Methods of Psychological Research Online 1999, Vol.4, No.2
Internet: http://www.ipn.uni-kiel.de/mpr/

Confidence Intervals as an Alternative to Significance Testing

Eduard Brandstätter[1]
Johannes Kepler Universität Linz

Abstract

The article argues to replace null hypothesis significance testing by confidence intervals. Correctly interpreted, confidence intervals avoid the problems associated with null hypothesis statistical testing. Confidence intervals are formally valid, do not depend on a-priori hypotheses and do not result in trivial knowledge. The first part presents critique of null hypothesis significance testing; the second part replies to critique against confidence intervals and tries to demonstrate their superiority to null hypothesis significance testing.

Introduction

Significance testing has become a standard tool in psychological methodology. Although commonly used, it has been the focus of an increased deal of criticism and the subject of discussions (see Baril & Cannon, 1995; Cohen, 1994; Cohen, 1995; Cortina, 1997; Frick 1995; Frick, 1996; Gigerenzer, 1993; Gigerenzer, et al., 1989; Hagen 1997; Harlow, Mulaik, & Steiger, 1997; Hubbard, 1995; Kleiter, 1969; McCraw, 1995; Parker, 1995; Schmidt, 1996; Sedlmeier, 1996; Svyantek & Ekeberg, 1995). In order to avoid the problems posed by significance tests, various methods like graphic data analyses (Cohen, 1994; Tukey, 1977), meta-analyses (Schmidt, 1996), replications of studies, and confidence intervals (Cohen, 1994; Sedlmeier, 1996) have been proposed as alternatives to significance testing. Whereas negative criticism of the three former methods is seldom found in the literature, the use of confidence intervals as a substitute for significance tests has prompted both negative (e.g., Frick, 1996; Hagen, 1997) and positive criticism (e.g., Schmidt, & Hunter, 1997; Steiger & Fouladi, 1997). Critics claim that confidence intervals are subject to the same logical misinterpretation as significance tests. That is, confidence intervals are not able to solve the problems created by significance tests. On the other hand, supporters

[1] Address:
Eduard Brandstätter; Johannes-Kepler-Universität; Institut für Pädagogik und Psychologie; Tel: ++43/+732/2468/578 - Fax: /228; mail: e.brandstaetter@jk.uni-linz.ac.at, Altenbergerstr. 69; A-4040 Linz, Österreich

claim that confidence intervals do not require a priori hypotheses and therefore avoid testing of trivial null hypotheses.

This paper attempts to test the validity of the various pros and cons of confidence intervals. The question as to whether confidence intervals represent a suitable alternative to significance tests is crucial, because an improper use of confidence intervals in scientific research would cause more harm than good. On the other hand, if confidence intervals are abandoned without reason or if significance tests are abolished, the range of methods available to scientists will be restricted without due cause.

In the first section I will briefly present the most important critical arguments regarding classical significance testing (for an extensive review see e.g., Harlow et al., 1997). In the second part I will suggest confidence intervals as an alternative to significance tests. I will discuss (i) possible interpretations, (ii) criticism and misunderstandings of and (iii) arguments against confidence intervals. Furthermore I will examine (iv) the question as to whether confidence intervals should replace significance tests, and finally I would like to draw attention to (v) possible misinterpretation of effect sizes, as the centres of confidence intervals.

Criticism of Significance Tests

According to critics, significance tests (1) furnish irrelevant information and are based on (2) trivial null hypotheses. Both arguments will now be discussed in detail.

(1) Significance Tests Furnish Irrelevant Information

One of the basic principles of correct logical reasoning is the modus tollens. According to the modus tollens, the statement "If A then B" leads to the inference "If non-B then non-A". The inference "If B then A", however, is false. Consider the following example:

A person who lives in Lichtenstein also lives in Europe (if A then B).

From this it is correct to conclude:

A person who doesn't live in Europe doesn't live in Lichtenstein (if non-B then non-A).

On the other hand, the reasoning

A person who lives in Europe also lives in Lichtenstein (if B then A)

is logically false. Cohen (1994) showed that the modus tollens becomes invalid when a proven if-then sequence is replaced by a probable one. For example:

If one throws a dice, a number greater than 1 will *probably* appear (if A then probably B).

