Effect sizes:
Statistical, practical and theoretical
Significance of empirical findings

George Lind [1]

2021 [2]

"A picture is more worth than thousands of p-values: On the irrelevance of hypothesis testing in the computer age." (Loftus, G. R., 1993)

"What's wrong with significance testing? Well, among many other things, it does not tell us what we want to know, and, out of desperation, we nevertheless believe in that it does!" (Cohen, 1994, p. 997)

"For the meaningful use of inferential statistics, it is necessary that a theoretically well-founded hypothesis or question has been formulated prior to the start of the investigation." (Bortz, 1994, p. 2)

"What we should teach our students is statistical thinking: how to formulate bold hypotheses, derive precise alternative hypotheses, design experiments to minimize real measurement error (rather than just measuring it and putting it in the F-break), analyze data for each individual separately rather than automatically combining them into means, and use meaningful descriptive ratios (statistics) and exploratory data analysis" (Gigerenzer, 1998, p. 200; my translation).

"When productivity analyses are foregone or poorly carried out, the result can be the promotion of policies with little if any evidentiary support. The result can also be the promotion of a large-scale implementation of unproven strategies equally likely to do harm as to do good." (Baker & Welner, 2012, p. 99)

Contact: apl. Prof. em. Dr. Georg Lind, e-mail: georg.lind@moralcompetence.net.

2This paper grew out of my dissertation: https://moralcompetence.net/pdf/Lind-1985_Inhalt-und-Struktur.pdf
and the lecture I gave at the Catholic University of Eichstätt in 1991 as part of my habilitation:
https://moralcompetence.net/pdf/Lind-1991_Empirischer-Gehalt-von-Hypothesen.pdf.
The first version of this paper has since undergone several revisions and additions. Major corrections: additions to absolute ES (27.10.2008). Editorial additions (11/16/2009); correction of typos on page 4 (6/29/2010). Change of links to new URL: http://moralcomptence.net
1

[2] Textbooks of statistics usually deal only with the case where the variances of the dependent variables are different between categories (homoscedasticity). The much more important problem of different variances of the independent variables seems to be hardly treated, since one is obviously more interested in inferring populations from sample data than in providing suitable measures for the evaluation of causal hypotheses or treatment effects.

"Exploratory studies, on the other hand, give the experimenters considerably more freedom, which is why their results are always taken with a grain of salt as a matter of principle. The fact that someone in a publication summarily re-declared such a study to be a hypothesis test is usually only noticed when someone else repeats the experiment." (Nuzzo 2014)

How big does an effect have to be to be significant?

To answer these questions, one can consider several possibilities. In psycho logy and many social sciences, a formal, statistical answer is considered almost exclusively: Is the finding "statistically significant"? What is rarely (far too rarely!) considered, however, is the possibility of examining findings for their theoretical, substantive significance: What difference in values is significant for our subjective feelings and actions? At what effect size can we speak of a therapy method or an educational intervention really bringing something and being worth the effort that all participants have to invest? Does the effect always occur or only under certain conditions? Is it tied to specifics of the study (sample size, dispersion of independent variables)? Does the effect fit with what we already know about the variables we are studying, or does it challenge well-founded theories?

In short, even if we assume for once that the determination of statistical significance ("significance") is carried out correctly, this procedure does not relieve the scientist and the practitioner who wants to interpret empirical data and make them the basis of decisions for action from the question of whether the finding is also significant in terms of content, in terms of subject matter. If we read from the thermometer that the average temperature has risen five degrees today compared to yesterday, we can ask whether this difference is statistically significant (a question we can answer, of course, only if we have taken several measurements at each of the two measurement points). In fact, in everyday life we are hardly interested in this question, but perhaps in the question of whether five degrees of difference necessitates certain behavioral decisions, such as putting on slightly less warm clothing. In this sense, we are also less interested in whether two countries in which a few thousand students participated in school achievement tests differ "significantly," but perhaps more interested in whether these differences are so large that they noticeably affect the economic power of the countries or the quality of life of the people who live in them. The question of statistical significance is more of interest to the researcher who is considering how large a study group must be in order for a particular difference that is important to him (for substantive reasons!) to be demonstrated beyond a reasonable doubt, that is, to be larger than the precision of his measurement instrument allows. Thus, if a difference has not become statistically "significant," it means that it is smaller than the difference that was determined to be significant. (However, this substantive determination is not to be confused with the so-called "significance level", e.g. $\alpha = 5\%$!)

It is therefore highly problematic to infer directly from a *statistical* significance of findings their *theoretical* and *practical* relevance. So-called measures of "practical significance" or "relative effect size" are more suitable for this purpose, although they are also purely statistical procedures that cannot replace a theoretical and practical evaluation of a finding. Depending on the question and the context of application, even very low effect sizes can be of great practical importance. For example, in medicine, drugs are approved when their effect size is relatively small (r = 0.15) if they can be used against particularly serious diseases. In the case of absolute effect strengths, the observation of very small effect strengths (as found, for example, in the deflection of light by large masses) can have a very great practical significance, for example, when it comes to calculating the exact launch direction of a Mars probe.

This paper presents two important alternatives to the questionable concept of “statistical significance”: (1) The concept of *relative effect size* and the formula for calculating it from conventional measures of statistical significance. (2) Because this concept is not optimal, I also discuss the concept of *absolute effect size* and in order to determine the practical and theoretical significance of empirical and experimental findings.

An example

Let's take an example: We want to know whether a certain teaching method is suitable for improving moral judgment. Before and after using the method, students are measured with the Moral Competence Test (MCT; Lind, 2019). The following mean differences and standard deviations are obtained:

M1 = 25.5, s = 12.1

$_{M2}$ = 31.7, s = 12.1

Difference = M1 - $_{M2}$ = 31.7 - 25.5 = 6.2 (C points).

