `controls`.

9.1.1.3. Function `RSDT()`. This is the revised standardised difference test for assessing dissociations, developed in Garthwaite & Crawford (2004) and Crawford & Garthwaite (2005). Takes two (case) observations, standardise them and compares the size of the difference between them to the distribution of the standardised differences estimated by a control sample. The function calculates an approximate point estimate of the difference as well as the proportion of the population that would be expected to exhibit a more extreme task discrepancy.

Usage:

```
RSDT(case_a, case_b, controls_a, controls_b, sd_a = NULL,
sd_b = NULL, sample_size = NULL, r_ab = NULL, alternative
= c("two.sided", "greater", "less"), exact.method = T, alpha
= 0.05, na.rm = FALSE)
```

Argument descriptions:

- `case_a`: Single value from case observation on task A.
- `case_b`: Single value from case observation on task B.
- `controls_a`: Controls' scores on task A. Takes either a vector of observations or a single value interpreted as mean. Note: you can supply a vector as input for task A while mean and SD for task B.
- `controls_b`: Controls' scores on task B. Takes either a vector of observations or a single value interpreted as mean. Note: you can supply a vector as input for task B while mean and SD for task A.
- `sd_a`: If single value for task A is given as input the standard deviation of task A must be supplied.

- `sd_b`: If single value for task B is given as input the standard deviation of task B must be supplied.
- `sample_size`: If A or B is given as mean and SD the sample size must be supplied. If controls_a is given as vector and controls_b as mean and SD, sample_size must equal the number of observations in controls_a.
- `r_ab`: If A or B is given as mean and SD, the correlation between the tasks must be supplied.
- `alternative`: A character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter. Since the direction of the expected effect depends on which task is set as A and which is set as B, be very careful if changing this parameter.
- `na.rm`: If set to `TRUE` removes NA from `controls`.
- `exact.method`: If set to FALSE generates an approximate t-statistic used to derive the exact. The exact method can only generate an absolute t-statistic. The approximate method also requires a pre-set alpha-value, see Crawford & Garthwaite (2005) for more information. Not using the exact method is only appropriate when a precise p-value is not needed.
- `alpha`: Chosen risk of Type I errors. This is only relevant if setting exact.method = FALSE due to the test statistic depending on the critical value chosen.

9.1.1.4. Function `BTD()`. This is the Bayesian analouge of `TD()`, developed by Crawford & Garthwaite (2007). Just as `TD()` it takes a single observation and compares it to a distribution estimated by a control sample. Calculates standardised difference between case and controls. The p-value is produced by simulating a distribution of possible means based on the prior and the mean from the sample, as more thoroughly explained in Bayesian approaches. The function calculates the point estimate of the standardised difference between the case score and the mean of the controls and the point estimate of the p-value (i.e. the percentage of the population that would be expected to obtain a lower or higher score, depending on the alternative hypothesis) as well as associated credible intervals.

Usage:

```
BTD(case, controls, controls.sd = NULL, sample_size = NULL,
alternative = c("less", "greater", "two.sided"), int.level
= 0.95, iter = 1000, na.rm = FALSE)
```

Argument description:

- `case`: Single value from a case observation, on some task.
- `controls`: Numeric vector of observations from the control sample. If single value, treated as mean.
- `controls.sd`: If input of controls is single value, the standard deviation of the sample must be given as well.
- `sample_size`: If input of controls is single value, the size of the sample must be given as well.
- `alternative`: A character string specifying the alternative hypothesis, must be one of "less" (default), "greater" or "two.sided". You can specify just the initial letter.
- `int.level`: Level of confidence for credible intervals.
- `iter`: Number of simulations to run for the test. Greater number gives better estimation but takes longer to calculate.
- `na.rm`: Defaults to `FALSE` if set to `TRUE` removes any `NA´`s from controls.

