Theory & Psychology

Article

A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough?

Theory & Psychology
23(1) 98–122

DOI: 10.1177/0959354312436870
tap.sagepub.com
SAGE

Astrid Fritz, Thomas Scherndl, and Anton Kühberger
University of Salzburg

Abstract

Over-reliance on significance testing has been heavily criticized in psychology. Therefore the American Psychological Association recommended supplementing the *p* value with additional elements such as effect sizes, confidence intervals, and considering statistical power seriously. This article elaborates the conclusions that can be drawn when these measures accompany the *p* value. An analysis of over 30 summary papers (including over 6,000 articles) reveals that, if at all, only effect sizes are reported in addition to *p*'s (38%). Only every 10th article provides a confidence interval and statistical power is reported in only 3% of articles. An increase in reporting frequency of the supplements to *p*'s over time owing to stricter guidelines was found for effect sizes only. Given these practices, research faces a serious problem in the context of dichotomous statistical decision making: since significant results have a higher probability of being published (publication bias), effect sizes reported in articles may be seriously overestimated.

Keywords

confidence interval, effect size, NHST, publication bias, significance testing, statistical power

Introduction

Psychology is an empirical science. Hence knowledge must be based on observation. Generalizations about unknown *population* parameters are *based on sample* statistics in most cases. In the social and behavioral sciences these inferences from the sample to the population are the domain of statistical significance testing. Psychology has an especially long tradition in significance testing. According to Hubbard and Ryan (2000), the

Corresponding author:
Anton Kühberger, Department of Psychology, University of Salzburg, 5020 Salzburg, Austria.
Email: anton.kuehberger@sbg.ac.at

percentage of empirical articles published in American Psychological Association (APA) journals that use statistical tests increased from about 17% from 1911 to 1929 to over 90% in the 1970s and beyond. Gigerenzer and Murray (1987) even claimed an *inference revolution* had taken place in psychology between 1940 and 1955.

Despite the widespread use and long tradition of significance testing, there is an old and continuing controversy about it (e.g., Nickerson, 2000). One of the first criticisms—nearly as old as the method itself—was by Boring (1919), who questioned the transferability of mathematical significance into a scientific conclusion. Over the past 90 years, a variety of articles from other authors followed. Kline's (2004) excellent book discusses the most prominent among them. In addition, Thompson (1997–2001) compiled a list of 402 citations that question the use of the so-called *null hypothesis significance testing* (NHST), but unfortunately he stopped updating in 2001. The present paper is another one on that list, but it adds something special: namely it discusses what is wrong with recent suggestions that were intended to change what was wrong with the traditional method. Our goal is to discuss not the limitations of the traditional method, but the shortcomings that are introduced by recent suggestions for solving the problems of traditional NHST.

Contemporary NHST is a hybrid of Fisherian testing of a single null hypothesis (Fisher, 1925/1950, 1935/1951) and the Neyman–Pearson idea of a decision between two hypotheses, H_0 and H_1 (Neyman & Pearson, 1933), embodying these approaches as if they were a single, uncontroversial method of inductive inference (e.g., Anderson, Burnham, & Thompson, 2000; Gigerenzer, 1993; Gigerenzer, Krauss, & Vitouch, 2004; Gigerenzer & Murray, 1987). However, past and contemporary statisticians—and even the founders of the approaches themselves—disagree. We refer to Halpin and Stam (2006) for an elaborate comparison of the two approaches. Controversy over NHST, however, is not our topic (for reviews see, e.g., Balluerka, Gomez, & Hidalgo, 2005; Kline, 2004; Nickerson, 2000). Rather, this paper investigates additional measures beyond *p* values aiming at qualifying results obtained by NHST.

First we will outline three additional methods to NHST, namely power analysis, effect size, and confidence interval, since they were recommended by the APA Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999), and in the two latest editions of the *Publication Manual* (APA, 2001, 2010; Capraro & Capraro, 2003; Fidler, 2002; Finch, Thomason, & Cumming, 2001). For instance, the latest version of the *Publication Manual* notes:

> Historically, researchers in psychology have relied heavily on null hypothesis statistical significance testing (NHST) as a starting point for many (but not all) of its analytic approaches. APA stresses that NHST is but a starting point and that additional reporting elements such as effect size, confidence intervals, and extensive description are needed to convey the most complete meaning of the results. (APA, 2010, p. 33)

Since, as Budge and Katz (1995) pointed out, the Manual is "the single text which virtually every psychologist, of whatever sub-specialty, has contact with at some point in their career" (p. 218), we will focus on these three additions to NHST. Of course, other ideas have also been proposed as solutions to the NHST controversy. Among those are Bayesian methods (e.g., Edwards, Lindman, & Savage, 1963; Hays & Winkler, 1970; Kruschke,

2010; Wagenmakers, 2007), replication (e.g., Killeen, 2006; Thompson, 1993, 1996, 1999a), and meta-analysis (e.g., Cohn & Becker, 2003; Dalton & Dalton, 2008). Moreover, the formulation of better, or more precise, hypotheses (e.g., the good-enough principle; Serlin & Lapsley, 1985, 1993) has been suggested as an alternative solving at least some of the problems associated with NHST. The interested reader is referred to Denis (2003) and Harlow, Mulaik, and Steiger (1997) for a more comprehensive review on these ideas.

Power

> *When applying inferential statistics, take seriously the statistical power considerations associated with the tests of hypothesis …. In that regard, routinely provide evidence that the study has sufficient power to detect effects of substantive interest.* (APA, 2010, p. 30)

This statement pronounces two important issues. On the one hand, a priori power calculations to obtain sufficient power are crucial for the individual researcher who is eager to publish since statistically significant findings have a higher probability of being published in scientific journals (publication bias; e.g., Atkinson, Furlang, & Wampold, 1982; Begg, 1994; Dickersin & Min, 1993; Easterbrook, Berlin, Gopalan, & Matthews, 1991; Ferguson & Brannick, 2012; Gerber & Malhotra, 2008; Greenwald, 1975; Mahoney, 1977; McDaniel, Rothstein, & Whetzel, 2006; Sterling, Rosenbaum, & Weinkam, 1995; Stern & Simes, 1997; Sterne, Gavaghan, & Egger, 2000). Thus, low statistical power is a waste of money and time because researchers unwittingly redo similar experiments again and again but hardly find significant and thus publishable results. On the other hand, too much power can result in the detection of effects that lack of substantive interest. Inappropriate high sample sizes can yield many statistically significant, but trivial effects (cf. Ferguson, 2009). That is, appropriate power is fundamental for creating comprehensive knowledge within a discipline. Rossi (1990) suggested that low statistical power not only could in combination with small effect size lead to a "large number of Type II errors, but low power also suggests the possibility of a proliferation of Type I errors in the research literature" (p. 652). To illustrate, assume a hypothesis that is not true. If an experiment is repeated several times, then the error rate (obtaining a statistically significant result given that the H_0 is actually true) is 5% in the long run. By publication bias a statistically significant result—here obtained by chance—would probably be published, whereas the inconclusive non-significant results would not. Then "100% of all published significant results would be Type I errors, despite a Type I error rate of 5%" (Rossi, 1990, p. 652). With that example in mind, the importance of power consideration—to prevent low power—is obvious. Hence formal instructions of power exist (e.g., Cohen, 1988; Hallahan & Rosenthal, 1996; Kraemer & Thiemann, 1987; Maxwell, Kelley, & Rausch, 2008; Murphy & Myors, 2004). However, power surveys of psychological research literatures have examined that the probability of finding a significant effect of medium size (i.e., $r = .30$, $d = .50$), if it does exist in reality, is in the 0.40–0.60 range (Schmidt & Hunter, 1997). This is much lower than the acceptable power level of .80 suggested by Cohen (1988). In sum, these findings reveal that the concept of statistical power is either unregarded or misunderstood.