N1 = 15, N2 = 15

If, as in many studies, one calculates only statistical significance (here, for simplicity, the t-test for independent samples is used, although of course a t-test for dependent samples would be correct), then the answer to the question of whether this is "significant" is the following:

$$= 6{,}2 / 7{,}1 = 0{,}87$$

This t-value has a p = 0.086, i.e., the value increase of 6.2 points in the MCT is "not significant at the 5 percent level." For this, *p* would have had to be equal to or less than *p* = 0.05 (α-level).

However, the experimenter thinks that a C-value difference of 6.2 points is a relatively large amount. Why did the difference not become significant? The answer is clear: the standard deviation is relatively large and the sample size relatively small. The test variable for statistical significance, i.e. the probability associated with the *t-value* (*p-value*), becomes all the larger (i.e. rather "significant"),

a) the larger the *t-value* (which in turn becomes larger *the smaller the standard deviation* S_e of the means), and

b) the larger the sample size N is.

Yes, says the experimenter to himself, now I know how to get a "significant result" next time. I increase the sample size and take 50 instead of 15 students *per group*. The fictitious experiment led to the following result:

$M_1 = 25.5$ $s = 12.1$

$M_2 = 31.7$ $s = 12.1$

Difference = $M_2 - M_1 = 31.7 - 25.5 = 6.2$ (C points)

N1 = 50, N2 = 50

In fact, the same difference becomes "highly significant": $p = 0.006 | < 0.01$ (α-level, two-sided).

But, the experimenter may object, increasing the number of subjects costs a lot of money and time. I could also try to make the scatter of C-values in my experimental group smaller by removing persons with very small and very large C-values from the experiment (elimination of so-called "outliers"). No sooner said than done. This measure, sometimes considered quite legitimate, reduces the standard deviation to s = 6.2.

$M_1 = 25.5$, s = 6.2

$M_2 = 31.7$, s = 6.2

Difference = $M_2 - M_1 = 31.7 - 25.5 = 6.2$ (C points)

N1 = 15, N2 = 15

And this result is also "significant": $p = 0.0106 | < 0.05$ (α-level).

As we have seen, the effect of the educational intervention has actually always remained the same: the test scores have grown exactly 6.2 points. But the "significance" of the result has changed each time. When the sample was increased or the spread of scores was decreased, a "non-significant" result became a "significant" one. One could also say that while the psychological significance remained the same,

only the statistical one changed. This can only surprise someone who does not know what statistical significance tests are actually good for.

Another example: How increasing the sample size changes the "significance" without changing anything in the effect size.

N = 26

The following relationship exists between variables (A and B):

	A = 0	A = 1
B = 0	10	5
B = 1	5	6

Correlation: Phi = **0.21;** statistical significance: n.s. (**p = 0.279**)

N = 260

The following relationship exists between variables (A and B):

	A = 0	A = 1
B = 0	100	50
B = 1	50	60

Correlation: Phi = **0.21**; statistical significance: sig. 1% level (**p = 0.00063**)

Why do we need statistical significance tests?

Statistical significance tests are used to determine the precision of a sample result with respect to a well-defined population and a threshold (based on prior content knowledge) for the theoretical significance of a difference score (see, among others, Hays, 1963). This is the case, for example, when we want to investigate the question of whether the test scores "moral judgment" differ in content significance between the populations of all 15-year-old Germans and French and we want to draw only a random sample from each population instead of the whole population. What constitutes a difference that is substantively or practically significant must be determined here *prior to the* study so that it can be determined how large the sample must be to measure this difference with sufficient precision. Statistical significance tests, on the other hand, are out of place a) if no population is defined at all (or can be

defined) b) if no true random sample has been drawn from such a well-defined (!) population, and if it is not clear which difference value is really significant. These conditions are often not given in the social sciences, which is why the question about the practical and theoretical significance of a difference, which should actually be clarified first, often comes last or is completely ignored, because one erroneously thinks that the statistical significance would give us information about it.
For these reasons, Thompson (1994), one of the leading experts in the field of social science statistics, recommends the following language, which I find very helpful:

"Overcoming three language habits can help prevent unconscious misinterpretation:

* Always speak of 'statistical significance' and not simply 'significance'. This can help break the erroneous association between the rejection of a null hypothesis and an important finding.

* Don't talk about things like "my result approaches statistical significance." Such language makes little sense in the context of statistical test logic.

* Don't talk about things like 'the statistical significance test tells us whether the results were random.' This language creates the impression that statistical significance tests tell us something about the replicability of a finding." (Thompson, 1994)

Even the introduction of "power analysis" (Cohen, 1988) does little to change this misuse of statistical significance testing. True, it corrects the bias of the usual use of significance tests toward the null hypothesis (which they always try to disprove, as if that were an important goal in research). But even this analysis has little to do with the practical or theoretical significance of research findings.

As an alternative to statistical significance tests, measures of "practical significance" (Bredenkamp, 1970) or "effect size" have been proposed. These measures have great advantages over the pure use of statistical significance tests, but they are not optimal either, since - despite the appearance of the term "practical significance" - they say nothing about either practical or theoretical significance. I therefore prefer to use the term effect size, which has also gained acceptance in the literature.

Measures of relative effect size

Effect size measures are constructed to be independent of sample size, thus eliminating an important drawback of statistical significance measures. In other words, the effect size of an intervention is independent of how many subjects it was tested on. Thus, these measures also make studies comparable that used different sized samples. However, these measures still depend heavily on the spread of the

values of the dependent variable in the respective study group, which is why we should also refer to them as *relative* effect size measures.

Two measures are most commonly used, the so-called *d-value* and the correlation coefficient *r*. The *d-value* was proposed by Glass et al. (1978), who used it in meta-analyses, where it is still widely used today. It is defined as the difference of means in relation to the "common" (pooled) standard deviation (common SD from first and second series of measurements):

$$d = \frac{M_2 - M_1}{SD_{pooled}}$$

This formula also shows the dependence of this measure on the scatter of the values in the sample. The *d-value* is not limited downwards and upwards, which is why it is often difficult to estimate.