9.1.1.5. Function `BSDT()`. This is the Bayesian analouge of `RSDT()` and `UDT()`, developed by Crawford & Garthwaite (2007). Takes two case observations on task A and B, compares the size of the difference between them to the distribution of differences estimated by a control sample and tests if there is a significant abnormality. BSDT provides an exact point estimate of a case's abnormality of task discrepancy on some tasks, in contrast to RSDT which only approximates it. Furthermore, it solves the problem of setting interval limits on the estimations, which have proved difficult with RSDT.

If the tasks are measured on the same scale one have the oppurtunity to run the function without standardisation, making it comparable to UDT. Two types of priors can be used. Either the standard theory/Jeffrey's prior (Jeffreys, 1998) or a calibrated prior, which is similar to the standard theory but an accept/reject algorithm is applied on the initial random. Crawford et al. (2011) recommends this latter prior as it has been shown to have better frequentist properties for estimating ρ_{ab} and the size of the discrepancy, compared to Jeffrey's. The function calculates the point estimate of the task difference for the case, as well as the proportion of the population that would be expected to exhibit a more extreme task discrepancy, along with associated credible intervals.

Usage:

```
BSDT(case_a, case_b, controls_a, controls_b, sd_a = NULL,
sd_b = NULL, sample_size = NULL, r_ab = NULL, alternative
= c("two.sided", "greater", "less"), int.level = 0.95, iter
= 1000, unstandardised = FALSE, calibrated = FALSE, na.rm
= FALSE)
```

Argument descriptions:

- `case_a`: Single value from case observation on task A.
- `case_b`: Single value from case observation on task B.
- `controls_a`: Controls' scores on task A. Takes either a vector of observations or a single value interpreted as mean. Note: you can supply a vector as input for task A while mean and SD for task B.
- `controls_b`: Controls' scores on task B. Takes either a vector of observations or a single value interpreted as mean. Note: you can supply a vector as input for task B while mean and SD for task A.
- `sd_a`: If single value for task A is given as input the standard deviation of task A must be supplied.
- `sd_b`: If single value for task B is given as input the standard deviation of task B must be supplied.
- `sample_size`: If A or B is given as mean and SD the sample size must be supplied. If controls_a is given as vector and controls_b as mean and SD, sample_size must equal the number of observations in controls_a.
- `r_ab`: If A or B is given as mean and SD, the correlation between the tasks must be supplied.
- `alternative`: A character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter. Since the direction of the expected effect depends on which task is set as A and which is set as B, be very careful if changing this parameter.
- `int.level`: Level of confidence for credible intervals.
- `iter`: Number of simulations to run for the test. Greater number gives better estimation but takes longer to calculate.
- `unstandardised`: Estimate z-value based on standardised or unstandardised task scores.
- `calibrated`: Set to TRUE to use a calibrated prior distribution.
- `na.rm`: Remove NAs from controls.

9.1.1.6. Function `BTD_cov()`. This function runs the Bayesian test of deficit on a task of interest for the single case, while controlling for the effects of covariates.

A typical covariate in single case neuropsychology could for example be years of education. Say that a case has 12 years of education, he/she would then, when tested on some task A, be compared on this task to the participants in the control sample also with 12 years of education. I.e. the conditioned case score is, under the null hypothesis, an observation from the conditional task distribution estimated by the control population. The function calculates the standardised difference between case and the mean of the controls and proportions falling above or below the case score (p-value), as well as associated credible intervals.

Usage:

```
BTD_cov(case_task, case_covar, control_task, control_covar,
alternative = c("less", "two.sided", "greater"), int.level
= 0.95, iter = 1000, use_sumstats = FALSE, cor_mat = NULL,
control_n = NULL)
```

Argument descriptions:

- `case_task`: Single value from a case observation, on some task.
- `case_covar`: A vector containing the case scores on all covariates included. Can be of any length except 0, in that case use BTD.
- `control_task`: A vector containing the scores from the controls on the task of interest. Or a vector of length 2 containing the mean and standard deviation of the task. In that order.
- `control_covar`: A vector, matrix or dataframe cointaining the control scores on the covariates included. If matrix or dataframe each column represents a covariate. Or a matrix or dataframe containing summary statistics where the first column represents the means for each covariate and the second column represents the standard deviation(s).
- `alternative`: A character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter.
- `int.level`: The probability level on the Bayesian credible intervals.
- `iter`: Number of simulations to run for the test. Greater number gives better estimation but takes longer to calculate.
- `use_sumstats`: If set to TRUE, control_tasks and control_covar are treated as matrices with summary statistics. Where the first column represents the means for each variable and the second column represents the standard deviation.

