Psychiatry Research 270 (2018) 801–806

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Enhancing the understanding of clinically meaningful results: A clinical research perspective

Anders Nordahl-Hansen[a,1,*], Roald A. Øien[b,c,1,2], Fred Volkmar[c,3], Frederick Shic[d,e,4], Domenic V. Cicchetti[c,f,5]

[a] Faculty of Education, Østfold University College, B R A Veien 4, P.O. 700, Halden, Norway
[b] Department of Education, UiT – The Arctic University of Norway, Tromsø, Norway
[c] Child Study Center, Yale University School of Medicine, New Haven, USA
[d] Pediatrics, University of Washington, Seattle, WA, USA
[e] Seattle Children's Research Institute, Seattle, WA, USA
[f] Department of Biometry, Yale Home Office, North Branford, USA

ARTICLE INFO

Keywords:
Relative risk and risk ration
Clinical significance
P-value
Statistical hypothesis inference testing
Effect size

ABSTRACT

Published research often address aspects related to "statistical significance" but fail to address the clinical and practical importance and meaning of results. Our main objectives in this article are to investigate the merit of common measures of Effect Size in statistical research and to highlight the importance of the simple Relative Risk ratio. In this article we present data where we consider two widely utilized effect size measures (Cohen's d and Pearson's r) in relations to relative risk. We conclude that probability analyses of risk surpass the most commonly used statistical approach used in clinical trials today and should thus be the preferred compared to the misuse and misunderstanding of reporting for instance *p*-values alone.

1. Introduction

Research literature evaluating effects of treatment should guide clinical practice. It is then key that the summarization of data from treatment studies is presented in a manner that can be readily appreciated by doctors, clinicians and practitioners (Cook and Sackett, 1995). Due to a higher availability of research today than before, health consumers are now readers of evaluative science such as research reports (Eysenbach, 2000). Thus, it is as important as ever that research is presented in a clear and concise fashion and that inferences are in line with what the results of the study actually indicate. However, practicing and preaching are not always synonymous, and the misuse of the *p-value* continues to distort inferences in almost any discipline including psychiatry.

The objective of this article is three-fold: First to address the long-standing, still ongoing discussion regarding the *p*-value; second, to discuss the relative merits of two of the most widely utilized measures of Effect Size (ES) in bio-behavioral research; and third, to highlight the importance of the simple Relative Risk ratio (RR).

The most frequently used statistical approach in clinical trials is "Null Hypothesis Significance Testing" (NHST) and the highly debated *p*-value. When undertaking a significance test, the researcher selects a significance level called alpha. The significance level is typically chosen from a consensus in that particular field of research (e.g., alpha = 0.05). The *p*-value, under a specified statistical model, represents the probability that an observed alpha value or one more extreme has occurred. The usage of *p*-values continues to be debated, and has been so since well before the Second World War. In an attempt to clarify and elucidate the problems with NHST and the *p*-value, the American Statistical Association (ASA) issued a statement on its use in

* Corresponding author.
E-mail addresses: anders.nordahl-hansen@hiof.no (A. Nordahl-Hansen), roald.a.oien@uit.no (R.A. Øien), fred.volkmar@yale.edu (F. Volkmar), fshic@uw.edu (F. Shic), dom.cicchetti@yale.edu (D.V. Cicchetti).
[1] (Note: 1st and 2nd author are sharing first authorship and ordering of the first two authors was decided alphabetically, i.e., Nordahl-Hansen before Øien.
[2] Associate Professor Roald A. Øien, Ph.D.: Department of Education, UiT – The Arctic, University of Norway, 9037 Tromsø, Norway.
[3] Professor Fred R Volkmar, M.D.: Child Study Center, Yale University School of Medicine, 230 South Frontage Road, P.O. Box 207900 New Haven, CT USA.
[4] Associate Professor Frederick Shic: University of Washington, 2001 8th Ave Suite #400, Seattle WA USA 98121.
[5] Professor Dom Cicchetti, Ph.D.: Child Study Center and, Department of Biometry, Yale, University School of Medicine, New Haven, CT 06520, Yale Home office, Box 317, North Branford, CT 06471.