This shortcoming is remedied by the (non-linear) correlation coefficient *r,* familiar to most empirical social scientists*, which* can vary from -1.0 to +1.0. At +1 there is a maximum effect size, at 0 there is absolutely none, and at -1 there is a maximum negative effect size. The measure is much used by many today to express effect strength (Thompson 1996). It is not limited - in contrast to the so-called correlation research - to expressing the linear relationship between an intervention and the measured values found, but also to represent other, non-linear relationships. The correlation with the measure *d is* easily established via the following formula (Cohen, 1988, p. 23):

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

For small samples and for unequal sample sizes, the somewhat more complicated formula of Aaron, Kromrey, and Ferron (1998) is suggested:

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

In the appendix, further formulas are given to convert significance values (*Chi2*, *t, F,* etc.) into the correlation coefficient *r* as a measure of effect size, provided that the necessary information (especially the sample size N*) is* available (see appendix).

Measures of relative statistical effect size are better than those of statistical significance because they do not depend on sample size. This makes it easier to compare studies that used different sample sizes. This advantage is exploited in meta-analyses.

However, relative effect size measures are also not suitable for comparing effect sizes:

1. the problem that the scatter of the dependent variables often differs greatly from study to study.[3] This severely limits comparability (which, however, does not prevent many authors from making comparisons and calculating meta-analyses because they are not even aware of the prerequisites of this method or they think that this does not have a great influence on the result, which, however, would have to be proven in each case). Different relative effect sizes (*d* or *r) do* not necessarily indicate different effects. They can also be influenced by uncontrolled or uncontrollable or deliberately induced differences in the spread of the values of the dependent variables.

2. The widely used relative effect size measures "d" and "r" optimally represent only *symmetric* and *mono-causal* relationship. An empirical relationship is symmetric if it is equal in both directions (from the independent variable to the dependent variable and vice versa). This is usually the case only if the independent variable (e.g., schooling) is both a *necessary*and a *sufficient* condition for the dependent variable (e.g., moral judgment). This would mean, for example, that high moral judgment should never occur if education is low (since education is considered *necessary*), and that it does not require any other condition, so that it alone can trigger the effect (since education is considered *sufficient).* One can see that these assumptions are rather unrealistic, since relationship between variables are mostly asymmetric and multi-causal or conditioned by other variables.

3. This objection can be partly taken into account by using the statistical method of *analysis of variance* (in the case of experimental designs) or the method of *regression analysis (in* the case of field studies) to calculate the relative effect size. In regression analysis, the ß coefficient (ß = beta) indicates the number of units by which the dependent variable (in the example, moral judgment) increases when the independent variable (in the example, the amount of education) is increased by one unit. However, even these measures are not without problems, since they presuppose that the independent variable is a sufficient condition for the effect (i.e., they do not represent conditional effects or so-called "interaction effects") and because they also depend on the dispersion of the values of the independent variable. In the analysis of variance, the proportion of variance between groups can be inter preted in a similar way.
4. Another problem with all measures of *relative effect size* arises when one wants to compare the effect of several "independent"[4] variables with each other, since these independent variables can

"Independent" does not refer to the relationship of these variables or factors to the "dependent" variable, as some publications erroneously state, but to the statistical relationship of the factors to each other. By independent and dependent is also initially meant only a statistical relationship. The question whether this statistical relationship corresponds to a real

correlate with each other. In order to obtain unambiguous values, these correlations must be "controlled", i.e., brought to zero. This can be achieved in two ways, firstly by designing the study as an *experiment*, i.e. by making the independent variables statistically independent ("orthogonalized") by means of a suitable experimental design. For example, if the influences of gender and education on moral development are to be compared, one can neutralize the possible correlation between gender and education by including exactly the same number of cases for each combination of these variables, as illustrated in the following table:

	Men	Women
Without high school diploma	20	20
With high school diploma	20	20

Here, the correlation between gender and education is exactly zero. If one of the two variables, e.g. gender, correlates with a third variable (e.g. school success), this correlation can be attributed entirely to this variable, without having to consider that part of the correlation may be "explained" by the correlation between the two "independent variables". Thus, this is a purely methodological procedure to be able to assign statistical correlations more clearly. Such an *experimental* design ("experimental" because it was intentionally set up this way) also offers the possibility to measure *conditional* (interaction) effects by means of variance analysis.

If no experimental design of the study is possible, but a *field study is* available in which the independent variables correlate with each other, the correlations between the independent variables can be calculated out ("controlled"). However, in such cases, which are the rule rather than the exception, the magnitude of the effect size depends on the order in which one controls the independent variables and which variables one controls. Again, there are many influences that have nothing to do with the "true" effects. One can only avoid the suspicion of arbitrariness or even manipulation if precise content-related theories are available from which one can derive substantive hypotheses.

5. All measures of relative effect size (as well as measures of absolute effect size) depend on the distribution of the *independent* variables. In the extreme case, where there is no variance in the independent variable, there can be no effect. Example: If only boys are studied, but no girls, the

---

causal relationship can only be answered by consulting other findings and an explanatory theory. However, as Popper (1968) proved, absolute certainty can never be obtained. Every finding that is assumed to be certain may later turn out to be false on the basis of new research.

effect of the variable 'gender' on moral judgment or mathematics achievement cannot be studied. But also otherwise: the smaller the variance of the values of the independent variable, the smaller is usually the measured effect (mind you: not the 'true' effect), because the 'true' effect is the more overlaid by measurement and sampling errors, the smaller is the variance of the independent variable. One can try to solve this problem by carefully selecting the study groups with respect to certain independent variables (for example: equal numbers of men and women or equal numbers of people with low, medium and high education). However, this is not always possible. However, it should be tried in any case.