What are the reasons for the persistence of underpowered studies? Maxwell (2004) suggested that the necessity for high power has vanished since every article contains multiple hypotheses: Testing multiple hypotheses—if not controlled—increases Type I error rate and thus increases statistical power. Therefore the probability of a significant result of any specific test within one study might be low but the probability of obtaining a statistically significant result somewhere in the study could be substantial anyway. As another result, multiple hypotheses may enhance the selective reporting of only the positive trial outcomes within published studies (Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004; Williamson, Gamble, Altman, & Hutton, 2005), leading to a more consistent perception of a given field than is actually appropriate.

Another reason for ignorance of power is that it originates from the Neyman–Pearson approach of hypothesis testing—which is hardly ever applied. Halpin and Stam (2006) analyzed 678 research articles; only 4 used procedures from the Neyman–Pearson approach. This approach does not only demand a single hypothesis (H_0) as in Fisherian significance testing, but needs also a specification of a second hypothesis (H_1) or an effect size expression of it (indexing the degree of deviation from H_0 in the underlying population). But such alternative hypotheses would require good theories, which are unfortunately rare in psychology (Gigerenzer, 2010).

Effect size

> *For the reader to appreciate the magnitude or importance of a study's finding, it is almost always necessary to include some measure of effect size in the results section.* (APA, 2010, p. 34)

An effect size represents the strength or magnitude of a relationship between the variables in the population, or a sample-based estimate of that quantity (Cohen, 1988). "Whereas a test of statistical significance provides the quantified strength of evidence (attained p level) that a null hypothesis is wrong, an effect size measures the degree to which such a null hypothesis is wrong" (Grissom & Kim, 2005, p. 4). Therefore providing both a p value and an effect size as a measure of the "statistical" and the "practical" significance of a result (Thompson, 2002b) seems to be a good choice. For instance, Huberty (2002) suggested: "Consider the P value and the E [effect size] value jointly; if the P value is small and the E value is substantial, then a real effect is obtained" (p. 236).

Effect sizes can be expressed in three ways: either as an unstandardized estimation of the sample effect size (e.g., a simple difference of means), as a standardized estimation of the sample effect size (such as Pearson r, or Cohen's d), or as standardized estimation of the population effect size (Hedges' g). Some authors advocate the usage of simple effect sizes over standardized effect sizes, arguing that an interpretation based on the original scale is more intuitive and less prone to error (Baguley, 2009; Greenland, 1998; Lenth, 2007). However, standardized effect sizes allow comparisons over several studies using different scales and consequently constitute the basic elements of research synthesis. Dozens of standardized effect size measures are currently available for researchers (Henson, 2006). For the most part they can be classified into two main groups: standardized mean differences; and measures of association (Henson, 2006; Kirk, 1996; Rosnow & Rosenthal, 2009). Since effect size measures are calculated from

sample data, they are dependent on the individual sample design that causes sampling error (e.g., sample size, number of measured variables, population effect size; for details see Vacha-Haase & Thompson, 2004). Therefore, different correction formulas have been proposed to obtain unbiased standardized estimation of the population effect sizes (e.g., Richardson, 2011; Skidmore & Thompson, 2011; Wang & Thompson, 2007). The choice of the appropriate effect size measure depends on several factors, like research design, context of the study, or audience.

Confidence interval

> *The inclusion of confidence intervals (for estimation of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results. Because confidence intervals combine information of location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended. As a rule, it is the best to use a single confidence level, specified on an a priori level (e.g., a 95% or 99% confidence interval), throughout the manuscript. Wherever possible, base discussion and interpretation of results on point and interval estimates*. (APA, 2010, p. 34)

In contrast to point estimates of unknown parameters in the population, confidence intervals describe a likely range of these values. The width of a confidence interval based on a sample statistic depends on its standard error, and therefore on both the standard deviation and the sample size. Moreover it depends on the degree of "confidence" associated with the resulting interval (e.g., 95%, or 99%; Gardner & Altman, 1986).

The main advantage of confidence intervals is that they present information of how precisely, or how accurately, a population parameter can be estimated (Brandstätter, 1999; Cumming & Finch, 2001). They thus make the uncertainty of the estimation explicit (Fidler, 2006). Although confidence intervals may be easier to understand (Brandstätter, 1999; Fidler, 2006), they are subject to similar misinterpretations as are significance tests (e.g., Frick, 1995; Hagen, 1997; Kalinowski & Fidler, 2010). Empirical studies showed that specific *confidence interval misconceptions* exist in students as well as in researchers (Belia, Fidler, Williams, & Cumming, 2005; Cumming, Williams, & Fidler, 2004). These misconceptions include definitional misconceptions—for example, that the confidence interval would be an estimate of the sample mean rather than of the population mean (Fidler, 2006)—and misconceptions about the confidence level—for example, that a 95% confidence interval for an experiment has a 95% chance of capturing the population parameter of a replication of that experiment (Cumming et al., 2004). But this would only be true if the sample mean coincides exactly on the population mean. Cumming and Maillardet (2006) showed that a 95% confidence interval will include just about 83% of future replication means. Nonetheless, Harlow (1997) identified confidence interval as the most commonly recommended supplement to NHST by contributors to the book *What If There Were No Significance Tests?* (Harlow et al., 2007).

Interim summary

Table 1 presents an overview on the approaches recommended by the APA as complements to NHST together with an outline of what information is conveyed by adding power

analysis, effect size, or confidence interval, respectively. It also gives a concise summary of the benefit of each supplement. We distinguish two categorical cases: $p < .05$ (significant result), and $p > .05$ (non-significant result). The table begins with the conclusion that can be drawn from a p value. Formally speaking, the p value is the conditional probability of the test statistic assuming H_0 is true (Kline, 2004). Thus it is the probability of the data given that H_0 is true, but never a probability of any hypothesis. This last misconception of the p value is a serious one in a long and well-documented list (Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007; Dar, Serlin, & Omer, 1994; Finch, Cumming, & Thomason, 2001; Haller & Krauss, 2002; Hoekstra, Finch, Kiers, & Johnson, 2006; Kieffer, Reese, & Thompson, 2001; Oakes, 1986; Tversky & Kahneman, 1971). In a non-significant case no conclusion at all can be drawn from a p value since H_0 cannot be accepted in a pure Fisherian logic of statistical inference (Fisher, 1935/1951; Gigerenzer & Murray, 1987). The set of conclusions is larger if the concept of statistical power, as developed within the Neyman–Pearson approach of hypothesis testing, is added. In this approach, hypotheses (H_0 and H_1) can be accepted. The additional conclusion, given a significant result, is that the alternative hypothesis (H_1) can be accepted, and when observing a non-significant outcome we can act as if H_0 was true. Moreover, knowledge of statistical power facilitates decisions on how to proceed after obtaining a null result: if power was low, one has little evidence against H_1 and should probably redo the experiment, this time with adequate power; otherwise, if power was high, affirmation of H_0 is possible within limits because the probability of committing an error (Type II) is low (Rossi, 1990).

In general, the conclusions drawn from a p value are always restricted to the quality of the null hypothesis. If the H_0 is a nil-null hypothesis—for example, there is no difference between two samples or treatments—the information obtained is trivial, since "the effects of A and B are always different—in some decimal place—for any A and B" (Tukey, 1991, p. 100). Reporting effect size together with the p value gives information about precisely this amount: the difference between A and B.

Finally, providing confidence intervals for estimates of parameters, like means, or for effect sizes, gives information on the error associated with this estimate (Cumming & Fidler, 2009; Kirk, 1996). On that basis the researcher can evaluate whether the observed value is likely to be the "true" value (Denis, 2003).