The conclusion

If the number 1 appears, one probably will not have thrown the dice (if non-B then probably non-A)

is logically false. It is precisely this pattern that has mistakenly become the most popular interpretation for significance tests (Cohen, 1994), as the following example illustrates:

If H_0 is true, this difference between sample means (significance) is unlikely.

This difference has occurred.

Therefore H_0 is probably false.

Recall, the alpha error represents – correctly interpreted – the probability to obtain a specific or more extreme event (for instance, an observed or more extreme difference between two sample means), given that H_0 is true; that is, p (event / H_0)[2]. Significance tests provide no information about the probability of H_0. Strictly speaking, significance tests do not test hypotheses. In summary, significance tests provide information on the probability of finding a specific or more extreme event when the null hypothesis is true. Significance tests say nothing about the probability of a null hypothesis being true. However, according to the next point of criticism, the impossibility to know the probability of the null hypothesis being true, given the event – p(H_0 / event) – does not bear much weight, because the null hypothesis only reveals trivial information.

(2) Significance Tests Are Based on Trivial Null Hypotheses

According to this argument, null hypotheses are irrelevant because differences between two means nearly always exist. A study conducted by Bakan (1966) clearly illustrates this point. Bakan categorised 60,000 persons according to random criteria, for example, whether they lived to the east or the west of the Mississippi River and found significant differences for all the questions included in his questionnaire. Tukey (1991) summarised this criticism: "It is foolish to ask 'Are the effects of A and B different?' They are always different - for some decimal place" (p. 100). With the same lack of ambiguity, Nunnally (1960, cited in Gigerenzer et al., 1989, p. 210) stated: "If the null hypothesis is not rejected, it usually is because the *N* [sample size] is too small. If enough data is gathered, the hypothesis will generally be rejected. If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data." In sum, this criticism goes even one step further. Consequently, it does not matter whether significance tests are interpreted correctly or incorrectly, as the H_0 is of little scientific interest. Abandoning significance tests would therefore result in no loss.

To counter the criticism that significance tests are based on trivial null hypotheses, other ways of significance testing have been suggested. Most prominently, the testing of an al-

[2] The term "event" refers to the obtained and to the more extreme outcomes.

ternative hypothesis H_1 alongside a conventional null hypothesis H_0. However, this broadened concept of significance testing is also fraught with problems.

Recall, the conventional and most popular two-tailed significance test is only based on a single null hypothesis H_0, and the alternative hypothesis H_1 subsumes the rest of all the other possible hypotheses. If an additional, precisely specified, alternative hypothesis H_1 is included, one may calculate the power of a test (1-β) as well as the sample size required. From this broadened concept, however, two major problems arise: First, in actual psychological research it is rarely possible to state and justify a specific alternative hypothesis H_1. For example, it may be hard to justify a-priori why the alternative hypothesis for a correlation coefficient ρ equals .6 and not $\rho = .5$. But even when estimates of ρ can be derived from previous studies, a certain amount of discretion is needed.

Secondly and more importantly, the broadened concept of significance testing again reveals no information concerning the validity of the hypotheses H_0 and H_1. Consequently, this broadened conceptualisation is again prone to the same misinterpretation as the classical significance test, which is based on a single null hypothesis (see violation of modus tollens). In summary, testing a specific alternative hypothesis H_1 in addition to the conventional null hypothesis H_0 is fraught with two problems: First, specifying a single alternative hypothesis H_1 may be hard in practice and second, this broadened concept of testing two instead of one hypotheses reveals no information about the validity of both hypotheses. The next section attempts to offer one possible solution to these problems.

Confidence Intervals as an Alternative

The Interpretation of Confidence Intervals

Cohen (1994) proposed confidence intervals as an alternative to the above problems. But what do confidence intervals actually mean with regard to – for example – the difference between two population means? If a study is replicated an indefinite number of times and if a 95% confidence interval is computed for the difference of sample means each time, the true difference of the population means will fall within these intervals in 95% of all replications (Bleymüller, Gehlert, Gülicher, 1988; Cohen, 1995). Consequently, the true difference of population means will fall outside these intervals in 5% of all replications. Thus confidence intervals are random variables and the width and location of these confidence intervals vary from replication to replication.