What can be done if the study has already been conducted and therefore no more changes can be made to the study design? How can one analyze such studies as well as possible? Here are some tips that follow from the above explanations:

- Do not rely on statistical significance alone! They depend on the sample size and the variance of the dependent variables. Both variables vary strongly from study to study and may also have been deliberately "manipulated" to obtain large effects.
- Measures of relative effect size are better and they can usually be calculated retrospectively from statistical significance. But they also have their problems. Mostly they underestimate the 'true effect' because the variances of the dependent variables often differ strongly from each other, because the variance of the independent variables is often small, and because very restrictive ideas about the relationship between variables (symmetric and mono-causal) are assumed without the researcher often being aware of it.
- A better idea of the practical significance of an effect can be obtained by the *binomial effect size display (BESD)* of Rosenthal and Rubin (1982) and by absolute effect sizes such as (mean) differences.

Measures of absolute effect size

In everyday life, but also in the natural sciences, the procedure is usually different from that used in psychological research. Instead of calculating statistical significance values and relative correlations, the measurement differences are often simply read off (or mean values are calculated over small measurement series if individual observations are too inaccurate). The effect size is then the difference between two measurements or two averages of measurement series. Measures of absolute effect size are thus formed by calculating difference values between before and after measurements without using the dispersion of the dependent variable.

Example: We want to know if it is warmer today than yesterday. To do this, we compare the (average) temperature of yesterday with the (average) temperature of today. If it was 18 degrees Celsius yesterday and 23 degrees today, then the warming is 5 degrees (= 23 minus 18 degrees). In everyday life we are interested in the *practical significance of* this difference. 5 degrees difference seems to be significant for most people, because from this follows the change of various behaviors: One "feels the warming. You don't dress as warmly anymore. With a difference of two degrees, probably nothing would be felt and no other behavior would follow. In meteorology, on the other hand, even much smaller fluctuations in temperature play a role, for example the change in the mean annual temperature. Even a few tenths of a degree of warming can mean the melting of large glaciers and flooding in many coastal areas.

The calculation of the absolute effect size (aES) is based on simple mean comparisons or difference calculations: $M_2 - M_1$. Which mean values are compared depends on the research question and the research design. We are mainly interested in intervention studies in which the effect of a specific intervention (measure, therapy, teaching method, etc.) is to be measured. A certain diffusion has the comparison of the mean values of an experimental group with a comparison or control group: aES = $_{\text{Mexperimtal}}$ - Mcontrol. It is obvious that such a comparison gives us a reliable information about the effect only if the initial values in both groups were the same and also otherwise the participants in both groups did not show any differences that can favor or complicate the effect.

If a purely random allocation to experimental and control groups is not possible in a social science experiment, because it is not only very time-consuming and expensive, but also calls into question the external or ecological validity of the results, the above mean comparison is not useful. Instead, a *before-after comparison* of the mean values is appropriate:
$aES = M_{after} - M_{before}$, or more simply $M_2 - M_1$.

Estimating absolute effect sizes for social science scales.

Assessing the significance of a particular absolute effect size depends primarily on how much we know about the scale we are using to measure the effect. If the scale is still "young" and we have little research and everyday experience on it, only formal properties of the scale itself are available to us for orientation. For finite scales, as is common in the social sciences (including psychology and education), these are the absolute endpoints and the absolute midpoint. The longer a scale is in use and the more we know about the empirical meaning of a scale, the better we can assess the practical significance of scale values and value differences. "Empirical significance" here means both the conditions necessary to achieve certain scale values or for certain value differences, and the various effects that certain scale values or value differences have.
Example: The temperature scale has been familiar to us humans for a long time. We know, for example, how much energy is needed to raise the temperature of a liter of water from 20 degrees Celsius room temperature to 100 degrees boiling temperature. We also know the consequences of a rise in body temperature to 40 degrees and how we have to change our clothing when the outside temperature rises or falls by about 5 degrees.

Unlike in the natural sciences, where we have to deal with a few scales of measurement (for length, time, energy, mass) that have been known for a long time and have been improved again and again, in the social sciences there are an enormous number of scales and most of them are very young. New ones are constantly added and old ones disappear. In the natural sciences there is (apart from some exceptions in everyday life) always only one "operationalization" (measurement operation) for each measurement scale. Variations only concern the accuracy and scale of the measurement (micro, meso and macro). In the social sciences, on the other hand, accuracy plays a subordinate role. Here, on the other hand, different operationalizations of the same measurement dimension are often found, which do not correlate perfectly with each other - as one would expect - and often do not even appear to be very similar to each other. Example: moral judgment (MU). "Ability" is unambiguously defined in everyday life and in science as something that must prove itself in certain types of tasks (moral competence is the ability to perform moral tasks; Lind, 2019). Moral competence is operationalized by many authors as moral *attitude* (e.g., as preference for moral principles, as in the Defining-Issues Test, DIT). Attitudes, however, can be simulated in any direction and at best provide indirect evidence of ability. For example, there is usually a high empirical correlation between ability in a subject area and interest in it. However, this correlation is found only under favorable conditions. Interests and other attitudes are not *reliable* indicators of ability. Therefore, the measurement of ability cannot usually be replaced by the measurement of corresponding attitudes if the measurement is to be reliable.

In the field of moral psychology, the *Moral Competence Test* (MKT) has been around for more than 40 years (Lind 1978; 2019). This means that its meaning is relatively well researched in the two senses of the word (see above): (a) we can estimate well how much education is needed to achieve a given score on this scale, and how good that education must be to achieve a given absolute effect size; (b) we can also estimate relatively well how high the level of moral development must be to minimize the risk of dysfunctional behavior (such as crime, drug addiction, refusal to help, learning problems, inability to make decisions). Estimates are still subject to a relatively high degree of uncertainty, but this uncertainty has been significantly reduced by research with the MST compared to earlier times when this scale did not exist. Presumably, this uncertainty goes back not only to the quality of the scale, but also to the complexity of the behavior at issue. There are also many areas in the natural sciences, such as weather forecasting, where uncertainty exists despite highly precise measurement scales and despite very powerful computer models.

On the conventional estimation of absolute effect sizes for under-researched measurement scales.