- `cor_mat`: A correlation matrix of all variables included. NOTE: the first variable should be the tasks of interest.
- `control_n`: An integer specifying the sample size of the controls.

9.1.1.7. Function `BSDT_cov()`. This function takes two case observations on task A and B, standardise them and tests whether the difference between them is larger than the standardised differences in the control sample, while controlling for the effects of covariates. Using the same example as in the previous section, it could for example be education. The case would be compared to a the control population, having the same length of education. The function calculates the point estimate of the standardised task difference of the case, as well as the proportion of the population that would be expected to exhibit a more extreme task discrepancy, along with associated credible intervals.

Two types of priors can be used. Either the standard theory/Jeffrey's prior (Jeffreys, 1998) or a calibrated prior, a variant of a prior investigated in Berger & Sun (2008) developed in Crawford et al. (2011). The latter is recommended.

Usage:

```
BSDT_cov(case_tasks, case_covar, control_tasks, control_covar,
alternative = c("two.sided", "greater", "less"), int.level
= 0.95, calibrated = TRUE, iter = 1000, use_sumstats = FALSE,
cor_mat = NULL, control_n = NULL)
```

Argument descriptions:

- `case_tasks`: A vector of length 2. The case scores from the two tasks.
- `case_covar`: A vector containing the case scores on all covariates included.
- `control_tasks`: A matrix or dataframe with 2 columns and n rows containing the control scores for the two tasks. Or a matrix or dataframe containing summary statistics where the first column represents the means for each task and the second column represents the standard deviation.
- `control_covar`: A matrix or dataframe cointaining the control scores on the covariates included. Or a matrix or dataframe containing summary statistics where the first column represents the means for each covariate and the second column represents the standard deviation.
- `alternative`: A character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter. Since the direction of the expected effect depends on

which task is set as A and which is set as B, be very careful if changing this parameter.

- `int.level`: The probability level on the Bayesian credible intervals.
- `calibrated`: Whether or not to use the standard theory (Jeffreys) prior distribution (if set to FALSE) or a calibrated prior examined by Berger and Sun (2008) and sample size treated as n - 1. See Crawford et al. (2011) for further information. Calibrated prior is recommended.
- `iter`: Number of simulations to be performed. Greater number gives better estimation but takes longer to calculate.
- `use_sumstats`: If set to TRUE, control_tasks and control_covar are treated as matrices with summary statistics. Where the first column represents the means for each variable and the second column represents the standard deviation.
- `cor_mat`: If `use_sumstats` set to `TRUE` a correlation matrix of all variables included must be supplied. NOTE: the two first variables should be the tasks of interest.
- `control_n`: If `use_sumstats` set to `TRUE` an integer specifying the sample size of the controls must be supplied.

9.1.1.8. Function `TD_power()`. Calculates exact power given sample size or sample size given power, using analytical methods for the frequentist test of deficit for a specified case score and, mean and standard deviation for the control sample. The mean and standard deviation defaults to 0 and 1, so if no other values are given the case score is interpreted as deviation from the mean in standard deviations.

Usage:

```
TD_power(case, mean = 0, sd = 1, sample_size = NULL, power
= NULL, alpha = 0.05, spec = 0.005)
```

Argument descriptions:

- `case`: A single value from the expected case observation.
- `mean`: The expected mean of the control sample.
- `sd`: The expected standard deviation of the control sample.
- `sample_size`: The size of the control sample, vary this parameter to see how the sample size affects power. One of sample size or power must be specified, not both.

