

From Bayes Through Marginal Utility to Effect Sizes: A Guide to Understanding the Clinical and Statistical Significance of the Results of Autism Research Findings

Domenic V. Cicchetti · Kathy Koenig · Ami Klin ·
Fred R. Volkmar · Rhea Paul · Sara Sparrow

Published online: 19 May 2010
© Springer Science+Business Media, LLC 2010

Abstract The objectives of this report are: (a) to trace the theoretical roots of the concept clinical significance that derives from Bayesian thinking, Marginal Utility/Diminishing Returns in Economics, and the “just noticeable difference”, in Psychophysics. These concepts then translated into: Effect Size (ES), strength of agreement, clinical significance, and related concepts, and made possible the development of Power Analysis; (b) to differentiate clinical significance from statistical significance; and (c) to demonstrate the utility of measures of ES and related concepts for enhancing the meaning of Autism research findings. These objectives are accomplished by applying criteria for estimating clinical significance, and related concepts, to a number of areas of autism research.

Keyword Clinical significance in autism research

Historical Background

The Contributions of Thomas Bayes

A seminal article by the biostatistically informed Reverend Thomas Bayes was published posthumously, almost 250 years ago, in 1763. It focused, arguably for the first time, upon the crucially important role that prior information adds to the statistical and clinical meaning of a given research investigation.

It is of crucial importance for the reader to get a “feel” for the implications of the word “prior”, in a specific Bayesian sense. For example, when one assesses the level of inter-examiner agreement, say between two clinicians, on the presence or absence of a diagnosis of Autism, among 100 toddlers, with a resulting examiner agreement level of 80%, this takes on clinical meaning when one knows what prior level of agreement can be expected on the basis of chance alone. If the chance level of agreement is 50%, then kappa (Cohen, 1960) becomes: $(80-50\%)/(100-50\%) = 0.60$, or Good agreement, by the criteria of Cicchetti and Sparrow (1981). On the other hand, if the prior level of chance agreement turns out to be 70%, then kappa reduces to: $(80-70\%)/30\% = 0.33$, or Poor agreement by the aforementioned clinical criteria.

It should be noted at this point, that much later, the Reverend Bayes’ ranks were joined by other biostatistical titans, such as Neyman and Pearson (1928, 1933), and later still by Cohen (1965, 1968, 1977, 1988); Borenstein (1998); and by many, many others. The Bayesians fought in brutal statistical verbal warfare with the Frequentists (the Classical Statisticians) led by Sir Ronald A. Fisher, he of the concepts of randomization, Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA), as well as, The Fisher F test, Linear Discriminant Function Analysis, the eponymously named Fisher Exact Probability Test and a multitude of other important and time-honored statistical tests. The Frequentists came close to implying that statistical significance should probably be enshrined in Holy Grail Format as $p \leq 0.05$, while clinical significance be damned. The Bayesian argument was and still is that the Frequentists’ argument had to be invalid since given a large enough sample size, any result, no matter how trivial will be statistically significant at $p = 0.05$. Thus, a correlation or R value of only 0.06 will be statistically significant at the

D. V. Cicchetti (✉) · K. Koenig · A. Klin ·
F. R. Volkmar · R. Paul · S. Sparrow
Child Study Center, Yale University School of Medicine,
New Haven, CT 06520, USA
e-mail: Dom.Cicchetti@yale.edu; dc43@email.med.yale.edu

0.05 level, when based upon 1,000 cases. The thrust of this argument is that the Bayesians would much favor and support a replicable finding of considerable clinical relevance, based upon a small sample size than they would the same replicable clinically meaningful finding that is very large-sample based. This argument would also receive support for its obvious time and cost efficiency value.

The Contributions of the Economist Friedrich von Wieser (1851–1926)

von Wieser (1893) introduced the inter-related, dual concept of Marginal Utility/Diminishing Returns. The now famous construct is defined easily as the additional satisfaction an individual experiences from consuming one more unit of something desirable. However, at some point, the next unit that is added results in less of a desire to consume it further. This defines the “Diminishing Returns” feature of the Marginal Utility concept. Marginal Utility/Diminishing Returns translates, in everyday parlance, into the well-known negative effects of “too much of a good thing.”

The Contributions of the Psychophysicists

The psychophysicists focused upon the relationship between minimally perceived differences in physical sensations, such as levels of illumination, as they are perceived by humans. This idea then ushered into behavioral sciences the very useful notion of the “just noticeable difference” (jnd). The idea here is how much of a difference in a physical scale of measurement needs to occur before it is perceived as a just noticeable difference (jnd) in terms of human perception (Bolanowski and Gescheider 1991; Fechner, 1907; Stevens 1946, 1951, 1968; and Stone and Sidel 1993).