Use of power analysis, effect size, or confidence interval in psychological literature

The APA recommended the reporting of the above-presented supplements to rote NHST, but was this recommendation followed? We investigated empirically which of the three main supplements to NHST was really reported in psychological science. Note that the APA's recommendations regarding significance testing and its supplement became more strict with every edition (Finch, Thomason, et al., 2001; APA, 2010), as, for example, confidence intervals were not recommended until the 5th edition (Fidler, 2002). In the 6th edition, examples and advice on how to use and report confidence intervals are provided (APA, 2010). We therefore investigated whether the reporting frequency of the supplements was increasing over the years.

In the statistical reform debate, reformers frequently argue for more interdisciplinary communication. Across disciplines similar arguments are made, similar difficulties are

Table 1. Conclusions from the *p* value and its supplements.

Parameters	Conclusion	Benefit
p value	$p < .05$: If H_0 is true, it is very unlikely (< 5%) to get the observed (or more extreme) data;[a] reject H_0 $p \geq .05$: Nothing (suspend judgement: Fisher)	
p + power analysis	$p < .05$: If H_0 is true, it is very unlikely (< 5%) to get the observed (or more extreme) data; + accept H_1 $p \geq .05$: + Depends on power: if power is low, little evidence against H_1; if power is high, affirmation of H_0 is possible within limits (Rossi, 1990)	Declaration of H_1
p + effect size	$p < .05$: If H_0 is true, it is very unlikely (< 5%) to get the observed (or more extreme) data; + estimate of the extent to which sample results diverge from the expectations specified in the null hypothesis (Cohen, 1994; Thompson, 2002a) $p \geq .05$: + Estimate of the extent to which sample results diverge from the expectations specified in the null hypothesis	A single best estimate (of the degree of divergence of H_0)
p + confidence interval for means	$p < .05$: If H_0 is true, it is very unlikely (< 5%) to get the observed (or more extreme) data; we are very confident (95%) our interval includes the population mean $p \geq .05$: + we are 95% confident our interval includes the population mean	Estimate of population mean
p + confidence interval for effect size	$p < .05$: if H_0 is true, it is very unlikely (< 5%) to get the observed (or more extreme) data; + estimate of the extent to which sample results diverge from the expectations specified in the null hypothesis; + estimate of the precision and accuracy of observed effect $p \geq .05$: + Estimate of the extent to which sample results diverge from the expectations specified in the null hypothesis; + estimate of the (im)precision and (in)accuracy of the observed effect (Cumming & Fidler, 2009)	A single best estimate + information about the precision of that estimate

Note. [a]Kline (2004) provides a more comprehensive list of possible statements of $p < .05$.

identified, and similar solutions are proposed (Fidler & Cumming, 2007). For psychology the comparison with medicine is especially obvious. In medicine, similar recommendations were proposed, with the crucial difference that guidance and software were offered to help medical researchers to adopt the recommended practices (Altman, Machin, Bryant, & Gardner, 2000; Fidler & Cumming, 2007; Fidler, Thomason, Cumming, Finch, & Leeman, 2004). Faulkner, Fidler, and Cumming (2008) compared psychological and psychiatry randomized controlled trials and found that psychiatry papers more frequently reported statistical power (28% vs. 10%), effect size (64% vs. 43%), and confidence interval or standard error (49% vs. 19%), compared to psychology papers. We were intrigued whether the more successful implementation of statistical reform in medicine also had an impact on clinical studies within psychology since clinical psychology is in immediate proximity to medicine. We therefore investigated whether

similar differences in reporting practices could also be found between *clinical* and *non-clinical* studies within psychology.

Method

To identify relevant studies that specifically investigated the reporting practice of power analysis, effect size, or confidence interval, a literature search in the Web of Knowledge database was performed. Keywords were "statistical power analys*AND psychology," "effect size report*AND psychology," or "confidence interval*AND psychology," respectively, in title, abstract, or keyword section. All German- or English-language studies published between 1990 and 2010 were accessed. The resulting 245 hits were scanned for their appropriateness. If they investigated the reporting practice of power analysis, effect size, or confidence interval, the reference sections of the articles were investigated for additional studies on this topic according to the ancestry approach. Additionally, all appropriate articles were further searched for citing articles (descendency approach).

Note that our analysis is based on the results of analyses of several articles. We collected existing studies that investigated the frequency of usage of effect size, power, and confidence interval, and analyzed those, rather than the original articles. Thus, we reviewed existing studies in order to obtain a comprehensive picture of the reporting practices of the additional measures beyond the *p* value. Our data are percentages, weighted by the number of articles involved in each study.

In the original studies the three reporting practices where evaluated according to the following criteria: A study was coded "power analysis" when power was calculated or reported within the results section of an article. A posteriori power analysis or implicit suggestions of power, for example in the discussion, were not coded as power analysis. For coding of an effect size, the results section had to contain a standardized or unit-free measure of effect size. Whenever a confidence interval was present in the results section, regardless of whether it was presented in figures, tables, or in the text, it was coded as such.

For each study, the name of the journal and the number of articles included were recorded. Moreover, the time span in which the included articles were published was extracted. Some studies assessed changes over time and therefore analyzed articles from a broad time range. As the main measures of interest, the percentage of articles that reported a power analysis, an effect size, or a confidence interval, respectively, were calculated and the weighted average was computed. We distinguished between *clinical* and *non-clinical* studies within psychology by classifying articles as *clinical* if they were published in journals containing one of the following keyword(s) in their title*: clinical, counseling, disability, health, disorder, randomized controlled trials*. This procedure yielded an agreement rate of over 97% compared to the classification of an independent psychological researcher; the only disagreement was solved after discussion.

Results

The search revealed 11 studies (containing 1,164 articles) investigating the reporting rate of power analysis, 9 studies (including 1,046 articles) focused on confidence intervals, and 29 studies (including 6,366 articles) focused on effect size reporting. As can be seen in Table 2, the weighted average of articles that used or reported a power analysis was

only 2.9%. Confidence intervals were reported more frequently with a weighted average of 10.4%. Effect sizes seem to be the most accepted additional reporting method (weighted average = 38.4%), however, with a huge range (1–81%). This weighted average is probably an overestimation of the real usage of effect sizes since some measures (e.g., correlations) are test statistics and effect sizes at the same time. Consequently, they were counted as effect sizes in the investigated studies, although it remains questionable whether effect size reporting was intended by the authors initially.

The results for clinical and non-clinical studies are depicted in Figures 1 and 2, respectively. Power analyses were clearly more frequent in clinical articles, but effect sizes were more frequent in non-clinical articles. However, the reporting frequency of effect sizes seemed to increase over the years for clinical as well as for non-clinical articles. There were only two studies that analyzed confidence interval usage in clinical articles, but they suggest that confidence intervals are more common in clinical than in non-clinical articles. The potential impact that editors have on the reporting frequency of statistical measures can be seen in the reporting frequency of confidence intervals in non-clinical articles (Figure 2). One study (Finch et al., 2004) investigated articles from the journal *Memory & Cognition*, where a former Editor, Geoffrey Loftus (from 1994 to 1997), strongly encouraged the application of confidence intervals. This increased the reporting frequency of confidence intervals in a sustained manner so that after Loftus's time as Editor 27% of articles still reported confidence intervals.

Table 2. Summary results for studies of three reporting practices in psychology.