- `power`: A single value between 0 and 1 specifying desired power for calculating necessary sample size. One of sample size or power must be specified, not both.
- `alpha`: The specified Type I error rate. This can also be varied, with effects on power.
- `spec`: A single value between 0 and 1. If desired power is given as input the function will utilise a search algorithm to find the sample size needed to reach the desired power. However, if the power specified is greater than what is actually possible to achieve the algorithm could search forever. Hence, when power does not increase substantially for any additional participant in the sample, the algorithm stops. By default the algorithm stops when power does not increase more than 0.5% for any additional participant, this specificity can be changed.

9.1.1.9. Function `BTD_power()`. Calculates approximate power, given sample size, using Monte Carlo simulation for the Bayesian test of deficit for a specified case score, mean and standard deviation for the control sample. The mean and standard deviation defaults to 0 and 1, so if no other values are given the case score is interpreted as deviation from the mean in standard deviations. Returns a single value approximating the power of the test for the given parameters.

Usage:

```
BTD_power(case, mean = 0, sd = 1, sample_size, alternative
= c("two.sided", "one.sided"), alpha = 0.05, nsim = 1000,
iter = 1000)
```

Argument descriptions:

- `case`: A single value from the expected case observation.
- `mean`: The expected mean of the control sample.
- `sd`: The expected standard deviation of the control sample.
- `sample_size`: The size of the control sample, vary this parameter to see how the sample size affects power.
- `alternative`: The alternative hypothesis. A string of either “two.sided” (default) or “one.sided”.
- `alpha`: The specified Type I error rate. This can also be varied, with effects on power.
- `nsim`: The number of simulations for the power calculation. Defaults to 1000 due to BTD already being computationally intense. Defaults to 1000.

- `iter`: The number of simulations used by the `BTD`.

9.1.1.10. Function `UDT_power()`. Calculates exact power given sample size or sample size given power, using analytical methods for the frequentist test of deficit for a specified case scores, means and standard deviations for the control sample. The means and standard deviations defaults to 0 and 1 respecitvely, so if no other values are given, the case scores are interpreted as deviations from the mean in standard deviations. The returned value will approximate the power for the given parameters.

Usage:

```
UDT_power(case_a, case_b, mean_a = 0, mean_b = 0, sd_a = 1,
sd_b = 1, r_ab = 0.5, sample_size = NULL, power = NULL, alternative
= c("two.sided", "one.sided"), alpha = 0.05, spec = 0.005)
```

Argument descriptions:

- `case_a`: A single value from the expected case observation on task A.
- `case_b`: A single value from the expected case observation on task B.
- `mean_a`: The expected mean from the control sample on task A. Defaults to 0.
- `mean_b`: The expected mean from the control sample on task B. Defaults to 0.
- `sd_a`: The expected standard deviation from the control sample on task A. Defaults to 1.
- `sd_b`: The expected standard deviation from the control sample on task B. Defaults to 1.
- `r_ab`: The expected correlation between the tasks. Defaults to 0.5
- `sample_size`: The size of the control sample, vary this parameter to see how the sample size affects power. One of sample size or power must be specified, not both.
- `power`: A single value between 0 and 1 specifying desired power for calculating necessary sample size. One of sample size or power must be specified, not both.
- `alternative`: The alternative hypothesis. A string of either "two.sided" (default) or "one.sided".
- `alpha`: The specified Type I error rate. This can also be varied, with effects on power. Defaults to 0.05.

- `spec`: A single value between 0 and 1. If desired power is given as input the function will utilise a search algorithm to find the sample size needed to reach the desired power. However, if the power specified is greater than what is actually possible to achieve the algorithm could search forever. Hence, when power does not increase substantially for any additional participant in the sample, the algorithm stops. By default the algorithm stops when power does not increase more than 0.5% for any additional participant, but by varying spec, this specificity can be changed.