https://doi.org/10.1016/j.psychres.2018.10.069
Received 9 February 2018; Received in revised form 25 October 2018; Accepted 25 October 2018
Available online 31 October 2018

research in which it was highlighted that the *p*-value "does not provide a good measure of evidence regarding a model or hypothesis" (Wasserstein and Lazar, 2016). Further, the statement underscored that statistical significance does not measure size of effects or importance of results. We adhere to the ASA statement when arguing that the use of *p*-values alone is simply not enough when drawing inferences about, for instance, effects of treatment trials and must be supplemented, or replaced by other approaches. There are several approaches available other than *p*-value statistics and although these approaches typically rely on more assumptions, they address sizes of effect more directly or whether a hypothesis indeed is warranted, both aspects of which are not addressed by Null Hypothesis Significance Testing.

Enhancing the knowledge of evidence based practices (EBP) has been a key endeavor within the field of medicine and psychology for quite some time (Cicchetti, 2011; Cutspec, 2004). However, the statistics used to evaluate and interpret results from randomized clinical trials (RCT) should be chosen on the basis of which best inform clinical practice. Exaggeration of what can be inferred from *p*-values through NHST remains an issue (Borenstein, 1998; Cohen, 1995; Stern, 2016). Even though the *p*-value and its usage have been heavily questioned since first introduced by Ronald Fisher in the 1920 s, it is, as noted, still the most common statistical method used to address treatment effects (Nuijten et al., 2016). A main critique from the opponents of statistical significance testing is that the *p*-value can be viewed as a fallacy or a "conceptual error that has profoundly influenced how we think about the process of science and the nature of scientific truth" (Goodman, 1999). Others have addressed similar concerns (Burton et al., 1998; Kraemer et al., 2011). Simmons and colleagues note that the costliest error relates to the incorrect rejection of the null hypothesis (Simmons et al., 2011). This error concerns *false positive* findings, or falsely claiming the reliability or accuracy of an evaluative method. Such erroneous findings die hard as possible replications with null findings can be difficult to publish (Simmons et al., 2011). Readers of journals may misinterpret results reported as statistically significant to be "positive" evidence for treatment efficacy. In fact, approvals of new drugs by regulatory authorities are typically based on such positive results indicating little else than slight group differences (Ocana and Tannock, 2010). Borenstein (1998) goes further, citing an example of several studies all showing positive results for a particular drug. However, some of the studies produced statistically significant results while others did not. A more careful scrutiny of the data indicated that the studies producing statistically significant results derived from much larger sample sizes than those that failed to reach statistical significance. Rather than recognizing this as a small N phenomenon requiring further investigation with a larger sample size, the drug was not considered further (Borenstein, 1998).

Due to the role that the *p*-value has attained during its history, it continues to confuse. This then goes on to impact the practical field as readers misinterpret the actual inferences that can be drawn from studies. These issues obscure actual effects, or the lack thereof, in intervention trials that fail to go beyond merely testing whether the groups were different or not. Therefore, there is a need for research reporting not only statistical significance but also estimates of whether change following the intervention was clinically significant and meaningful. The ultimate value of a test of a treatment trial should not rest on statistical significance, but rather be based on results that are both statistically and clinically meaningful (Cicchetti et al., 2011), while also remaining easy to interpret for clinicians with sparse statistical knowledge.

1.1. What constitutes change/meaningful change/clinical significance?

Some of the reasons for the continuing misinterpretations of *p*-values are due to the mere coining of the term "significance", which implies that something is of a magnitude or of importance. Consistent with this fallacious reasoning is the all too frequent idea that a *p*-value of 0.01 is more clinically meaningful than one of 0.05 (Borenstein, 1998). This is a bona fide example of a too prevalent confusion of statistical significance with clinical or substantive significance. Nevertheless, the only feature of significant magnitude in many studies, despite "reaching" statistical significance, is the sample size. This echoes the criticism that under normal assumptions, any consistent difference between means can be made statistically significant at $p = 0.05$ with a large enough sample size (Bakan, 1966; Cicchetti et al., 2011) and underscores the arbitrariness of the norm value of 0.05 which was elegantly reflected upon by Rosnow and Rosenthal (1989), when stating that "surely, God loves the 0.06 nearly as much as the 0.05″. There is of course the chance she or he does not love any of them. A *p*-value below the 0.05 threshold does not establish that a hypothesis is true just as a *p*-value not reaching the same threshold proves the null hypothesis. Further, as noted by Cohen (1995) and Borenstein (1998), at the end of the day it is not whether the effect of a treatment is zero or not that is important, the answer one gets from a significance test, but the magnitude of the effect and thus including the element of interest, namely clinical significance (Borenstein, 1997).