When research in a new field is not very advanced and we cannot properly assess what is a large effect or a small effect, we can help ourselves by determining absolute effect sizes by *convention,* for example, by relating them to theoretical scale width. (I still call them "absolute" rather than "relative" because they are not relative to an *empirical* distribution of values). Since social science scales often have very different widths, it is advisable in a first step to convert the scores to a standard scale from 1 to 100 and to relate the differences or changes found to this. In the second step, it should be determined how large a difference should be in relation to this 100-point scale in order for it to be considered "significant" or "marked". Finally, in the third step, a theoretically based gradation of the significance or size of an effect would have to be made.

Here, we will only go as far as the second step and try to find clues for psychologically and pedagogically significant effects. As a basis for this, we draw on empirical studies that are highly regarded in the scientific community and whose findings have had some influence on practice. Since we only want to identify a rough dividing line between significant and insignificant differences in this step, a few studies suffice. Moreover, there are unfortunately only a few studies in which all the necessary information, namely mean differences *and* scale width, is reported. The school study by Fend and colleagues (Fend et al. 1976), the teacher-student study by Dann, Müller-Fohrbrodt, and others (Müller-Fohrbrodt, 1973), and the group interaction research by Oser (1981) were selected from. From each

study, we have selected a few evaluation results as examples, namely those that the authors described as statistically "highly significant" and to which they also attached great importance in their conclusions.

Thirteen such findings have been entered in the following table and converted to a reference scale of 1 to 100. The values in the rightmost column indicate how large the difference or change reported by the authors would be expressed on a scale of 100. For example, a value of 1.75 means that the statistically "highly significant" effect is 1.75% or 1/67 of the total scale width.

As we can see from the table, the authors vary greatly in their judgment of the size at which they consider an effect to be very significant, measured in terms of the absolute scale width. The values range from 1.75% to 15% of the scale width. The average value is 7.83 percentage points. This value seems typical for empirical social research. On this basis, we propose the following verbal descriptions for value differences on attitude scales:

Effect > 10% of scale width = "very significant" or "very clear".

Effect > 5% of scale width = "significant" or "marked".

There are therefore good reasons for speaking of *very significant* differences or changes only if they amount to *10 percent* or more of the theoretically possible scale values. With a scale of 5, for example, as is frequently used in social science research, 0.5 points, i.e. half a scale unit, would be a very significant difference. With a difference of a quarter point, we can still speak of a significant difference at this scale width. The limiting "good reasons" can be of many kinds. In particular, if a particular area is theoretically and empirically well penetrated, there may be reasons for a different (higher or lower) criterion value or for more differentiation in the score. In this case, we can dispense with purely conventionally based criteria.

Autor(en)	Interpretation der Effekte	Vergleich	Skalenbreite	Effekt absolut	Effekt in %
Fend et al. 1976	"Schüler mit hohem Leistungsstatus entsprechen den schulischen Disziplinforderungen eher als Schüler mit niedrigem Leistungsstatus. Die Unterschiede ... sind jeweils hochsignifikant." (S. 79)	mittel/ niedrig	16	0,28	1,75
		hoch/ mittel	16	0,37	2,31
	"Die Lern– und Leistungsmoral der Gymnasiasten ist signifikant (p<.01) niedriger als die der Hauptschüler." (S. 91 f.)	Leistungsmoral	16	0,65	4,06
		Lernmoral	16	0,57	3,56
	"Gesamtschüler befürworten eindeutig stärker eine Selektion nach Leistung in der Schule" (S. 214)		16	0,34	2,13
	"Lehrer an Gesamtschulen sind ... progressiver als die Lehrer an jeder Schulform des herkömmlichen Schulsystems." (S. 216)	GS/HS[5]	16	2,04	12,75
		GS/GY	16	2,36	14,75
	"[Eigen-Orientierungen] nehmen mit dem Herkunftsstatus *ab*." (S. 289)	US/MS	16	0,54	3,38
	Gymnasiasten haben ein höheres Selbstbewusstsein als Hauptschüler (S. 400)		16	1,74	10,88
Müller-Fohrbrodt, 1973	"Während des Studiums werden ... Studenten ... progressiver." (S. 108)	Abi/StF[6]männlich	90	7	7,78
	Studenten werden zwischen Abitur und 2. Studienhälfte "konformistischer" (S. 109)	Abi/StF männlich	30	4,5	15,00
Oser, 1981	"Die ... Stufen 1 und 2.5 lassen keine inhaltliche Integrierung zweier Treatments zu." (S. 401)	ohne/mit Regel	6	0,7	11,67
	"... hochsignifikanter Regelhaupteffekt... Der größte Unterschied liegt bei Stufe 3" (S. 402)	ohne/mit Regel	6	0,51	8,50
Gesamtskalenbreite / Gesamteffekte = relativer Effekt (in %):			276	21,60	**7,83**

On the theory-based estimation of absolute effect sizes for well-researched scales.

However, at an advanced stage of research in a particular field, we often ask whether a particular therapy or teaching method has a greater effect than previously successful therapies or teaching methods. Or we ask what practical implications certain differences on a scale of measurement may have. In this case, the evaluation design should include a comparison group in which the previous methods are applied. The effect of the new method then results from the comparison of two before-after differences:

$$aES = ([M_{exp,2} - M_{exp,1}] - [M_{kon,2} - M_{kon,1}])$$

Example: In Glasstetter's (2005) intervention study of juvenile offenders, the mean scores (moral judgment, C-score, MKT) in the experimental and control groups were as follows:

	Vorher	Nachher	Effekt
Experimentalgruppe (Dilemmadiskussionen u. a.)	$M_{exp,1} = 19,1$	$M_{exp,2} = 18,3$	**-0,8**
Kontrollgruppe (traditionelle Erziehungsmaßnahmen)	$M_{kon,1} = 16,7$	$M_{exp,2} = 12,3$	**-4**
$aES = ([M_{exp,2} - M_{exp,1}] - [M_{kon,2} - M_{kon,1}])$ = -0,8 – (-4,4) = **+3,6**			

Glasstetter (2005), p. 194

As can be seen, a seemingly paradoxical result emerges: In each of the two groups, moral judgment decreased over time, so both measures had a negative effect. But Glasstetter's intervention (experimental group) was apparently able to almost neutralize the clearly negative effect of staying in the institution (control group). Thus, it had an effect. The aES of 3.6 remarkable. This is approximately how strong the effect of an entire school year at the secondary level is (Lind, 2002, p. 159 ff.).