9.1.1.11. Function `RSDT_power()`. Calculates approximate power, given sample size, using Monte Carlo simulation, for specified case scores, meana and standard deviations for the control sample. The means and standard deviations defaults to 0 and 1 respectively, so if no other values are given the case scores are interpreted as deviations from the mean in standard deviations. Hence, the effect size of the dissociation (zdcc) would in that case be the difference between the two case scores. Returns a single value approximating the power of the test for the given parameters.

Usage:

```
RSDT_power(case_a, case_b, mean_a, mean_b, sd_a, sd_b, r_ab,
sample_size, alternative = c("two.sided", "one.sided"), alpha
= 0.05, nsim = 10000)
```

Argument descriptions:

- `case_a`: A single value from the expected case observation on task A.
- `case_b`: A single value from the expected case observation on task B.
- `mean_a`: The expected mean from the control sample on task A. Defaults to 0.
- `mean_b`: The expected mean from the control sample on task B. Defaults to 0.
- `sd_a`: The expected standard deviation from the control sample on task A. Defaults to 1.
- `sd_b`: The expected standard deviation from the control sample on task B. Defaults to 1.
- `r_ab`: The expected correlation between the tasks. Defaults to 0.5
- `sample_size`: The size of the control sample, vary this parameter to see how the sample size affects power.

- `alternative`: The alternative hypothesis. A string of either “two.sided” (default) or “one.sided”.
- `alpha`: The specified Type I error rate. This can also be varied, with effects on power. Defaults to 0.05.
- `nsim`: The number of simulations to run. Higher number gives better accuracy, but low numbers such as 10000 or even 1000 are usually sufficient for the purposes of this calculator.

9.1.1.12. Function `BSDT_power()`. Calculates approximate power, given sample size, using Monte Carlo simulation, for specified case scores, means and standard deviations for the control sample. The means and standard deviations defaults to 0 and 1 respectively, so if no other values are given the case scores are interpreted as deviations from the mean in standard deviations. Hence, the effect size of the dissociation (zdcc) would in that case be the difference between the two case scores. Is computationally heavy and might therefore take a few seconds.

Usage:

```
BSDT_power(case_a, case_b, mean_a, mean_b, sd_a, sd_b, r_ab,
sample_size, alternative = c("two.sided", "one.sided"), alpha
= 0.05, nsim = 1000, iter = 1000)
```

Argument descriptions:

- `case_a`: A single value from the expected case observation on task A.
- `case_b`: A single value from the expected case observation on task B.
- `mean_a`: The expected mean from the control sample on task A. Defaults to 0.
- `mean_b`: The expected mean from the control sample on task B. Defaults to 0.
- `sd_a`: The expected standard deviation from the control sample on task A. Defaults to 1.
- `sd_b`: The expected standard deviation from the control sample on task B. Defaults to 1.
- `r_ab`: The expected correlation between the tasks. Defaults to 0.5
- `sample_size`: The size of the control sample, vary this parameter to see how the sample size affects power.
- `alternative`: The alternative hypothesis. A string of either “two.sided” (default) or “one.sided”.

- `alpha`: The specified Type I error rate. This can also be varied, with effects on power. Defaults to 0.05.
- `nsim`: The number of simulations to run. Higher number gives better accuracy, but low numbers such as 10000 or even 1000 are usually sufficient for the purposes of this calculator.
- `nsim`: The number of simulations to run. Higher number gives better accuracy, but low numbers such as 10000 or even 1000 are usually sufficient for the purposes of this calculator. Defaults to 1000 due to the computationally intense `BSDT`.
- `iter`: The number of simulations used by `BSDT`. Defaults to 1000.