The use of the term clinical significance may be interpreted in different ways, and may depend on the condition at hand. For instance, a clinically significant change for a once fractured hand can be the regaining of full functionality. Although many psychiatric diagnoses are not necessarily life-long, many are. Thus, if the term clinical significance is to be used, it should also be operationalized as something other than a dichotomous diagnosis versus recovery divide. We now turn to some alternatives that in addition to providing statistical estimates of effect are more intuitive measures of effect.

1.2. Clinical significance in treatment trials

The concept of clinical significance is closely intertwined with effect size estimation, which in turn led to power analysis estimation (Cohen, 1977). A definition of effect size is provided by Kelley and Preacher stating that an effect size is "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (2012, p. 140). Making power analysis possible was the core intent for Cohen, and when he presented the cutoffs considering what constitutes small, medium and large effect sizes (ES) he strongly cautioned the many dangers of using such rules of thumb (Cohen, 1977). Following reports on *p*-values many researchers rely on the use of effect sizes such as Cohen's *d* or correlation coefficients (*r*) and statistics of this genre are also at the heart of the summarization of results in meta-analyses (McGrath and Meyer, 2006).

1.3. Two widely applied measures of ES in bio-behavioral research: r and d

Because *r* and *d* appear, arguably, to be the two most widely applied measures of ES, it will be instructive to briefly define each and then compare them across a wide range of values. *d* refers to the percentage of overlap between two mean values, with a range of 2.3 to 100%; *r* is the familiar Pearson Product Moment Correlation Coefficient (PPMC). As noted by Cohen (1988), *d* is defined as:

$$d = \frac{m_A - m_B}{\mathrm{SD}} \tag{1}$$

where:m_A = the first of two Mean or Average valuesm_B = the second of the two Mean values andSD = the standard deviation of the difference between the pair of Means

The formula for converting *d* to *r* is given by Cohen (1988) as:

$$r = d/\sqrt{(d^2 + \frac{(n_1 + n_2)^2}{n_1 n_2})} \tag{2}$$

where:*d* is defined as in Formula [1] and n_1 and n_2 are the population sizes of groups 1 and 2, respectively. For $n_1 = n_2$ the above reduces to:

Table 1
Cohen's criteria for *d*, *r* and % overlap.

d[a]	Percent overlap[c]	*r*[b]
0.0	100	0.000
0.1	92.3	0.050
0.2	85.3	0.100
0.3	78.7	0.148
0.4	72.6	0.196
0.5	66.6	0.243
0.6	61.8	0.287
0.7	57.0	0.330
0.8	52.6	0.371
0.9	48.4	0.410
1.0	44.6	0.447
1.1	41.1	0.482
1.2	37.8	0.514
1.3	34.7	0.545
1.4	31.9	0.573
1.5	29.3	0.600
1.6	26.9	0.625
1.7	24.6	0.648
1.8	22.6	0.669
1.9	20.6	0.689
2.0	18.9	0.707
2.2	15.7	0.740
2.4	13.0	0.768
2.6	10.7	0.793
2.8	8.8	0.814
3.0	7.2	0.832
3.2	5.8	0.848
3.4	4.7	0.862
3.6	3.7	0.874
3.8	3.0	0.885
4.0	2.3	0.894

[a] The Cohen (1988) criteria for *d* are: < 0.2 = No Effect; 0.2 = Small; 0.5 = Large; and ≥ 0.8 = Large.

[b] The Cohen (1988) for *r* are: < 0.10 = Trivial; 0.10–0.29 = Small; 0.30–0.49 = Medium; and ≥ .50 = Large. These were revised to read as: < 0.10 = Trivia l; 0.10–0.29 = Small; 0.30–0.49 = Medium; 0.50–0.69 = Large; and ≥ 0.70 = Very Large (Cicchetti, 2008).