Proceeding with this example at hand, the calculation of a confidence interval for the difference between two population means ($\mu_1 - \mu_2$) is straightforward and simple. Suppose a researcher draws two independent samples of sufficient size (n_1, $n_2 > 30$) and calculates

the sample means $\bar{x}_1$, $\bar{x}_2$ and standard deviations s_1 and s_2. The 95% interval for the true difference of the population means therefore is:

$$I = [(\bar{x}_1 - \bar{x}_2) - 1.96*\sigma_{diff}, (\bar{x}_1 - \bar{x}_2) + 1.96*\sigma_{diff}], \text{ or}$$

$$P[(\bar{x}_1 - \bar{x}_2) - 1.96*\sigma_{diff} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + 1.96*\sigma_{diff}] = .95$$

$$\text{with } \sigma_{diff} = (\sigma_1^2 / n_1 + \sigma_2^2 / n_2)^{1/2};$$

σ_1 and σ_2 are estimated on the basis of the standard deviations s_1 and s_2. The value of 1 - α is called the confidence coefficient and α the confidence level.[3]

However, researchers are generally not so much interested in the proportion of confidence intervals which comprise the true population parameter. Instead they are usually interested in the probability of finding the true population parameter within the calculated confidence interval. In other words, a researcher is interested in the probability that the true population parameter is part of a single, specific confidence interval, which has been calculated from sample data.

At this point our researcher is unfortunately let down, because there are two schools of thought concerning the assignment of a probabilities to a *single* event. Whereas the "classic theory of probability" (frequentists) only assigns the probabilities 0 and 1 to a single event (Kendall & Stuart, 1979; Mulaik, Raju & Harshman, 1997), the "subjective theory of probability" (DeFinetti, 1971; Wright & Ayton, 1994) permits the assignment of all possible probabilities between 0 and 1 (see Reichardt & Gollob, 1997). The classic theory of probability defines the term probability on the basis of repeatable events. Probability then is the (asymptotic) relative frequency of an event, repeated infinitely under identical conditions varying only by chance (Reichardt & Gollob, 1997). Accordingly, there are only two designated values which make sense. Either a single confidence interval contains the true population parameter ($\mu_1 - \mu_2$) or not. Thus, the probability that a specific confidence interval contains the true population parameter ($\mu_1 - \mu_2$) is either only 0 or 1.

---

[3] In contrast to confidence intervals which are usually computed around sample data, confidence intervals can also be computed around population parameters (see Menges, 1969; Witte, 1980), i. e.:
$I = [(\mu_1 - \mu_2) - 1.96*\sigma_{diff}, (\mu_1 - \mu_2) + 1.96*\sigma_{diff}]$, with
$P[(\mu_1 - \mu_2) - 1.96*\sigma_{diff} \leq (\bar{x}_1 - \bar{x}_2) \leq (\mu_1 - \mu_2) + 1.96*\sigma_{diff}] = .95$
for the difference between two means. If a difference between two sample means ($x_1 - x_2$) falls outside the above interval, then the data deviate significantly from the hypothesis ($\mu_1 - \mu_2$ as the true population parameter). This strategy is identical to that used for significance testing. Confidence intervals around population parameters are fixed (when σ_1 and σ_2 are known for calculating σ_{diff}), whereas confidence intervals around sample data are random variables. However, because computerisation allows the exact calculation of *p*-values nowadays, confidence intervals around population parameters have lost influence.

In contrast, the interpretation of the subjective theory of probability is more liberal[4]. If the true population parameter falls within the calculated confidence interval in 95% of all samples, then the probability *p*, that the true population parameter falls within the 95% confidence interval of the specific sample taken, is also .95; the true population parameter falls outside the confidence interval with $p = .05$. Therefore, the probability for a single event may have any value between 0 and 1.

This is not the right time and place to discuss the different viewpoints between frequentists and subjectivists (for a discussion see Stegmüller, 1973). One should only be aware that an interpretation like: "The true population parameter falls within this (one) confidence interval with probability $p = .95$" only represents one school of thought (subjective theory of probability). Another possible interpretation (classic theory of probability) also exists.