The use of absolute effect sizes thus presupposes that we are well acquainted with the measurement instrument or the measurement scale, which is not the case in a new field of research. In the social sciences, there are hardly any measurement scales with which we are as familiar as with physical quantities such as temperature, length or time. The difference of 3.6 C points in our example above hardly tells anyone - outside a small circle of experts - whether this is much or little, practically relevant or theoretically significant. The assessment of this difference depends first of all on the question of how much this value can vary at all. If the scale is infinite, this information is of little help; but in the social sciences we are mostly dealing with finite scales. For example, the C-value can only range from 0 to 100. Thus, a difference of 3.6 still represents an increase of 3.6 percent of the total scale. If the total length of the scale were only 20 points, this difference would be even more impressive.

Then the assessment of the significance of this difference depends on the question of how much the C score usually differs in different groups. Since it ranges from 10 to 40 for groups of people depending on their level of development, an effect of 3.6 C points on average means that it represents about 10% of the usual range of moral development, which we consider to be quite a lot.

In addition, we must see the difference found in relation to the time within which this change was achieved, measured by how much time is otherwise required for this. I have compiled all studies in which the C-value was examined as a function of time or educational processes. I found that the C-value of students at a general education school increases by about 3.5 C-points per year. An effect of 6.2 C points in an educational intervention experiment that lasted only three months can therefore be considered large (even if, as in the first case of our hypothetical example above, it was not statistically significant).

Finally, we must view an observed difference in relation to the behavior that may or may not trigger it. A given mean difference is neither practically nor theoretically significant if it is statistically "highly significant" but irrelevant to otherwise. Thus, it would still have to be shown that a difference of 6.2 C points leads to different behaviors in people. Based on many studies, such relevance can be assumed today (Lind, 2002).

Summing up

Statistical significance tests should only be used for what they were made for, namely to estimate the accuracy of a measurement process. The conclusion from (statistical) significance to practical significance is inadmissible. The term "significance" (which translates as meaningfulness) is misleading and should be changed.

For conventional reasons, the term significance will probably have to be used for some time even when it is inappropriate, since even the editors of prestigious journals have not yet recognized the problem and reject publications that report effect size measures instead of significance. But the author should always indicate that he/she is aware of the limitations of this concept, and should also always report effect sizes if that is what they want to report.

As long as it is not precisely defined as a norm how large the samples must and may be, the results of such purely statistical procedures for determining "significance" are uncertain(because they reflect not only differences in the mean values, but also different sample sizes and variances) and open to arbitrary intervention by the experi mentator. And anyone familiar with the scientific community knows that this possibility of determining the outcome of a study is also used extensively, without the scientific community seeing this as a violation of its ethos.

The use of measures of *relative effect size (rES) such as "r" and "d" represents* a significant advance over statistical significance tests, but are ambiguous for several reasons and do not allow a simple conclusion about the 'real' effect of a treatment or educational intervention (see above). It should also always be stated whether r or d is being reported. Unfortunately, this indication is often missing, leaving the reader to guess. The disadvantage of rES is that, although they are independent of sample size, they are, as the name implies, "relative" to the dispersion of values in the data under study. Since this dispersion (variance, standard deviation) can differ greatly from study to study, a direct comparison of relative effect sizes is often not possible.

Measures of *absolute effect size (aES)* obtained from simple difference calculations eliminate this drawback. However, they also do not allow a mechanical interpretation. A difference found between pretest and posttest values, or between experimental and control group, does not allow a simple conclusion on the effectiveness of a treatment or an educational measure.

To interpret the significance of an aES, one should always refer to the current state of research and, if necessary, conduct meta-analyses. Coefficients for effect size cannot replace the expert scientific assessment of findings. The practical significance of an intervention effect depends, among other things, on the cost of the intervention, the importance of the target, and the relevance of the effect to specific behaviors. Example: If method A can achieve the same effect with less effort than method B, then method A is 'more efficient'. If only a few patients' lives can be saved with one drug, this numerically small effect may be more significant than if very many people can be helped to cure their cold with another drug. If even a small increase in measured ability has a large long-term impact, even this small effect can be highly significant. The theoretical significance of an effect depends on the research to date. At an early stage of a new research field, even a small effect may provide an important clue; if large effects are already present, new approaches should be able to surpass these effects.

However, this by no means exhausts the possibilities of an effect analysis. For example, the question can be asked:

- How large are the effects of a new psychological therapy or educational intervention compared to "natural" changes or compared to the effects of previous therapies and interventions?
- At what effect size can we expect the effects to be sustainable, i.e., stable over a longer period of time or even to increase? For example, the desired effect of a new teaching method may be that it does not impart more knowledge, but enables the learner to increase his or her own knowledge.

- What is the minimum effect size that must be achieved for people's (future) behavior to be noticeably influenced by it? For example, we can ask at what level of difference in school performance tests students have a better chance of finding a job or earning a certain income in the future. This relationship is called "prognostic validity.
- What other conditions must be met for a particular effect to occur? For exampl, is a particular teaching method always effective or only effective when used by teachers who are adequately trained in that method?
- Finally, when assessing the significance of an outcome, it is important to know the relationship between (therapeutic or educational) effort on the one hand and the benefit for the individual and society on the other. This is called the *efficiency of* the measure under review.