[c] Percent overlap is the extent to which two populations of research interest are "superimposed" upon each other (Cohen, 1988); Suppose the Boston Naming Test was administered to patients with Parkinson's Disease (PD) and healthy controls; a *d* of 3 would mean that only about 7% of the PD patients would obtain scores obtained by healthy controls (Zakzanis, 2001).

$$r = d/\sqrt{(d^2 + 4)} \quad (3)$$

We note that PPMC and correlation coefficients in general are typically used to assess the strength of bivariate relationships between random variables (Kraemer et al., 1999), and as such the use of PPMC in the context of group difference effect sizes is a special case (Cohen, 1988).

The aforementioned comparison of *d* and *r* is given in Table 1 below. The ES values for both statistics are given in the Footnote below the Table.

The data in Table 1, derive from two sources: Cohen (1988), and Zakzanis (2001), as combined into a single Table.

Since *d* and *r* are valid measures of ES, a question is which one should the clinical research scientist apply? McGrath and Meyer (2006) address this important issue in an article appropriately entitled "When effect sizes disagree: The case of *r* and *d*." In making the comparison, the two authors recommend reporting both measures of ES:

"Doing so has several benefits, including simplicity and the fact that it does not require adjusting interpretive benchmarks. An additional benefit is that when base rates diverge, reporting both *r* and *d* will juxtapose the seemingly discrepant inferences about magnitude of effect and will highlight the importance known for some time of deciding whether the natural base rates should be given credence or be discounted" (McGrath and Meyer, 2006).

1.4. Assessing inter-examiner agreement and correlations for binary data deriving from a 2 × 2 table

The concept of association is pertinent in the context of both examiner agreement and correlation. One recently published report assessed the reliability of various measures of agreement for bio-behavioral disorders in general, with a specific investigation of the reliability of the presence or absence of personality disorders. The results indicated that from both a probabilistic and clinical perspective, Cohen's kappa coefficient (Cohen, 1960) is to be preferred over competitors, such as Gwet's 2002 and 2008 AC1-coefficient (Cicchetti et al., 2017; Gwet, 2002; Gwet, 2008).

One example of correlations from binary data is the association between aspirin therapy (yes/no) compared to the probability of a heart attack or myocardial infarct, also defined as yes or no (Hennekens, 1988; Rosenthal et al., 1994). When analyzed using the familiar binary version of the standard correlation coefficient, *phi*, the result was a paltry 0.03, a trivial result by the criteria developed by Cohen (1988). However, when the Relative Risk (RR) statistic was applied to the same data, this resulted in a value of 1.82 favoring the aspirin therapy group. This translates to mean that the risk of suffering a heart attack was almost twice as high in the group not taking aspirin. One is forced to conclude that this is hardly a trivial result. In words, the RR can be defined as the ratio of two probabilities, each based upon a binary variable and is defined as "the probability of an event in the active treatment group divided by the probability of an event in the control group" (Cook and Sackett, 1995). In the aforementioned study, 189 persons of 11,034 not on aspirin therapy suffered a heart attack, as compared to 104 of the 11,037 who received aspirin therapy (Hennekens, 1988). Here the RR becomes:

$$RR = (189/11034) \div (104/11037) = 1.82$$

The aspirin therapy study introduces a very subtle additional problem, namely, the need to select an appropriate statistic to correctly evaluate treatment effects. While the standard correlation coefficient is a very valuable and even venerable statistic, it was not appropriate for assessing the effects of aspirin therapy, while RR was ideal.

1.5. From the odds ratio to relative risk (RR): a brief historical perspective

The Odds ratio (OR) as a measure of effect size, provides information about the power of an association and an outcome. The OR can be calculated using the ratio between two odds, and provides information about the odds that a specific outcome will occur when a specific exposure is present compared to the odds when the specific outcome is not present. Hence, the OR can be viewed as what are the odds that a group of interest will have a disease, relative to a comparison group: for example, is it a 50–50 bet or might it be 3 to 1 favoring the group of clinical interest. OR is the most widely used statistic in epidemiological research and is also often presented in treatment trials (Bland and Altman, 2000).