The calculation of confidence interval as stated above is valid for the most common case, when there is no precise information regarding the prior distribution of the population parameter (Reichardt & Gollob, 1997) while at the same time, a uniform distribution can be ruled out (*no usable* prior distribution). If previous knowledge of the distribution for a population parameter exists (*nonuniform* prior distribution), Bayes theorem can be applied (see Edwards, Lindmann, Savage, 1963; Kleiter, 1980; Winkler, 1972); this is both true for conclusions which are based on a confidence interval and on a significance test.

Finally, confidence intervals offer more information than significance tests. A confidence interval reveals, how precisely a population parameter can be estimated (accuracy of estimation). Wider intervals permit less accurate estimations than smaller intervals. Significance tests, on the other hand, do not permit this estimation. One can only tell that the probability, that this or a more extreme event has occurred, given H_0, equals alpha. Therefore, based on the subjective theory of probability, confidence intervals provide more information than significance tests. After discussing possible interpretations of confidence intervals, we will now turn to criticism of and misunderstandings caused by confidence intervals.

Criticism and Misunderstandings

As mentioned above, confidence intervals have also received criticism. Accordingly, it is not logical to reject significance tests and, at the same time, recommend confidence intervals as an alternative:

"... a confidence interval can function to indicate which values could not be rejected by a two-tailed test with alpha at .05. In this function, the confidence interval could replace the

[4] Both the classic and the subjective theories of probability acknowledge Bayes' theorem. Therefore, I will not equate the subjective theory of probability with Bayes statistics (see also Reichardt & Gollob, 1997).

report of null hypothesis for just one value, instead of communicating the outcome of the tests of all values as null hypotheses ... Cohen (1994) was illogical when he criticized the logic of null hypothesis testing and then advocated using the confidence interval because it reported the results of all statistical tests" (Frick, 1996, p. 383).

A similar statement was made by Hagen (1997, p. 22): "We cannot escape the logic of NHST [null hypothesis statistical testing] by turning to point estimates and confidence intervals".

This criticism appears to be false and unfounded. The reason why it appears false is that confidence intervals can be interpreted as significance tests, however, they do not have to be, as Schmidt and Hunter (1997) indicated: "The assumption underlying this objection is that because confidence intervals *can* be interpreted as significance tests, they *must* be so interpreted. But this is a false assumption" (p. 50). Confidence intervals, however, comprise specific information concealed by significance tests.

Confidence intervals clearly show how exact the estimate of a population parameter will turn out to be, with small confidence intervals permitting more exact estimates than larger ones. Effect sizes provide valuable information on the magnitude of the effect of interest. It is precisely this information, such as the exactness of parameter estimates and the size of the effect of interest, that is concealed by *p*-values.

Confidence intervals are easier to understand than significance tests and therefore have a definite instructional advantage over significance tests. Anyone who has taught statistics is familiar with the fact that students come to understand confidence intervals much quicker than significance tests. If a confidence interval includes the value 0, it is not possible to predict with any great degree of certainty the direction of the effect: The effect can be positive, negative and, at least theoretically, null. If the value 0 falls outside the 95% confidence interval, then one knows the sign of the most likely (95% confidence) population parameters. Every student clearly understands this at once. The logic of significance tests is all the more twisted: Assuming H_0, the probability of finding this or a more extreme event is equal to *p*. And now? Even though confidence intervals can be interpreted as significance tests, there is little reason to do so (Schmidt & Hunter, 1997). In summary, the above mentioned argument that confidence intervals must be interpreted as significance tests is erroneous and misses the point.

In contrast to significance tests, confidence intervals reveal the precision of parameter estimates and it is of no particular interest whether an effect is zero or not. As previously mentioned, confidence intervals comprise more information than significance tests. It is precisely this argument, that significance tests might be more economical in many situations because they contain less information, that is mentioned by many authors as an advantage for significance tests (see Schmidt & Hunter, 1997). Often exact information from a point estimate including the confidence interval is not needed and a quick probing of sig-

nificant results would suffice, at least for the initial orientation. Who isn't familiar with the probing of significant values in order to quickly and comfortably find one's way about in a large correlation matrix?