It is indispensable for the progress of social science research to strive for hypotheses that make sense in terms of content and to determine practically and theoretically what, for example, *psychologically significant* differences are (Meehl, 1978). Only then can the question arise of how to plan a study to confirm or refute with sufficient precision and unambiguity a hypothesis or, even better, to choose the correct one from two alternative hypotheses. In the context of such a question, it would also quickly become apparent which statistical ratios are adequate and helpful for gaining scientific knowledge, and which are not. The mechanical application of statistical concepts does not help us scientifically nor practically (Gigerenzer, 1998; Meehl, 1958; Hoffrage, 1998; Sedlmeier, 1998; Lind, 2002; Haller & Krauss, 2002; Thompson, 2006; Bracey, 2006; Nuzzo 2014).

Literature

Aaron, B., Kromrey, J.D., & Ferron, J.M. (1998). *Equating r-based and d-based effect size indices: problems with a commonly recommended formula*. Paper presented at the meeting of the Florida Educational Research Association, Orlando, FL, (ERIC Document No. ED 433 353).

American Psychological Association, APA (1994). *Publication guidelines*. Washington, D.C., 4th edition.

Baker, B. & Welner, K. G. (2012). Evidence and rigor: Scrutinizing the rhetorical embrace of evidence-based decision making. *Educational Researcher, 41*, 3, 98-101.

Bortz, J. (1994). *Statistics*. Berlin: Springer-Verlag.

Bracey, G. W. (2006). *Reading Educational Research: How to Avoid Getting Statistically Snookered*. Heinemann.

Bredenkamp, J. (1970). On measures of practical significance. *Journal of Psychology, 177,* 310-318.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

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

Cohen, J. (1994). The Earth Is Round (p < .05). *American Psychologist, 49* (12), 997-1003.

Cooper, H. & Hedges, L. V., eds (1994). *The handbook of research synthesis*. New York: Russell Sage.

Fend, H., Knörzer, W., Nagl, W., Specht, W. & Väth-Szusdziara, R. (1976). *Socialization effects of school*. Weinheim: Beltz.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103*, 592-96.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences, 21*, 199-200.

Glass, G. V. & Stanley, J. C. (1970). Statistical methods in education and psychology. Englewood-Cliffs, NJ: Prentice Hall.

Glass, G. V., McGaw, B., & Smith, M. L. (1978). Meta-analysis in social research. London: Sage Publications.

Glasstetter, S. (2005). *Moral education according to Lawrence Kohlberg - The effects of the Just-Community in a closed home for delinquent adolescents*. Diploma thesis, Department of Psychology, University of Landau.

Haller, H. & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? Methods of Psychological Research Online 2002, Vol.7, No.1. Retrieved from http://www.uni-landau.de/~agmunde/mpr/issue16/art1/haller.pdf on 3/22/2004.

Hays, W. (1963). *Statistics for psychologists*. New York: Holt.

Hepach, R. (2007). *Intervention studies for* fostering *moral judgment* competence*; a meta-analysis of studies in the years 1985 to 2006* [Interventions studies for fostering moral judgment competence. Meta-analysis of studies in the years1985 to 2006]. Bachelor thesis, Department of Psychology, University of Konstanz.

Hoffrage, (1998). Understanding statistics. BerliNews. Retrieved from http://www.berlinews.de/archiv/1580.shtml (3/22/2004).

Jacobs, B. (n.d.). Some ways of calculating effect sizes. http://www.phil.uni-sb.de/~jakobs/seminar/vpl/meaning/effektstaerketool.htm (Aug. 20, 2007).

Journal of Experimental Education. (1993). Special Issue "*The role of statistical significance testing in contemporary analytic practice: Alternatives with comments from journal editors.* " Washington, DC: Heldref Publications. (Available from ERIC/AE).

Kendall, M. G. & Stuart, A. (1973). *The advance theory of statistics*. Vol. 2. London: Griffin.

Law, K. S. (1995). The use of Fisher's Z in Schmidt-Hunter-type meta-analyses. *Journal of Educational and Behavioral Statistics, 20,* 287-306.

Lind, G. (1985). The content and structure of moral judgment. Theoretical, methodological, and empirical investigations of judgment and democratic competence in college students. Dissertation, Sozialwiss. Faculty of the University of Konstanz.
http://kops.ub.uni-konstanz.de/bitstream/handle/urn:nbn:de:bsz:352-opus-5042/Lind-Diss.pdf?sequence=1

Lind, G. (1991). *The concept of "empirical content" of hypotheses according to POPPER and a practicable operationalization*. Habilitationsvortrag an der Katholischen Universität Eichstätt.

Lind, G. (2002). *Is morality teachable? Results of modern moral psychological research*. Berlin: Logos-Verlag.

Lind, G. (2004). Jenseits von PISA - Für eine neue Evaluationskultur, pp. 1 - 7. In: Institut für Schulentwicklung, PH Schwäbisch Gmünd, eds, *Standards, Evaluation und neue Methoden. Reactions to the PISA Study*. Baltmannsweiler: Schneider Verlag Hohengehren.

Lind, G. (2008). The meaning and measurement of moral judgment competence revisited - A dual-aspect model. In D. Fasko & W. Willis, eds, *Contemporary Philosophical and Psychological Perspectives on Moral Development and Education, pp.* 185 - 220. Cresskill. NJ: Hampton Press. For new studies see; http://moralcompetence.net

Lind, G. (2019). *How to teach moral competence*. New discussion theatre. Berlin: Logos.

Loftus, G. R. (1993). A picture is more worth than thousands of p-values: On the irrelevance of hypothesis testing in the computer age. *Behavior research methods, Instrumentation and computers, 25*, 250-256.

Meehl, P. (1958). When to use your head instead of the formula? In: H. Feigl, M. Scriven & G. Maxwell, eds, *Minnesota studies in the philosophy of science, pp.* 498-506.

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.

Müller-Fohrbrodt, G. (1973). *What are teachers really like? Ideal, advantages, facts*. Stuttgart: Klett.

Nuzzo, R. (2014). When researchers fail the significance test. *Spectrum.com* (from 'nature'). Source:
http://www.spektrum.de/alias/umstrittene-statistik/wenn-forscher-durch-den-signifikanztest-fallen/1224727

Oser, F. (1981). *Moral judgment in groups. Social action*. Frankfurt: Suhrkamp.