Cornfield (1951) developed the Odds Ratio. However, Fleiss et al., (2003) note that for Cornfield the OR was mainly a step along the way as it "provided a good approximation to another measure he proposed, the *relative risk*, also called the *rate ratio*." Fleiss and colleagues go on to mention *"Because of its great advantages over other measures of binary association, Edwards (1963) recommended that the odds ratio or derivatives of it such as the relative risk (RR) be used as the preferred statistic"* (Paik et al., 2003). Others have also noted the difficulties of using and reporting OR for clinical decision-making, favoring instead Relative Risk (Ferguson, 2009) and Numbers Needed to Treat (Kraemer and Blasey, 2015). In addition, OR and relative risk are similar only in specific circumstances, e.g. when the frequencies of outcomes are small in cohort studies (Viera, 2008). Notwithstanding, the Relative risk or Risk Ratio (RR either way) is often confused with OR, thus the importance of the distinction between the two is shortly addressed. In

research, mistakes frequently occur when OR is reported as RR. While the OR is the ratio of two odds, the RR is the ratio of two probabilities or risks. Hence, RR is the risk of an event happening in one group compared to the risk of that same event happening in another group. From a clinical vantage point deploying RR asks the question of how a group of interest compare to a comparison group in terms of their relative risks of a particular disease. Being an index using the understanding of ratios RR is readily intuitive for clinicians and practitioners.

2. Method and results

In the next section, we shall make a direct comparison between values of the three measures we have just discussed, namely: Kappa, the Phi Coefficient and the RR coefficient when each is applied to a theoretical data set. The method to be defined is Hypothetics. It will be recalled that the Phi Coefficient is the conceptual equivalent of the standard correlation coefficient.

2.1. How the three measures Kappa, phi and RR inter-correlate

Hypothetics, before it was defined formally, was first applied in an earlier study by one of the authors (Cicchetti, 1988). Confirmatory follow-up studies were published two years later. That earlier investigation (Cicchetti, 1988) and two later studies (Cicchetti and Feinstein, 1990; Feinstein and Cicchetti, 1990) explained and resolved the seeming paradoxes that occur when the level of observed agreement is high (say, 80% or higher), but Kappa is low and unacceptable [i.e., below 0.40 by the criteria of Cicchetti, (1994) and Cicchetti and Sparrow (1981)].

The concept of Hypothetics can be defined as the study of the hypothetical results that would occur if a scientific researcher were able to vary or hold constant a number of critical binary variables. With respect to the examiner reliability case, the authors could vary the following: The total number of cases diagnosed as positive and negative by each of the two examiners [e.g., 50:50 for examiner 1 and 45:55 for Examiner 2 (these are also designated as the Rater Marginals); the range of agreement levels on positive cases (say 40% to 80%)] ; the agreement levels on the number of negative cases (say, 10 to 20); the numbers of cases for which the first examiner diagnosed positive, while the second diagnosed the same cases as negative; the reverse phenomenon, whereby the first examiner defined 10 cases as, say, negative that the second examiner diagnosed as positive; the resulting values of the RR, Kappa and Phi coefficient; and, say the probability levels of each hypothetical case.

Following this line of reasoning, the following hypothetical Table was devised:

Table 2

2.2. Correlates of RR

The correlation between Kappa and RR is 0.62; and the correlation between Phi and RR is 0.59; these represent Large Effect Sizes (ES) by the criteria of Cohen (1988) and by the expanded criteria of Cicchetti (2008). The Cohen (1988) ES criteria are: < 0.10 = Trivia l; 0.10 = S mall; 0.30 = Medium; and 0.50 = Large; the expanded criteria are: < 0.10 = Trivia l; 0.10–0.29 = Small; 0.30–0.49 = Medium; 0.50–0.69 = Large; and ≥ 0.70 = Very Large. Applying the clinical criteria suggested by Cicchetti (1994) and earlier by Cicchetti and Sparrow (1981), Kappa values below 0.40 can be considered Poor; 0.40 to 0.59 is Fair; 0.60–0.74 represents Good; and Kappa's of ≥ 0.75 can be considered Excellent. It is also of interest that RR values below 1.50 are not statistically significant while those of 1.5 and higher are statistically significant at or beyond the time-honored probability (p) level of 0.05. It is probably fair to consider an RR of ≥ 1.5 to be clinically significant. With respect to the Kappa values, every value that is clinically significant is also statistically significant; the reverse is not true; thus, there are three Kappa's that are well below the desideratum of 0.40 (specifically 0.218, 0.289 and 0.375). This is a welcome result and mirrors the findings published recently by two of the authors and a third colleague (Cicchetti et al., 2017).