Schmidt and Hunter (1997) disagree with this strategy and support the use of effect sizes (*eta*, Cohen's *d*, *r*) instead of *p*-values for quick orientation. It is true that there is a prefect relationship between *p*-values and effect sizes when the sample size is constant. Most of the time, however, sample sizes differ within a study and, to an even greater extent, between studies. In such cases *p*-values provide more trivial information than effect sizes and misinterpretation is often the result: For instance, a significant correlation coefficient $r = .1$ ($N = 1{,}000$) may not be significantly "confirmed" in another study $r = .4$ ($n = 30$) thus producing conflicting results.

Concentrating on effect sizes avoids this kind of fallacy. Both studies found a positive relationship and thus support (or don't support) hypothesis X. The second study ($r = .4$) does not contradict but support the first. This information is adequate for an initial survey and confidence intervals then provide additional information, taking the different sample sizes into consideration. In summary, effect sizes are more suitable for quick orientation than *p*-values (Schmidt & Hunter, 1997). Considering the great number of advantages that confidence intervals have over significance tests, the question as to why confidence intervals are still considered inferior in the representation of statistical events arises. The next section attempts to answer this question.

Arguments Against Confidence Intervals

Several assumptions can be made to explain why confidence intervals are still relatively seldomly used today (see Reichardt & Gollob, 1997; Steiger & Fouladi, 1997): Inadequate availability of confidence intervals in software packages, the need to carry out significance tests in order to get published in scientific journals, the often small effect sizes which are concealed by emphasising a "significant result" and the fact that confidence intervals are often very broad and hence allow inexact estimates – are just a few of the most important reasons. The heuristic that a large number of people will probably not be mistaken (social proof) may be another reason for the widespread use of significance tests. Heuristics, like those of social proof often deliver good, prompt predictions, yet they sometimes lead astray (Tversky & Kahneman, 1974).

Should Confidence Intervals Replace Significance Tests?

Following a debate about the usefulness of confidence intervals (see above), in the meantime some consensus has been reached. In the recently edited volume entitled "What if there were no significance tests?" (Harlow et al., 1997), all authors - even supporters of significance tests – recommended the use of confidence intervals (Harlow, 1997). The im-

portant question that arises no longer asks whether confidence intervals should be reported or not, but whether there is enough room for both confidence intervals and significance tests, or whether confidence intervals should replace significance tests. The latter is the more extreme standpoint, because it implies the abolishment of significance tests altogether. The former standpoint implies looking for conditions under which it is better to use confidence intervals instead of significance tests. Hunter and Schmidt (1997), supporters of the latter view which favours the replacement of significance tests by confidence intervals, generally doubt the contributions significance testing has made to the development of cumulative scientific knowledge. Abelson (1997), in response, lists two situations, where significance testing contributed to psychology as a science: First, decisive experiments between two rival theories, and second, testing the congruence of a model with empirical data ("Goodness-of-Fit" testing). However, both of these problems – as I will proceed to demonstrate - can be better managed without significance tests.

Let us examine the situation with significance tests as a decisive aid between two rival theories first. For example, theory A predicts a positive and theory B a negative correlation for an experiment. In this case, a statistically significant correlation in any direction that supports one of the two theories would suffice: a statistically significant positive correlation would support theory A, a statistically significant negative correlation, theory B. Generally speaking, the magnitude of the effect is of little interest because one is only concerned with deciding between two theories.

The critical arguments I initially mentioned, like the fact that significance tests provide no information about the probability of the null hypothesis or the triviality of the null hypothesis, still apply to significance tests. The confidence interval, however, allows to make a decision between the two rival theories without these shortcomings: A decision in favour of theory A ($r > 0$) or of theory B ($r < 0$) is possible as long as the confidence interval rules out the value zero. Therefore, there is no reason to use significance tests instead of confidence intervals to decide between two theories. The direction of the corresponding effect size (r, d, etc.) and its confidence interval may be able to support a decision in favour of one of the theories.