10. Appendix B

Table 6

	Dependent variable:
	True discrepancy
Found_disc1	−3.169 (1.653)
Found_disc2	0.982* (0.459)
Found_disc3	−0.072 (0.041)
Sample_size1	−2.319*** (0.540)
Sample_size2	0.059 (0.039)
Sample_size3	−0.001 (0.001)
Correlation1	11.575 (12.060)
Correlation2	−41.300 (23.731)
Correlation3	24.406 (14.151)
Found_disc1:Sample_size1	1.362** (0.455)
Found_disc2:Sample_size1	−0.255* (0.125)
Found_disc3:Sample_size1	0.016 (0.011)
Found_disc1:Sample_size2	−0.028 (0.033)
Found_disc2:Sample_size2	0.004 (0.009)
Found_disc3:Sample_size2	−0.0001 (0.001)
Found_disc1:Sample_size3	0.0005 (0.001)
Found_disc2:Sample_size3	−0.0001 (0.0002)
Found_disc3:Sample_size3	0.00000 (0.00002)
Found_disc1:Correlation1	−11.435 (10.737)
Found_disc2:Correlation1	3.396 (3.122)
Found_disc3:Correlation1	−0.304 (0.295)
Found_disc1:Correlation2	32.778 (22.027)
Found_disc2:Correlation2	−7.681 (6.692)
Found_disc3:Correlation2	0.551 (0.659)
Found_disc1:Correlation3	−16.084 (13.614)
Found_disc2:Correlation3	2.920 (4.284)
Found_disc3:Correlation3	−0.138 (0.435)
Sample_size1:Correlation1	−3.045 (3.223)
Sample_size2:Correlation1	0.185 (0.229)
Sample_size3:Correlation1	−0.002 (0.005)
Sample_size1:Correlation2	10.619 (6.156)
Sample_size2:Correlation2	−0.410 (0.431)
Sample_size3:Correlation2	0.004 (0.009)
Sample_size1:Correlation3	−5.137 (3.602)
Sample_size2:Correlation3	0.159 (0.249)
Sample_size3:Correlation3	−0.001 (0.005)
Found_disc1:Sample_size1:Correlation1	3.281 (2.813)
Found_disc2:Sample_size1:Correlation1	−0.936 (0.806)
Found_disc3:Sample_size1:Correlation1	0.080 (0.075)
Found_disc1:Sample_size2:Correlation1	−0.162 (0.198)
Found_disc2:Sample_size2:Correlation1	0.041 (0.056)
Found_disc3:Sample_size2:Correlation1	−0.003 (0.005)
Found_disc1:Sample_size3:Correlation1	0.001 (0.004)
Found_disc2:Sample_size3:Correlation1	−0.0003 (0.001)
Found_disc3:Sample_size3:Correlation1	0.00002 (0.0001)
Found_disc1:Sample_size1:Correlation2	−7.263 (5.595)
Found_disc2:Sample_size1:Correlation2	1.466 (1.676)
Found_disc3:Sample_size1:Correlation2	−0.088 (0.164)
Found_disc1:Sample_size2:Correlation2	0.240 (0.387)
Found_disc2:Sample_size2:Correlation2	−0.038 (0.115)
Found_disc3:Sample_size2:Correlation2	0.001 (0.011)
Found_disc1:Sample_size3:Correlation2	−0.002 (0.008)
Found_disc2:Sample_size3:Correlation2	0.0001 (0.002)
Found_disc3:Sample_size3:Correlation2	0.00002 (0.0002)
Found_disc1:Sample_size1:Correlation3	2.509 (3.393)
Found_disc2:Sample_size1:Correlation3	−0.252 (1.054)
Found_disc3:Sample_size1:Correlation3	−0.008 (0.106)
Found_disc1:Sample_size2:Correlation3	−0.046 (0.232)
Found_disc2:Sample_size2:Correlation3	−0.007 (0.072)
Found_disc3:Sample_size2:Correlation3	0.002 (0.007)
Found_disc1:Sample_size3:Correlation3	−0.0002 (0.005)
Found_disc2:Sample_size3:Correlation3	0.0003 (0.001)
Found_disc3:Sample_size3:Correlation3	−0.00004 (0.0001)
Constant	4.617* (1.934)
Observations	3,640
R^2	0.998
Adjusted R^2	0.998
Residual Std. Error	0.065 (df = 3576)
F Statistic	30,603.150*** (df = 63; 3576)
Note:	*p<0.05; **p<0.01; ***p<0.001 Standard errors in parentheses