3. Discussion

The American Statistical Association (2016) and other biostatisticians (see e.g. Borenstein, 1998; Cohen, 1995; Kraemer 2017) highlight the problem of using erroneous inferential statistics in medical research, and in particular statistical significance testing. Although the peer review system decreases the chance of statistical errors, statistical significance testing with p-values has become a systemic error where reviewers may even demand the introduction of statistical errors (Kraemer, 2017). Some have suggested that a solution to the p-value problem could be to use an even higher threshold (e.g. $p < 01.$ or to 0.005). However, as noted by Kraemer (2017) and Ioannidis (2018) this does not address the clinical significance nor add strength to the inferences of the study, but rather address aspects related to the design and execution of the study (e.g., sample size and the reliability of the measures used).

In this paper, we have highlighted the issue of faulty inferences due to misinterpretations of p-value statistics and propose a solution that is a) more correct to use in most instances, and gives the information one typically wants, and b) appears more intuitively easy to understand. Presentation of results is not necessarily an easy task as there are degrees of uncertainty in addressing effects of treatment; and wording such as probability, risk and chance are subject to misinterpretation (Kong et al., 1986). However, probability analyses of risk surpass the most commonly used statistical approach used in clinical trials today and should thus be preferred compared to p-values alone. We note that there exist multiple alternate formulations, for example one could consider as effect sizes area under the curve AUC = P(T1 > T2) + ½P (T1 = T2), success rate differences (2 AUC −1), or number needed to treat (1/(2AUC-1), all of which (,[6]see Kraemer and Frank, 2010, Kraemer and Kupfer, 2006). These formulations, while reflecting different intuitive perspectives, are often highly (even perfectly) correlated with one another. More generally, one could consider tradeoffs regarding the clinical importance of weight of false positives and negatives on a case-by-case basis.

Yet, the use of p-values remains as the most frequent choice for many researchers. However, there is a general trend in the broader field of psychiatry that Bayesian statistics are rising in popularity (van de Schoot et al., 2017). Although there are various approaches within Bayesian statistics, and disagreement on what types of models are best in different circumstances, the above-mentioned trend is welcoming. Nevertheless, the need for sober inferences is as important when using Bayesian statistics as well as the need for communicating the results to the broad community of readers of research today. The development of statistical computer software (e.g., R and JASP) allow for running analyses using both a classical- (frequentist) as well as Bayesian frameworks. Due to its intuitive nature, Relative Risk statistics pose as a viable candidate for use in clinical trials within the field of psychiatry.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[6] Where T1 and T2 refer to two different treatments, P(T1 > T2) + ½P (T1 = T2) the probability that a patient drawn from T1 would have a preferable or equal outcome compared to T2.

Table 2
Rater margins for Kappa, RR and phi.

Rater marginals:	Overall agreement:	(+ +)	(- -)	(+ -)	(- +)	Kappa[1]	Phi[2]	RR probability[3]- *p*	
50:50 vs 80:20	50	40	10	40	10	0.000	0.000	1.000	NS
55:45 vs 80–20	55	45	10	35	10	0.043	0.050	1.125	NS
60:40 vs 80:20	60	50	10	30	10	0.091	0.102	1.250	NS
65:35 vs 80:20	65	55	10	25	10	0.146	0.157	1.375	NS
70:30 vs 80:20	70	60	10	20	10	0.211	0.218	1.500	0.05
75:25 vs 80:20	75	65	10	15	10	0.286	0.289	1.625	0.008
80:20 vs 80:20	80	70	10	10	10	0.375	0.375	1.750	0.001
85:15 vs 80:20	85	75	10	5	10	0.483	0.490	1.875	<0.001
86:14 vs 80:20	86	76	10	4	10	0.507	0.519	1.900	<0.001
87:13 vs 80:20	87	77	10	3	10	0.532	0.550	1.925	<0.001
88:12 vs 80:20	88	78	10	2	10	0.559	0.585	1.950	<0.001
89:11 vs 80:20	89	79	10	1	10	0.586	0.623	1.975	<0.001
90:10 vs 80:20	90	80	10	0	10	0.615	0.667	2.000	<0.001
89:11 vs 80:20	91	80	11	0	9	0.662	0.703	2.222	<0.001
88:12 vs 80:20	92	80	12	0	8	0.706	0.739	2.500	<0.001
87:13 vs 80:20	93	80	13	0	7	0.748	0.773	2.587	<0.001
86:14 vs 80:20	94	80	14	0	6	0.789	0.807	3.333	<0.001
85:15 vs 80:20	95	80	15	0	5	0.828	0.840	4.000	<0.001
84:16 vs 80:20	96	80	16	0	4	0.865	0.873	5.000	<0.001
83:17 vs 80:20	97	80	17	0	3	0.901	0.905	6.667	<0.001
82:18 vs 80:20	98	80	18	0	2	0.935	0.939	10.000	<0.001
81:19 vs 80:20	99	80	19	0	1	0.968	0.942	20.000	<0.001