The second possible advantage of significance testing deals with the congruency of a theoretically postulated model with empirical data. Structural equation modelling (see for instance Jöreskog & Sörbom, 1989) may serve as a good example. One first specifies a model and then tests whether the data significantly deviate from the theoretically postulated model. By doing so, it is possible to compare different models with each other, and usually the model with the highest p-value is chosen. For this purpose the researcher is interested in the null hypothesis and not in the alternative hypothesis. In principle, this type of model testing is based on the same kind of logic as a simple t-test. If the null hypothesis for a t-test states that two sample means come from the same population, then

within structural equation modelling the null hypothesis equals the specified model. However, the same problems still exist. Large samples lead to maximum power for each test and indicate small deviations from the theoretically postulated model; on the other hand, small samples lead to weak power for each test and are therefore unable to differentiate between different models.

Steiger and Fouladi (1997) present an excellent summary on alternative methods (e.g. Goodness-of-Fit-Index (GFI) or Root-Mean-Square-Error-of-Approximation (RMSEA) for structural equation models), which are not based on significance tests and hence avoid the problems of significance testing. So, alternative, better indices, which are not based on the calculation of *p*-values, are able to replace significance tests for the purpose of model testing. In summary, both possible advantages of significance testing listed by Abelson (1997) can be replaced by better methods.

Possible Misinterpretation of Effect Sizes

In the preceding discussion I have mentioned some advantages of effect sizes and of confidence intervals compared to significance tests. Despite these advantages, effect sizes may also be misinterpreted, which seems particularly important for experimental designs. Effect sizes rely heavily on the manipulation of independent variables, but there is no direct measure of the strength of manipulation for the independent variables (Ronis, 1981). For example, an experimenter presents subjects with hypothetical scenarios (vignettes) that are embedded in a 2x2 ANOVA design. Imagine in the vignette two persons who communicate with each other and the experimenter manipulates the emotional relationship between these two persons. Therefore, the first factor "emotional relationship" has two conditions, namely "positive" and "negative" emotional relationship. It then will make a great difference whether a positive emotional relationship is operationalised as "best friends on earth" or as "friends who like each other". Depending on the degree of the manipulation, the effect sizes for the first factor "emotional relationship" turn out differently. Given a strong manipulation ("best friends on earth"), the main effect for the second factor and the interaction effect will probably be weak, because the first factor may explain most of the variance of the dependent variable. On the other hand, a weak manipulation ("friends who like each other") might increase the importance of the second main as well as the interaction effect. The same reasoning holds true for the negative relationship condition. Thus effect sizes must always be interpreted relative to the size of the manipulation of the independent variables. This is less a problem for independent variables like gender, age or personality variables, which exist in natural variation. In summary, effect sizes and confidence intervals represent an improvement over significance tests; however, they must also be carefully interpreted, as effect sizes depend on the magnitude of the experimental manipulation. It follows that psychological experiments can often only show psy-

chological *principles* or *mechanisms* and statements like "effect X exists, but is small in value" bear little weight. These problems also apply to meta-analyses.

Conclusion

Confidence intervals avoid the problems of classic significance tests. They do not require a-priori hypotheses, nor do they test trivial hypotheses. Confidence intervals comprise the information of a significance test and are considerably easier to understand, which results in their didactic superiority.

When interpreted with regard to the subjective theory of probability, confidence intervals provide information about the probability of the sign of an effect. If null falls outside the 95% confidence interval, one knows the sign of the most likely (95% confidence) population parameters. If null falls within the 95% confidence interval, nothing can be said about the sign of an effect with any great degree of certainty. The effect can be positive, negative and, at least theoretically, null. The latter is extremely unlikely. It can also be assumed that the effect might be very small (close to zero).

The question as to whether significance tests should replace confidence intervals or not can be answered with a guarded "yes". Confidence intervals contain the information of a significance test, therefore there is no loss of information and no risk involved when confidence intervals replace significance tests. Taken together, confidence intervals in addition to replications, graphic illustrations and meta-analyses seem to represent a methodically superior alternative to significance tests. Hence, in the long run, confidence intervals appear to promise a more fruitful avenue for scientific research.

References

[1] Abelson, R. P. (1997). A retrospective on the singnificance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (S. 117-141). Hillsdale: Lawrence Erlbaum.