Study	Journal	Area	N[a]	Year	%
Reports of power analysis use					
Bezeau & Graves (2001)	*JCEN, JINS, N*	Clinical	66	1998–1999	3.0
Clark-Carter (1997)	*BJP*	Non-clinical	54	1993–1994	1.9
Conzelmann & Raab (2009)	*ZS*	Non-clinical	28	2004–2007	0.0
Crosby et al. (2006)	Eating & anxiety disorder publications	Clinical	152	2000	2.0
Faulkner et al. (2008)	Randomized controlled trials	Clinical	104	1999–2003	10.0
Finch, Cumming et al. (2001)	*BJP, JAP*	Non-clinical	60	1999	6.7
Hager (2005)	*D, KE, PB, PEU, PR, ZAOP, ZDDP, ZEPPP, ZGP, ZKPPT, ZMP, ZP, ZPP, ZSP*	Non-clinical	436	2001–2002	2.3
Keselman et al. (1998)	*AERJ, CD, CEP, CI, DP, ETRD, JAP, JCP, JECR, JECP, JEE, JEP, JPSP, JRB, JRME, RRQ, SE*	Non-clinical	106	1994–1995	0.0
Kosciulek & Szymanski (1993)	*JARC, RCB, RE, RP*	Clinical	32	1990–1991	3.1

(Continued)

Table 2. (Continued)

Study	Journal	Area	N[a]	Year	%
Osborne (2008)	*BJEP, CEP, JEP*	Non-clinical	96	1998–1999	2.1
Woods et al. (2006)	Neuropsychological studies	Non-clinical	30	1997–2004	0.0
Overall (11)			*1,164*		*2.9*[b]
Reports of confidence interval use					
Byrd (2007)	*EAQ*	Non-clinical	73	1997–2006	0.0
Crosby et al. (2006)	Eating & anxiety disorder publications	Clinical	152	2000	11.8
Cumming et al. (2007)	*AP, C, CD, JAbP, JACP, JCCP, JEPG, JPSP, PS, QJEP*	Non-clinical	40	2005–2006	10.6
Fidler et al. (2005)	*JCCP*	Clinical	60	2000–2001	17.0
Finch, Cumming, et al. (2001)	*BJP, JAP*	Non-clinical	60	1999	3.3
Finch et al. (2004)	*MC*	Non-clinical	228	1998–2000	27.0
Hoekstra et al. (2006)	*PBR*	Non-clinical	259	2002–2004	5.0
Keselman et al. (1998)	*AERJ, CD, CEP, CI, DP, ETRD, JAP, JCP, JECP, JECR, JEE, JEP, JPSP, JRB, JRME, RRQ, SE*	Non-clinical	106	1994–1995	0.0
Kieffer et al. (2001)	*ERJ, JCP*	Non-clinical	68	1997	0.0
Overall (9)			*1,046*		*10.4*[b]
Reports of effect size use					
Alhija & Levy (2009)	*CEP, EC, ECRQ, ER, ICD, JEE, JEP, JLD, JSP, LDRP*	Non-clinical	183	2003–2004	71.0
Andersen, McCullagh, & Wilson (2007)	*JASP, JSEP, SP*	Non-clinical	54	2005	81.0
Bezeau & Graves (2001)	*JCEN, JINS, N*	Clinical	66	1998–1999	9.1
Byrd (2007)	*EAQ*	Non-clinical	73	1997–2006	72.7
Clark-Carter (1997)	*BJP*	Non-clinical	54	1993–1994	38.9
Conzelmann & Raab (2009)	*ZS*	Non-clinical	28	2004–2007	85.7
Crosby et al. (2006)	Eating & anxiety disorder publications	Clinical	152	2000	23.3
Dunleavy, Barr, Glenn, & Miller (2006)	*JAP, JEP, JEPLM, JPSP*	Non-clinical	736	2002–2003	62.5
Faulkner et al. (2008)	Randomized controlled trials	Clinical	104	1999–2003	43.0
Fidler et al. (2005)	*JCCP*	clinical	60	2000–2001	38.0
Hager (2005)	*D, KE, PB, PEU, PR, ZAOP, ZDDP, ZEPPP, ZGP, ZKPPT, ZMP, ZP, ZPP, ZSP*	Non-clinical	436	2001–2002	29.8

Table 2. (Continued)

Study	Journal	Area	N[a]	Year	%
Hoekstra et al. (2006)	*PBR*	Non-clinical	259	2002–2004	1.0
Ives (2003)	*JLD, LDQ, LDRP*	Clinical	526	1990–1999	25.0
Keselman et al. (1998)	*AERJ, CD, CEP, CI, DP, ETRD, JAP, JCP, JECR, JECP, JEE, JEP, JPSP, JRB, JRME, RRQ, SE*	Non-clinical	411	1994–1995	10.0
Kieffer et al. (2001)	*ERJ, JCP*	Non-clinical	68	1997	51.5
Kirk (1996)	*JAP, JEP, JEPLM, JPSP*	Non-clinical	391	1995	42.2
Matthews et al. (2008)	*GCQ, GTI, JEG, JSGE, RR*	Non-clinical	149	2001–2005	45.9
McMillan, Lawson, Lewis, & Synder (2002)	*CEP, JEE, JEP, JER*	Non-clinical	508	1997–2000	29.13
Osborne (2008)	*BJEP, CEP, JEP*	Non-clinical	96	1998–1999	16.7
Paul & Plucker (2004)	*GCQ, JEG, RR*	Non-clinical	325	1995–2000	28.9
Plucker (1997)	*GCQ, JEG, RR*, and articles about "giftedness"	Non-clinical	157	1992–1995	21.7
Snyder & Thompson (1998)	*SPQ*	Non-clinical	35	1990–1996	54.3
Sun, Pan, & Wang (2010)	*AERJ, ECRQ, EEPA, DR, ICD, JEE, JEP, JEPA, JEPLMC, JEPPP, JLD, JSP, LDRP, SPQ*	Non-clinical	1,243	2005–2007	49.07
Thompson (1999b)	*EC*	Non-clinical	23	1996–1998	13.0
Thompson & Snyder (1997)	*JEE*	Non-clinical	22	1994–1996	63.6
Thompson & Snyder (1998)	*JCD*	Clinical	25	1996	60.0
Vacha-Haase & Ness (1999)	*PPRP*	Non-clinical	16	1997	31.3
Vacha-Haase & Nilsson (1998)	*MECD*	Clinical	83	1990–1996	35.3
Vacha-Haase, Nilsson, Reetz, Lance, & Thompson (2000)	*JCP, PA*	Clinical	83	1997	51.8
Overall (29)			*6,366*		*38.4*[b]

Note. AERJ = American Educational Research Journal; AP = Acta Psychologica; BJEP = British Journal of Educational Psychology; BJP = British Journal of Psychology; C = Cognition; CD = Child Development; CEP = Contemporary Educational Psychology; CI = Cognition and Instruction; D = Diagnostica; DP = Developmental Psychology; DR = Dreaming; EAQ = Educational Administration Quarterly; EC = Exceptional Children; ECRQ = Early Childhood Research Quarterly; EEPA = Educational Evaluation and Policy Analysis; ER = Educational Research; ERJ = Educational Research Journal; ETRD = Educational Technology, Research and Development; GCQ = Gifted Child Quarterly; GTI = Gifted and Talented International; ICD = Infant and Child Development; JAbP = Journal of Abnormal Psychology; JACP = Journal of Abnormal Child Psychology; JAP = Journal of Applied Psychology; JARC = Journal of Applied Rehabilitation Counseling; JASP = Journal of Applied Sport Psychology; JCCP = Journal of Consulting and Clinical Psychology; JCD = Journal of Counseling & Development; JCEN = Journal of Clinical and Experimental

(Continued)

Table 2. (Continued)