Popper, K. (1968). The logic of scientific discovery. London: Hutchinson (original 1934).

Rosenthal, R. & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experi men tal effect. *Journal of Education Psychology, 74,* 166-169.

Rosenthal, R. & Rosnow, R.L. (1984). *Essentials of behavioral research*. New York: McGraw-Hill.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L.V. Hedges, eds, The *handbook of research synthesis, pp.* 231-244. New York: Russell Sage Foundation.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*, 1173-1181.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540.

Sedlmeier, P. (1996). Beyond the significance test ritual: Complements and alternatives. Methods of *Psychological Research Online 1996,* Vol.1, No.4.

Sedlmeier, P. (1998). What are good reasons for significance testing? *Methods of Psychological Research Online,* Vol. 3, No. 1; obtained from:

http://www.uni-landau.de/~agmunde/mpr/issue4/art4/&e=747 (Mar. 22, 2004).

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experi mental Education 61*, 293-316.

Statistical significance testing in contemporary practice (1993). The *Journal of Experimental Education, 61(4),* September 1993.

Thompson, B. (1993). The use of statistical significance tests in research: bootstrap and other alternatives. *Journal of Experimental Education* 6(4), 361-377.

Thompson, B. (1994): http://ericae.net/pare/getvn.asp?v=4&n=5

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (2006). *Foundations of behavioral statistics. An insight-based approach*. Guilford Publications.

Wuttke, J. (2007). The insignificance of significant differences. In: T. Jahnke & W. Meyerhöfer, eds, Pisa *& Co. critique of a program.* 2nd, expanded edition. pp. 99-246. Hildesheim: Franzbecker.

Appendix: Formulas for converting various indices to Relative Effect Size *r*

Last revision: Aug. 2012

Conventions for using symbols (if not otherwise stated): N_i is the size of the *i*-th sample, whereby *i* may be a number between 1 and *k*, the total number of samples. For example, if the first sample has 6 members then $N_1 = 6$. x_i is a variable, which can represent a set of numbers, e.g., the subjects' scores in a math test: x_i= {72, 64, 45, 34, 95, 93}; specifically 2nd value of x is $x_2 = 64$. Σ is the summation symbol. *k* is the number of groups which are compared.		
Combining coefficients of correlation[7]	$M_r = \frac{\sum_{i=1}^{k} N_i * r_i}{\sum_{i=1}^{k} N_i}$	**Example:** In two studies, these correlations were found between moral development scores and level of education (the Ns are in parentheses): $r = 0.45$ (50) and 0.65 (230). The estimated mean of correlations is $M_r = (0.45 * 50 + 0.65 * 230) / 280 = 0.625$.
X² (Chi-square)[8]	$r_{xy} \approx C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$	**Example:** If $\chi^2 = 14.34$ and N = 80 then r_{xy} = sqrt (14.34 / (14.34 + 80)) = sqrt (0.152) = **0.39**.
Effect size measure *D*[9]	$r_{xy} = \frac{d}{\sqrt{d^2 + 4}}$	with $d = \frac{M_1 - M_2}{s}$
***t*-statistic**[10]	$r_{pb} = \sqrt{\frac{t^2}{t^2 + df}}$	**Example**: If $t = 3.5$ and $n_1 + n_2 = N = 250$ then $df = 249$ and $r_{pb} = 0.12$. **Note**: If you use *t* to compare repeated measurements as in follow-up studies, make sure that you use the *t*-statistics for *dependent* (or paired) groups.

7 Law (1995) showed that Z-transformations are not needed for averaging *r*; the results are almost identical.

8 Kendall & Stuart (1967, p. 557)

9 Glass et al. (1978)

10 Glass & Stanley (1970, p. 318)

Point-biserial correlation[11]	$r_{xy} = r_{pb} * 1.25$ This formula can be used only if the ratio n_1 / n_2 is bigger than 0.2 and smaller than 0.8.	r_{pb} denotes the coefficient of point-biserial correlation and df the degrees of freedom: $df = n_1 + n_2 - 1$. n_1 and n_2 are the number of subjects in each of the two groups that are compared. **Example:** If $r_{pb} = 0.12$, $n_1 = 125$, and $n_2 = 125$, then $r_{xy} = 0.12 * 1.25 = 0.15$
***F*-statistic**[12]	$r_{xy} = \sqrt{\frac{df_j * F}{df_j * F + df_i}}$	with df_i being the degrees of freedom *within* groups, and df_j the degrees of freedom *between* groups (number comparisons - 1). **Example**: Three groups (k=3) are compared with n=50 subjects. $F = 23.45$, $df_j = k-1 = 1$, and $df_i = n - k = 47$. Then $r_{xy} = 0.70$.
Variance-model	$r = \sqrt{\frac{S^2_{between}}{S^2_{between} + S^2_{error}}}$	with $S^2_{between}$ being the variance due to the treatment or effect-variable. The variance ration r^2 is also called the *coefficient of determination*. Its square root is the coefficient of correlation, however of total correlation, not only of linear correlation.
Binomial effect size display, BESD[13]	$r = 2 * BESD - 1$ $BESD = .50 + \frac{r}{2}$	**Example:** If r = .30 then BESD = 0,65 If BESD is .75, then r is 0.50
Mann-Whiteney U[14]	$r_{pb} = 1 - 2\frac{U}{n_1 * n_2}$	whereby r_{pb} is the point-biserial correlation coefficient.

Further conversion formulas can be found in Cooper & Hedges, 1994.
Conversion tools for d (effect size) can be found on the Internet, by Bernd Jacobs, Uni Saarbrücken:
http://www.phil.uni-sb.de/~jakobs/seminar/vpl/meaning/effektstaerketool.htm

11 Magnusson (1966, p. 205)
12 Rosenthal & Rosnow (1984, p. 249)
13 Rosenthal & Rubin (1982)
14 Wilson (1976)