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

[3] Baril, G. L. & Cannon, J. T. (1995). What is the probability that null hypothesis testing is meaningless? *American Psychologist, 50,* 1098-1099.

[4] Bleymüller, J., Gehlert, G. & Gülicher, H. (1988). *Statistik für Wirtschaftswissenschaften* (5. Aufl.). München: Vahlen.

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

[6] Cohen, J. (1995). The earth is round ($p < .05$): Rejoinder. *American Psychologist, 50,* 1103.

[7] Cortina, J. M. & Dunlap, W. P. (1997). On the logic and the purpose of significance testing. *Psychological Methods, 2,* 161-172.

[8] DeFinetti, B. (1971). *Theory of probability: A crititcal introductory treatment* (Vol. 1). New York: Wiley.

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

[10] Frick, R. W. (1995). A problem with confidence intervals. *American Psychologist, 50,* 1102-1103.

[11] Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1* 379-390).

[12] 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. Methodolocical issues* (pp. 311-339). Hillsdale: Lawrence Erlbaum.

[13] Gigerenzer, G.; Swijtink, Z.; Porter, T.; Daston, L.; Beatty, J. & Krüger, L. (1989). *The empire of chance*. Cambridge: Cambridge University Press.

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

[15] 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?* (S. 1-17). Hillsdale: Lawrence Erlbaum.

[16] Harlow, L. L; Mulaik, S. A. & Steiger, J. H. (Eds.), *What if there were no significance tests?* Hillsdale: Lawrence Erlbaum.

[17] Hubbard, M. (1995). The earth is highly significantly round ($p < .0001$). *American Psychologist, 50,* 1098.

[18] Jöreskog, K. G. & Sörbom, D. (1989). LISREL 7. *A guide to the program and applications* (2nd ed.). Chicago, IL: SPSS.

[19] Kendall, M. & Stewart, A. (1979). *The advanced theory of statistics. Vol. 2. Inference and relationship*. London: Charles Griffin & Co.

[20] Kleiter, G. D. (1969). Krise der Signifikanztests in der Psychologie. *Jahrbuch für Psychologie, Psychotherapie und medizinische Anthropologie, 17,* 144-163.

[21] Kleiter, G. D. (1980). *Bayes Statistik: Grundlagen und Anwendungen*. Berlin: De-Gruyter.

[22] McCraw, K. O. (1995). Determining false alarm rates in null hypothesis testing research. *American Psychologist, 50,* 1099-1100.

[23] Menges, G. (1968). *Grundriß der Statistik. Teil 1: Theorie*. Köln: Westdeutscher Verlag.

[24] Mulaik, S. A., Raju, N. S. & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (S. 65-115). Hillsdale: Lawrence Erlbaum.

[25] Parker, S. (1995). The "difference of means" may not be the "effect size". *American Psychologist, 50,* 1101-1102.

[26] Reichardt, C. S. & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical significance tests, and vice versa. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (S. 259-284). Hillsdale: Lawrence Erlbaum.

[27] Ronis, D. L. (1981). Comparing the magnitide of effects in ANOVA designs. *Educational and Psychological Measurement, 41,* 993-1000.

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

[29] 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?* (S. 37-64). Hillsdale: Lawrence Erlbaum.

[30] Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research Online, 1*, 41-63.

[31] Steiger, J. H. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (S. 221-257). Hillsdale: Lawrence Erlbaum.

[32] Stegmüller, W. (1973). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie IV: Personelle und statistische Wahrscheinlichkeit*. Berlin: Springer.

[33] Svyantek, D. J. & Ekeberg, S. E. (1995). The earth is round (So we can probably get there from here). *American Psychologist, 50*, 1101.

[34] Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

[35] Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

[36] Tversky, A & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

[37] Winkler, R. L. (1972). *Introduction to Bayesian Inference and Decision*. New York: Holt, Rinehart & Winston.

[38] Witte, E. H. (1980). *Signifikanztest und statistische Inferenz*. Stuttgart: Enke.

[39] Wright , A. & Ayton, P. (1994). *Subjective probability*. Chichester: Wiley.