Neuropsychology; JCP = Journal of Counseling Psychology; JECP = Journal of Experimental Child Psychology; JECR = Journal of Educational Computing Research; JEE = Journal of Experimental Education; JEG = Journal for the Education of the Gifted; JEP = Journal of Educational Psychology; JEPA = Journal of Experimental Psychology: Applied; JEPG = Journal of Experimental Psychology: General; JEPLM = Journal of Experimental Psychology, Learning, and Memory; JEPLMC = Journal of Experimental Psychology: Learning, Memory, and Cognition; JEPPP = Journal of Experimental Psychology: Human Perception and Performance; JER = Journal of Educational Research; JINS = Journal of the International Neuropsychology Society; JLD = Journal of Learning Disabilities; JPSP = Journal of Personality and Social Psychology; JRB = Journal of Reading Behavior; JRME = Journal for Research in Mathematics Education; JSEP = Journal of Sport & Exercise Psychology; JSGE = Journal of Secondary Gifted Education; JSP = Journal of Special Education; KE = Kindheit und Entwicklung; LDQ = Learning Disability Quarterly; LDRP = Learning Disabilities Research & Practice; MC = Memory & Cognition; MECD = Measurement and Evaluation in Counseling and Development; N = Neuropsychology; PA = Psychology and Aging; PB = Psychologische Beiträge; PBR = Psychonomic Bulletin & Review; PEU = Psychologie in Erziehung und Unterricht; PPRP = Professional Psychology: Research and Practice; PR = Psychologische Rundschau; PS = Psychological Science; QJEP = Quarterly Journal of Experimental Psychology; RCB = Rehabilitation Counseling Bulletin; RE = Rehabilitation Education; RP = Rehabilitation Psychology; RR = Roeper Review; RRQ = Reading Research Quarterly; SE = Sociology of Education; SP = The Sport Psychologist; SPQ = School Psychology Quarterly; ZAOP = Zeitschrift für Arbeits- und Organisationspsychologie; ZDDP = Zeitschrift für Differentielle und Diagnostische Psychologie; ZEPPP = Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie; ZGP = Zeitschrift für Gesundheitspsychologie; ZKPPT = Zeitschrift für Klinische Psychologie und Psychotherapie; ZMP = Zeitschrift für Medienpsychologie; ZP = Zeitschrift für Psychologie; ZPP = Zeitschrift für Pädagogische Psychologie; ZS = Zeitschrift für Sportpsychologie; ZSP = Zeitschrift für Sozialpsychologie.

[a]Number of articles included.

[b]Weighted average.

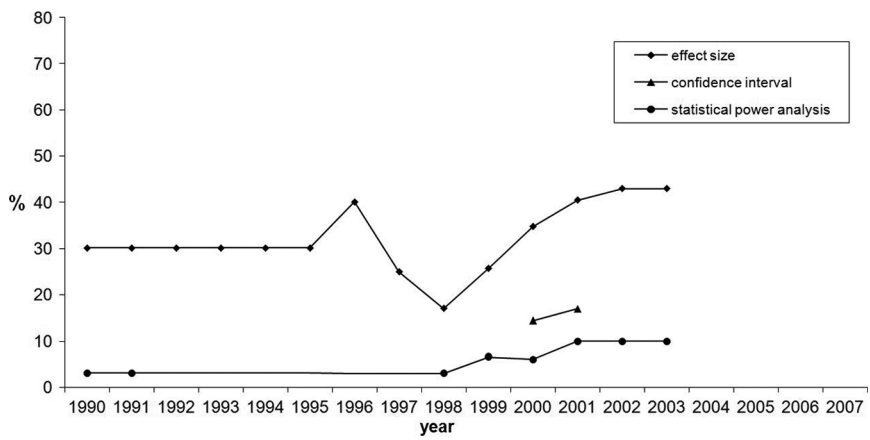

Figure 1. Average reporting frequency of effect size, power analysis, and confidence interval in clinical psychological studies from 1990 to 2007.

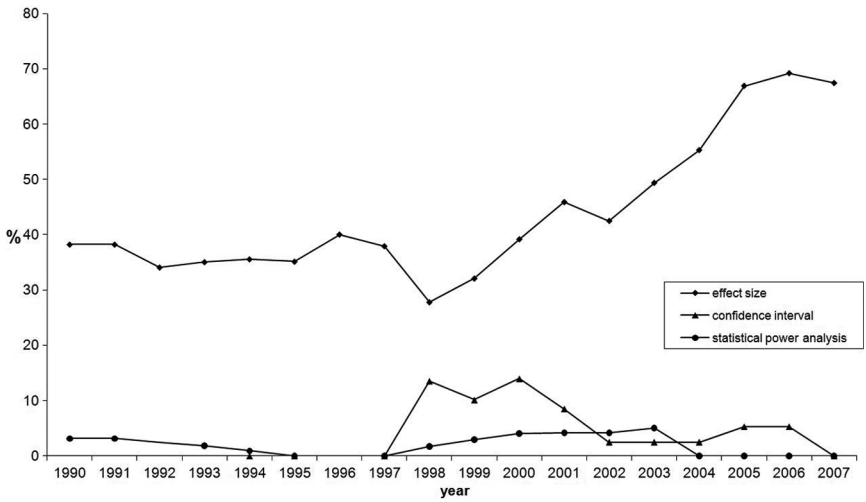

Figure 2. Average reporting frequency of effect size, power analysis, and confidence interval in non-clinical psychological studies from 1990 to 2007.

To examine if reporting practices of additional measures to NHST have increased over the years owing to stricter guidelines of the APA, a multiple linear regression analysis was computed for each reporting practice. We examined whether the percentage of articles that reported effect size, confidence interval, or power analysis could be predicted by year of publication (as the mean of analyzed years) of the underlying articles and research area (clinical vs. non-clinical). This analysis revealed an increase in effect size reporting over the years ($b = 2.09$, 95% CI [1.99; 2.19], $SE\ b = .05$, beta = .49, $R^2 = .26$) and a negligible influence of research area ($b = –3.58$, 95% CI [–4.72; –2.44], $SE\ b = .58$, beta = –.07, $\Delta R^2 < .01$). That is, every year witnessed an increase of about 2% more reports of effect size, irrespective of research area.

The percentage of reported confidence interval and power analysis was only influenced by research area (confidence interval: $b = 3.60$, 95% CI [2.11; 5.07], $SE\ b = .76$, beta = .15, $R^2 = .02$; power analysis: $b = 2.54$, 95% CI [2.25; 2.83], $SE\ b = .15$, beta = .45, $R^2 = .20$) with *clinical* articles reporting about 3.5% more confidence intervals and about 2.5% more power analyses than *non-clinical* articles.

Discussion

This survey integrates the results of several earlier surveys for the purpose of creating generalizations. Although this implies summarizing over different areas as well as averaging over different methodological applications, it provides a rough estimate of the usage of power analysis, effect size, and confidence interval in the psychological literature over recent years. In this way, the current study follows the epistemic culture of meta-analysis with the dominant focus on a comprehensive quantitative review rather than a detailed discussion of the quality of the aggregated empirical data (Shercliffe,

Stahl, & Tuttle, 2009). The analysis revealed that effect sizes are the most frequently used additional reporting method (approximately 38%) and their dissemination increased over the years. Confidence interval and power analysis are reported in about 10% and 3% of articles, respectively. For these two practices no increase in reporting frequency over the years could be detected. While confidence interval and power analysis were more common in clinical studies, this was not the case for effect sizes, not even in tendency.

Neglect of power analysis

Only 3% of articles report power analysis. It might be that these articles specifically tested null relationships. Recall that if someone wants to demonstrate that there is no relationship between two variables or no difference between two groups regarding a specific variable, it is not enough to come up with an insignificant result at a specific Type I error rate for the simple reason that with a small enough sample one will almost always find an insignificant result. Therefore, Type II error, acceptance of a false null hypothesis, has to be controlled for (Cohen, 1977; Sedlmeier & Gigerenzer, 1989). It is proposed that these studies should achieve a power of .95 rather than the common level of .80 (Cohen, 1988). However, an analysis of 54 such articles from Cashen and Geiger (2004) revealed that these standards are not adhered to at all. They found an average power of these studies of .29, which means the Type II error rate was 14 times greater than what is advocated in the literature.