Declaration of interests

None of the authors declare any conflict of interests.

Ethical approval

No ethical approval was sought as data do not include human or animal participants.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2018.10.069.

References

Bakan, D., 1966. The test of significance in psychological research. Psychol. Bull. 66, 423.
Bland, J.M., Altman, D.G., 2000. The odds ratio. BMJ 320, 1468.
Borenstein, M., 1997. Hypothesis testing and effect size estimation in clinical trials. Ann. Allergy Asthma Immunol. 78, 515–1216.
Borenstein, M., 1998. The shift from significance testing to effect size estimation. Res. Methods Compr. Clin. Psychol. 3, 319–349.
Burton, P.R., Gurrin, L.C., Campbell, M.J., 1998. Clinical significance not statistical significance: a simple Bayesian alternative to *p* values. J. Epidemiol. Community Health 52, 318–323.
Cicchetti, D.V., 1988. When diagnostic agreement is high, but reliability is low: some paradoxes occurring in joint independent neuropsychology assessments. J. Clin. Exp. Neuropsychol. 10, 605–622.
Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6, 284.
Cicchetti, D.V., 2008. From Bayes to the just noticeable difference to effect sizes: a note to understanding the clinical and statistical significance of oenologic research findings. JWE 3, 185–193.
Cicchetti, D.V., 2011. On the reliability and accuracy of the evaluative method for identifying evidence-based practices in autism evidence-based practices and treatments for children with autism. In: Reichow, B., Doehring, P., Cicchetti, D., Volkmar, F. (Eds.), Evidence-Based Practices and Treatments for Children with Autism. Springer, Boston, pp. 41–51.
Cicchetti, D.V., Feinstein, A.R., 1990. High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 43, 551–558.
Cicchetti, D.V., Klin, A., Volkmar, F.R., 2017. Assessing binary diagnoses of bio-behavioral disorders: the clinical relevance of Cohen's Kappa. J. Nerv. Ment. Dis. 205, 58–65.
Cicchetti, D.V., Koenig, K., Klin, A., Volkmar, F.R., Paul, R., Sparrow, S.A., 2011. From Bayes through marginal utility to effect sizes: a guide to understanding the clinical and statistical significance of the results of autism research findings. J. Autism Dev. Disord. 41, 168–174.
Cicchetti, D.V., Sparrow, S.A., 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am. J. Ment. Defic. 86, 127–137.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.
Cohen, J., 1977. Statistical Power Analysis for the Behavioral Sciences, revised ed. Academic Press, New York.
Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Lawrence Earlbaum Associates, Hilsdale. NJ, pp. 2.
Cohen, J., 1995. The earth is round (p<. 05): rejoinder. Am. Psychol. 49, 997–1003.
Cook, R.J., Sackett, D.L., 1995. The number needed to treat: a clinically useful measure of treatment effect. BMJ 310, 452.
Cornfield, J., 1951. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. J. Natl. Cancer Inst. 11, 1269–1275.
Cutspec, P.A., 2004. Bridging the research-to-practice gap: Evidence-based education. Centerscope: evidence-based approaches to early childhood development 2 (2), 1–8.
Eyesenbach, G., 2000. Recent advances: consumer health informatics. BMJ 320, 1713.
Feinstein, A.R., Cicchetti, D.V., 1990. High agreement but low kappa: I. The problems of two paradoxes. J. Clin. Epidemiol. 43, 543–549.
Ferguson, C.J., 2009. An effect size primer: a guide for clinicians and researchers. Prof. Psychol. Res. Pract. 40 (5), 532–538. https://doi.org/10.1037/a0015808.
Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical Methods for Rates and Proportions 203. J. Wiley-Interscience, Hoboken, N.J., pp. 151.
Goodman, S.N., 1999. Toward evidence-based medical statistics. 1: the p value fallacy. Ann. Intern. Med. 130, 995–1004.
Gwet, K.L., 2002. Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. Stat. Methods Inter Rater Reliab. Assess. Ser. 2, 1–9.
Gwet, K.L., 2008. Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol. 61, 29–48.
Hennekens, C.H., 1988. Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. N. Engl. J. Med. 318, 262–264.
Ioannidis, J.P., 2018. The proposal to lower *p* value thresholds to. 005. JAMA 319, 1429–1430.
Kelley, K., Preacher, K.J., 2012. On effect size. Psychol. Methods 17, 137–152.
Kong, A., Barnett, G.O., Mosteller, F., Youtz, C., 1986. How medical professionals evaluate expressions of probability. N. Engl. J. Med. 315, 740–744.
Kraemer, H.C., 2017. Evidence-based medicine in eating disorders research: the problem of "confetti *p* values. Int. J. Eat. Disord. 50, 307–311.
Kraemer, H.C., Frank, E., 2010. Evaluation of comparative treatment trials: assessing the clinical benefits and risks for patients, rather than statistical effects on measures. JAMA 304, 1–2.
Kraemer, H.C., Blasey, C., 2015. How Many Subjects? Statistical Power Analysis in Research. Sage Publications.
Kraemer, H.C., Frank, E., Kupfer, D.J., 2011. How to assess the clinical impact of treatments on patients, rather than the statistical impact of treatments on measures. Int. J. Methods psychiatr. Res. 20, 63–72.
Kraemer, H.C., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S., Kupfer, D.J., 1999. Measuring the potency of risk factors for clinical or policy significance. Psychol. Methods 4, 257–271.
Kraemer, H.C., Kupfer, D.J., 2006. Size of treatment effects and their importance to clinical research and practice. Biol. Psychiatry 59, 990–996.
McGrath, R.E., Meyer, G.J., 2006. When effect sizes disagree: the case of r and d. Psychol. Methods 11, 386.
Nuijten, M.B., Hartgerink, C.H., van Assen, M.A., Epskamp, S., Wicherts, J.M., 2016. The prevalence of statistical reporting errors in psychology (1985–2013). Behav. Res. Methods 48, 1205–1226.
Ocana, A., Tannock, I.F., 2010. When are "positive" clinical trials in oncology truly

positive? J. Natl. Cancer Inst. 103, 16–20.
Rosenthal, R., Cooper, H., Hedges, L., 1994. Parametric measures of effect size. In: Cooper, H., Hedges, L.V. (Eds.), The Handbook of Research Synthesis. Russel Sage Foundation, New York, pp. 231–244.
Rosnow, R.L., Rosenthal, R., 1989. Statistical procedures and the justification of knowledge in psychological science. Am. Psychol. 44, 1276.
Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1359–1366.
Stern, H.S., 2016. A test by any other name: *p* values, Bayes factors, and statistical inference. Multivar. Behav. Res. 51, 23–29.
Van De Schoot, R., Winter, S.D., Ryan, O., Zondervan-Zwijenburg, M., Depaoli, S., 2017. A systematic review of Bayesian articles in psychology: the last 25 years. Psychol. Methods 22, 217.
Viera, A.J., 2008. Odds ratios and risk ratios: what's the difference and why does it matter. South. Med. J. 101, 730–734.
Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on *p*-values: context, process, and purpose. Am. Stat. 70, 129–133.
Zakzanis, K.K., 2001. Statistics to tell the truth, the whole truth, and nothing but the truth: formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. Arch. Clin. Neuropsychol. 16, 653–667.