Our findings are also in line with a survey by Mone, Mueller, and Mauland (1996) in which they asked authors of applied psychology and management journals about their usage of statistical power. Sixty-four percent of respondents reported *never* using power analysis at all. The main reason for non-usage was that editors or reviewers do not require power analysis. Therefore, it is unlikely that the requirement of testing a null relationship is the main explanation for the power use we found.

Recall that there are three main consequences of neglecting power. First, if researchers design studies that have little chance of detecting an effect that actually exists, they risk wasting time and resources for research which is likely not to get published. Second, power analysis enables sound decisions as to what to do when non-significant results have been obtained (see Table 1). Without knowledge of power, no conclusion can be drawn when failing to reject the null hypothesis. Third, and most importantly, low statistical power may be a threat for the entire field of research. As methodologists have pointed out at various times (e.g., Bakan, 1966; Overall, 1969; Rossi, 1990; Selvin & Stuart, 1966; Sterling, 1959; Tversky & Kahneman, 1971), low power increases the probability of Type I error because the probability of rejecting a true null hypothesis is in this case only slightly smaller than the probability of rejecting the null hypothesis when the alternative is true. Therefore a substantial part of all significant results may be due to false rejections of valid null hypotheses (Rossi, 1990). Taken together, the current practice of not considering power analysis has serious implications for the entire research discipline. Without power analysis, nothing can be said about $H_{1,}$ which typically is the hypothesis the researcher is interested in. If, however, researchers test H_0 but talk about $H_{1,}$ this is invalid.

Effect sizes, but without confidence intervals

Our results suggest that, if supplements to NHST are reported at all, it is in the form of effect sizes. Reporting effect sizes is important for at least two reasons. First, it switches the focus from the *p* value to the practical significance of research findings (Fritz, Lermer, & Kühberger, 2011; Kirk, 1996). Since some authors argue that—at least in non-experimental research (e.g., Meehl, 1978)—virtually all point null hypotheses are false, the size of an effect, and not statistical significance, should be the focus of interest. This also changes the focus from H_0 to H_1, which typically is the interesting hypothesis. Second, effect sizes are essential for comparing and aggregating research findings and therefore for cumulative science.

Nonetheless, there may be a serious problem of effect sizes in the context of the dichotomous decision making of NHST. Effect sizes should be relatively independent from sample size (e.g., Hunter & Schmidt, 1990; Levine, Weber, Hullett, Park, & Lindsey, 2008). An effect should have the same magnitude regardless of whether it has been assessed by a sample size of 50 or a sample size of 5,000. However, this theoretical advantage may be violated when the publication of research findings is a function of the *p* value (e.g., under publication bias). Given publication bias, studies with small effect sizes are significant only when they have a large sample size, which guarantees significance. In contrast, large effect sizes only need small samples in order to be significant. That is, practically there exists a negative correlation between sample size and effect size. Indeed, Levine, Asada, and Carpenter (2009) analyzed 51 published meta-analyses in the area of communication research. They found a negative correlation between effect size and sample size in approximately 80% of the meta-analyses examined. They explain: "Smaller studies, however, report, on average, larger effects because smaller effects are non-significant and don't make it into print, and because larger effects occur more often (because of wider confidence interval). The net result is a negative n-r correlation" (p. 290). They concluded that this finding "most likely stems from a bias against non-significant findings, and it likely results in meta-analyses overestimating effect sizes" (p. 298). This is a serious problem in meta-analyses: even the common practice of including unpublished studies as a control for publication bias appears to be ineffective or even counterproductive (Ferguson & Brannick, 2012).

In detail, the sampling error explanation goes like this: an experiment's sample effect size is an estimate of the real effect size, which we unfortunately do not know. The experiment will sometimes overestimate and sometimes underestimate the true effect size. Smaller studies have more sampling error than larger ones. Using a small sample size, only those experiments that find (by chance, but wrongly) a large effect size will, in the end, be significant. As a consequence of publication bias, only the set of big effect sizes will make it into the published literature, thus leading to an overestimation of the effect size in any meta-analysis. Schmidt (1996) provided a hypothetical example of this fact: assume that the true effect of some treatment is medium sized (Cohen's $d = .50$). If this effect is examined with a small sample (total $n = 30$), it will frequently be insignificant, since with $n = 30$ it reaches significance only in studies that find a biased effect size of $d \geq .62$ ($p = .05$, one-tailed test). Under the assumption that only significant studies are published, all published studies in this example overestimate the true effect size!

Calculating the mean of the published studies will yield a still higher mean effect size of $d = .89$. That is, owing to publication bias the true effect would be overestimated by nearly 80%!

Given publication bias, the pursuit of statistically significant results within NHST will ironically undermine effect size reporting. Consequently, simple reporting of the point estimate of the effect size is not enough. Rather, it is necessary to report a confidence interval around the estimate (Baguley, 2009). Of course, confidence intervals cannot guarantee that effect sizes are not overestimated, but the reader can tell from the width of the intervals how accurate, and therefore trustworthy, the estimation is (Cumming, 2012). As Hedges (2008) pointed out: "Effect size estimates with very wide confidence intervals may be of less practical value than estimates with less uncertainty" (p. 170). Accordingly, a small effect with a narrow confidence interval can be of more practical relevance than a big effect size with great uncertainty. Our findings suggest, though, that this information is hardly ever provided.

Statistical reform in psychology compared to medicine

The distinction between *clinical* and *non-clinical* psychological studies led to a similar pattern of results to an earlier survey by Faulkner et al. (2008) comparing psychology and psychiatry randomized controlled trials. We found that *clinical* articles report about 3% more confidence intervals and power analyses than *non-clinical* articles. The better implemented statistical reform in medicine therefore seems to affect the immediate area of clinical psychology. Fidler, Cumming, Burgman, and Thomason (2004) and Fidler and Cumming (2007) analyzed statistical reform attempts in medicine and compared them to the efforts in psychology. According to their analysis, in the 1950s and 1960s NHST became standard practice in medicine as well, but after serious criticisms editorial policies were changed, and accessible guidance and software were published to help medical researchers to adopt the recommended practices. Furthermore many medical journals have statistical editors as well as substantive editors, which is, according to Fidler (2006), simply feasible owing to the larger budgets of medical journals for organizing these associate editors. Psychology internalized the responsibility for statistical analysis, whereas medicine appropriated outside expertise. Statistical failings, like lack of statistical power in clinical trials, were in medicine considered as ethical concerns (e.g., Altman, 1982; May, 1975). For instance, the exposure of patients as experimental subjects to a new treatment is only to be approved if the study has enough statistical power to find a positive treatment effect, if it does exist. The same argument appears flawed in fields where students have to fill in non-hazardous questionnaires—although for the progress of science it would be on any account indispensable. Another determinant of the different success of statistical reforms may lie in the nature of the policies: medicine required but psychology only encouraged supplementing the p value with additional measures (Fidler & Cumming, 2007; Fidler, Cumming, et al., 2004; Fidler, Thomason, et al., 2004). Moreover, medical editors presented these stricter guidelines more in unison and consistently, as in the guidelines of the International Committee of Medical Journal Editors (1988). In sum, these factors led to the result that medicine has better implemented the statistical reform recommendations than has psychology.

Limitations

It should be acknowledged that by aiming to provide a comprehensive review of reporting practices in psychological journals, the current study combines articles over a broad discipline including various journals. Previous studies showed that reporting practices can differ between journal types: for example Sun et al. (2010) found that journals edited by the American Educational Research Association (AERA) had higher effect size reporting rates compared to APA and independent journals. Thus whenever the reporting practice of specific journals, such as AERA journals, is under scrutiny, the original articles should be consulted for the specific results.

Notice that the articles combined in the current analysis only capture whether the supplements were reported in journal articles, not whether they were interpreted. We are afraid that the interpretation rates might actually be much lower. As Ferguson (2009) observed: "I can report only on my subjective perception here, but to the extent that effect sizes are reported in social science literature (and this remains imperfect), they are seldom interpreted" (p. 135).

Conclusions

According to the APA, NHST should only be a starting point for inference, but should be supplemented with additional reporting elements—like effect size and confidence interval—to get across the comprehensive meaning of study results. Unfortunately it seems that only the additional reporting of effect size has made it into reporting practice in empirical psychological research, at least to some extent. However, without additional reporting of confidence intervals in combination with the current practice of dichotomous decision making, this practice has the potential to be highly misleading, since it leads to an overestimation of effect sizes. For any research area where a negative correlation between sample size and effect size can be shown, the advice to report effect size in addition to *p* level does probably more harm than good and should therefore be abandoned. To the degree that number of—significant and therefore published—studies is used as the indicator of scientific success rather than substance of the papers, these problems might even get worse.

Funding

This research was partly supported by an Excellentia-Dissertations-Scholarship (30.002/2-2007) from the University of Salzburg to the first author. Astrid Fritz is recipient of a DOC-fFORTE-fellowship (23171) of the Austrian Academy of Sciences.

Acknowledgements

We would like to thank two anonymous reviewers for suggestions that have improved the manuscript.

Note

References marked with an asterisk indicate studies included in the analysis.

References

*Alhija, F. N., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement, 69*, 245–265.

Altman, D. G. (1982). Misuse of statistics is unethical. In D. G. Altman & S. M. Gore (Eds.), *Statistics in practice* (pp. 1–2). London, UK: BMJ Books.

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (Eds.). (2000). *Statistics with confidence: Confidence intervals and statistical guidelines* (2nd ed.). London, UK: BMJ Books.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

*Andersen, M. B., McCullagh, P., & Wilson, G. J. (2007). But what do the numbers really tell us? Arbitrary metrics and effect size reporting in sport psychology research. *Journal of Sport & Exercise Psychology, 29*, 664–672.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and alternatives. *Journal of Wildlife Management, 64*, 912–923.

Atkinson, D. R., Furlang, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology, 29*, 189–194.

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

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423–437.

Balluerka, N., Gomez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*, 55–70.

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399–409). New York, NY: Russell Sage Foundation.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*, 389–396.

*Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology, 23,* 399–406.

Boring, E. G. (1919). Mathematical vs. scientific importance. *Psychological Bulletin*, *16*, 335–338.

Brandstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods of Psychological Research Online, 14*, 33–46.

Budge, G., & Katz, B. (1995). Constructing psychological knowledge: Reflections on science, scientists and epistemology in the *APA Publication Manual. Theory & Psychology, 5*, 217–231.

*Byrd, J. K. (2007). A call for statistical reform in *Educational Administration Quarterly. Educational Administration Quarterly, 43,* 381–391.

Capraro, M. M., & Capraro, R. M. (2003). Exploring the *APA Fifth Edition Publication Manual*'s impact on the analytic preferences of journal editorial board members. *Educational and Psychological Measurement, 63*, 554–565.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods, 7*, 151–167.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*, 98–113.

Chan, A. W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials—Comparison of protocols to published articles. *Journal of the American Medical Association, 291*, 2457–2465.

*Clark-Carter, D. (1997). The account taken of statistical power in research published in the *British Journal of Psychology. British Journal of Psychology, 88,* 71–83.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York, NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*, 243–253.

*Conzelmann, A., & Raab, M. (2009). Datenanalyse: Das Null-Ritual und der Umgang mit Effekten in der Zeitschrift für Sportpsychologie [Data analysis: The null ritual and the use of effect sizes]. *Zeitschrift für Sportpsychologie, 16*, 43–54.

*Crosby, R. D., Wonderlich, S. A., Mitchell, J. E., deZwaan, M., Engel, S. G., Connolly, K., … Taheri, M. (2006). An empirical analysis of eating disorders and anxiety disorders publications (1980–2000) - Part II: Statistical hypothesis testing. *International Journal of Eating Disorders, 39*, 49–54.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Journal of Psychology, 217*, 15–26.

*Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., … Wilson, S. (2007). Statistical reform in psychology: Is anything changing. *Psychological Science, 18,* 230-233.

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*, 532–574.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods, 11*, 217–227.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 299–311.

Dalton, D. R., & Dalton, C. M. (2008). Meta-analyses: Some very good steps toward a bit longer journey. *Organizational Research Methods, 11*, 127–147.

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*, 75–82.

Denis, D. (2003). Alternatives to null hypothesis significance testing. *Theory & Science, 4*, 1. Retrieved from http://theoryandscience.icaap.org/content/vol4.1/02_denis.html

Dickersin, K., & Min, Y. I. (1993). Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences, 703*, 135–146.

*Dunleavy, E. M., Barr, C. D., Glenn, D. M., & Miller, K. R. (2006). Effect size reporting in applied psychology: How are we doing? *The Industrial-Organizational Psychologist, 43*, 29–37.

Easterbrook, P. J., Berlin, J. A., Gopalan, R., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet, 337*, 867–872.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

*Faulkner, C., Fidler, F., & Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. *Behaviour Research and Therapy, 46*, 270–281.

Ferguson, C. J. (2009). Is psychological research really as good as medical research? Effect size comparisons between psychology and medicine. *Review of General Psychology, 13,* 130–136.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*, 120–128. doi: 10.1037/a0024445

Fidler, F. (2002). The fifth edition of the *APA Publication Manual*: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement, 62*, 749–770.

Fidler, F. (2006). *From statistical significance to effect size estimation: Statistical reform in psychology, medicine and ecology.* Unpublished Ph.D. thesis, University of Melbourne, Australia.

Fidler, F., & Cumming, G. (2007). Lessons learned from statistical reform efforts in other disciplines. *Psychology in the Schools, 44*, 441–449.

Fidler, F., Cumming, G., Burgman, M., & Thomason, N. (2004). Statistical reform in medicine, psychology and ecology. *The Journal of Socio-Economics, 33*, 615–630.

*Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., … Schmitt, R. (2005). Toward improved statistical reporting in the *Journal of Consulting and Clinical Psychology. Journal of Consulting and Clinical Psychology, 73*, 136–143.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. *Psychological Science, 15*, 119–126.

*Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61*, 181–210

*Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., … Goodman, O. (2004). Reform of statistical inference in psychology: The case of memory and cognition. *Behavior Research Methods, Instruments, & Computers, 36*, 312–324.

Finch, S., Thomason, N., & Cumming, G. (2001). Past and future APA guidelines for statistical practice. *Theory & Psychology, 12*, 825–853.

Fisher, R. A. (1950). *Statistical methods for research workers* (11th ed.). Edinburgh, UK: Oliver & Boyd. (Original work published 1925)

Fisher, R. A. (1951). *The design of experiments* (5th ed.). Edinburgh, UK: Oliver & Boyd. (Original work published 1935)

Frick, R. W. (1995). A problem with confidence intervals. *American Psychologist, 50*, 1102–1103.

Fritz, A., Lermer, E. M., & Kühberger, A. (2011). *The significance fallacy in inferential statistics.* Manuscript submitted for publication.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal, 292*, 746–750.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research, 37,* 3–30.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory & Psychology, 20*, 733–743.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gigerenzer, G., & Murray, D. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

Greenland, S. (1998). Meta-analysis. In K. Rothman & S. Greenland (Eds.), *Modern epidemiology* (pp. 287–318). Philadelphia, PA: Lippincott-Raven.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15–24.

*Hager, W. (2005). Vorgehensweise in der deutschsprachigen psychologischen Forschung: Eine Analyse empirischer Arbeiten der Jahre 2001 und 2002 [Procedures in German empirical research: An analysis of some psychological journals of the years 2001 and 2002]. *Psychologische Rundschau, 56*, 191–200.

Hallahan, M., & Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behaviour Research and Therapy, 34*, 489–499.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online, 7*, 1–20.

Halpin, P. F., & Stam, H. J. (2006). Inductive inference or inductive behavior: Fisher and Neyman–Pearson approaches to statistical testing in psychological research (1940–1960). *American Journal of Psychology, 119,* 625–653.

Harlow, L. L. (1997). Significance testing introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Erlbaum.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hays, W. L., & Winkler, R. L. (1970). *Statistics: Probability, inference and decision*. New York, NY: Holt, Rinehart & Winston.

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

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist, 34*, 601–629.

*Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review, 13*, 1033–1037.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement, 60*, 661–681.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*, 227–240.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

International Committee of Medical Journal Editors. (1988). Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine, 108*, 258–265.

*Ives, B. (2003). Effect size use in studies of learning disabilities. *Journal of Learning Disabilities, 36*, 490–504.

Kalinowski, P., & Fidler, F. (2010). Interpreting significance: The differences between statistical significance, effect size, and practical importance. *Newborn and Infant Nursing Reviews, 10*, 50–54.

*Keselman, H. J., Huberty, C. J, Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., … Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350–386.

*Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1977: A methodological review. *Journal of Experimental Education, 69*, 280–309.

Killeen, P. R. (2006). Beyond statistical significance: A decision theory for science. *Psychonomic Bulletin & Review, 13*, 549–562.

*Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746–759.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

*Kosciulek, J. F., & Szymanski, E. M. (1993). Statistical power analysis of rehabilitation counseling research. *Rehabilitation Counseling Bulletin, 36*, 212–219.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Beverly Hills, CA: Sage.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14,* 293–300.

Lenth, R. V. (2007). Statistical power calculations. *Journal of Animal Science, 85*, E24–E29.

Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs, 76*, 286–302.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. M. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research, 34*, 171–187.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1,* 161–175.

*Matthews, M. S., Gentry, M., McCoach, D. B, Worrell, F. C., Matthews, D. M., & Dixon, F. (2008). Evaluating the state of a field: Effect size reporting in gifted education. *Journal of Experimental Education, 77*, 55–68.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences and remedies. *Psychological Methods, 9*, 147–163.

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

May, W. W. (1975). The composition and function of ethical committees. *Journal of Medical Ethics, 1*, 23–29.

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology, 59*, 927–953.

*McMillan, J. H., Lawson, S., Lewis, K., & Snyder, A. (2002, April). *Reporting effect size: The road less traveled*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology, 49,* 103–120.

Murphy, K. R, & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Erlbaum.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Statistical Society, Series A, 231*, 289–337.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York, NY: Wiley.

*Osborne, J. W. (2008). Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology, 28*, 151–160.

Overall, J. E. (1969). Classical statistical hypothesis testing within the context of Bayesian theory. *Psychological Bulletin, 71*, 285–292.

*Paul, K. M., & Plucker, J. A. (2004). Two steps forward, one step back: Effect size reporting in gifted education research from 1995–2000. *Roeper Review, 26*, 68–72.

*Plucker, J. A. (1997). Debunking the myth of the "highly significant" result: Effect sizes in gifted education research. *Roeper Review, 20*, 122–126.

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6,* 135–147.

Rosnow, R. L., & Rosenthal, R. (2009). Effect sizes: Why, when, and how to use them. *Journal of Psychology, 217*, 6–14.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646–656.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115–129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Erlbaum.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 107,* 309–316.

Selvin, H. C., & Stuart, A. (1966). Data-dredging procedures in survey research. *American Statistician, 20*, 20–23.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*, 73–83.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook of data analysis in behavioural sciences: Vol. 1. Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

Shercliffe, R. J., Stahl, W., & Tuttle, M. P. (2009). The use of meta-analysis in psychology: A superior vintage or the casting of old wine in new bottles? *Theory & Psychology, 19,* 413–430.

Skidmore, S. T., & Thompson, B. (2011). Choosing the best correction formula for the Pearson r^2 effect size. *Journal of Experimental Education, 79*, 257–278.

*Snyder, P. A., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. *School Psychology Quarterly, 13*, 335–348.

Sterling, T. D. (1959). Publication bias and their possible effects on inference drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54*, 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication bias revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*, 108–112.

Stern, J. M., & Simes, R. J. (1997). Publication bias: Evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal, 315*, 640–645.

Sterne, A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis. *Journal of Clinical Epidemiology, 53,* 1119–1129.

*Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102*, 989–1004.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*, 361–377.
Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*, 26–30.
Thompson, B. (1997–2001). *402 citations questioning the indiscriminate use of null hypothesis significance tests in observational studies*. Retrieved from http://www.warnercnr.colostate.edu/~anderson/thompson1.html
Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 10*, 167–183.
*Thompson, B. (1999b). Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional Children, 65*, 329–337.
Thompson, B. (2002a). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64–71.
Thompson, B. (2002b). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*, 25–32.
*Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in the *Journal of Experimental Education. Journal of Experimental Education, 66*, 75–83.
*Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent *Journal of Counseling & Development* research articles. *Journal of Counseling & Development, 76*, 436–441.
Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100–116.
Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105–110.
*Vacha-Haase, T., & Ness, C. N. (1999). Statistical significance testing as it relates to practice: Use within professional psychology: Research and practice. *Professional Psychology: Research and Practice, 30*, 104–105.
*Vacha-Haase, T., & Nilsson, J. E. (1998). Statistical significance reporting: Current trends and uses in *MECD. Measurement and Evaluation in Counseling and Development, 31*, 46–57.
*Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology, 10*, 413–425.
Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*, 473–481
Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779–804.
Wang, Z., & Thompson, B. (2007). Is the Pearson r^2 biased, and if so, what is the best correction formula? *Journal of Experimental Education, 75*, 109–125.
Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanation. *American Psychologist, 54*, 594–604.
Williamson, P. R., Gamble, C., Altman, D. G., & Hutton, J. L. (2005). Outcome selection bias in meta-analysis. *Statistical Methods in Medical Research, 14*, 515–524.
*Woods, S. P., Rippeth, J. D., Conover, E., Carey, C. L., Parsons, T. D., & Troster, A. I. (2006). Statistical power of studies examining the cognitive effects of subthalamic nucleus deep brain stimulation in Parkinson's disease. *Clinical Neuropsychologist, 20*, 27–38.

Astrid Fritz is a doctoral student and lecturer in the Department of Psychology at the University of Salzburg, where she teaches philosophy of science, research methods, and psychometrics. Her research is concerned with the problematic use of null hypothesis significance testing, the prevailing research practice and publication process, and its consequences (e.g., publication bias, adaptive

sampling). Address: Department of Psychology, University of Salzburg, 5020 Salzburg, Austria. Email: astrid.fritz@sbg.ac.at

Thomas Scherndl is a graduate student and lecturer in the Department of Psychology at the University of Salzburg. He teaches courses comprising research methods, philosophy of science, and statistics. His research interests are research methods and their current application in psychological research as well as processes of predictions and decision making. Address: Department of Psychology, University of Salzburg, 5020 Salzburg, Austria. Email: thomas.scherndl@sbg.ac.at

Anton Kühberger is Associate Professor of Psychology at the University of Salzburg. His research interests include judgment and decision making; methods of process tracing; and the role of theory and simulation in mindreading. Address: Department of Psychology, University of Salzburg, 5020 Salzburg, Austria. Email: anton.kuehberger@sbg.ac.at