An example would be light measured in Lumens, on the Physical scale vs. light perceived in Brills on the Psychological scale. By pairing various levels of Lumens with a standard, the experimenter sought to answer how many changes in Lumen units were required to result in a jnd in Brill level? Such careful measurements and applications across various sense modalities resulted in the development of Psychophysics, a science that was placed in modern day perspective by the work of S. S. “Smitty” Stevens et al.

The Contributions of Jacob Cohen et al

Cohen (1965, 1977, and 1988) and Borenstein (1998) argued that an average or mean difference, an R/Multiple R value, an ANOVA Result, an F Value, or any other test statistic must have clinical as well as statistical meaning to be relevant. As Cohen (1988) argued cogently, the position

of the Bayesians, *not the Frequentists* paved the way for the all important area of Power Analysis.

In fact, Cohen (1977) summarized the contributions of the Bayesians as follows:

The Fisherian formulation posits the null hypothesis...i.e., the Effect Size (ES) is zero, to which the “alternative” hypothesis is that the ES is *not* zero, that is, *any* nonzero value. Without further specification, although null hypotheses may be tested and thereupon either rejected or not rejected, no basis for statistical power analysis exists. By contrast, the Neyman-Pearson formulation posits an *exact* alternative for the ES, i.e., the *exact* size of the effect the experiment is designed to detect. With an exact alternative hypothesis or specific nonzero ES to be detected, given the other elements in statistical inference, statistical power analysis may proceed” (pp. 10–11).

More recently, Borenstein (1998), in a tribute to the memory of Jacob Cohen, wrote a scholarly treatise entitled, appropriately, for this paper, “The shift from significance testing to Effect Size estimation.” Like Cohen, Borenstein contends, as do we, that the ultimate value of a test of a new treatment (as for Autism) is not whether it has produced a mere statistically significant result, but rather one that is *both* statistically and clinically meaningful. Concepts similar in meaning to ES include relative strength, and clinical or practical Significance.

Whatever the concept, there are a number of sets of criteria that can/should and sometimes have been applied to answer specific questions in the field of Autism and related developmental disorders. Each of these tests of clinical significance whether one of association, correlation, rater agreement, or tests of treatment group differences, shares in common criteria that define an effect as Trivial, Small, Medium, or Large.

Cohen (1965, 1977, 1988), Borenstein (1998) and Borenstein et al. (2001) argued that an average difference, an R/Multiple R value, an ANOVA result, an F value, and any other test statistic must have clinical as well as statistical meaning to be relevant.

The next section of this report will define and apply ES and similar concepts to recently published clinical research investigations. Some additional concepts and sets of criteria have been developed and applied by the first author and colleagues over the past several decades. The urgent need for this report is underscored by the unfortunate fact that too often clinical research investigators have and continue to confuse level of statistical significance with level of clinical significance. Thus, a statistically significant finding that reaches a level of 0.001 is too often *incorrectly* interpreted as more clinically meaningful than

one that is significant at the 0.05 level, this despite the fact that the difference is very often nothing more than a function of a large versus a small sample size (Borenstein 1998). Another common and stark error in reporting occurs when a correlation coefficient (R), of say, 0.30, is alternately reported as Small, Medium, or Large, again, depending primarily upon its level of statistical significance. These all too often reporting errors underscore the necessity for this report, in order to remedy this state of affairs both in the specific field of Autism and related developmental disorders, as well as in biobehavioral research, more generally.

Effect Sizes for R/Multiple R

Cohen (1977) presented recommended ESs for R and Multiple R values, such that: <0.10 = Trivial; 0.10 = Small; 0.30 = Medium; and ≥ 0.50 = Large. These were revised and expanded by Cicchetti (2008), as: <0.10 = Trivial; 0.10 – 0.29 = Small; 0.30 – 0.49 = Medium; 0.50 – 0.69 = Large; and ≥ 0.70 = Very Large. As an application, the R between Adaptive Behavior Composite scores and Full Scale WISC-III IQ scores for children between 6 and 16 years of age is 0.09, an ES that is Trivial (or <0.10 ; Sparrow, Cicchetti, and Balla, 2005).

Effect Sizes for Measures of Internal Consistency: Chronbach's (1950) Alpha (α)

The late Cronbach developed the Alpha statistic in 1950, as a measure of the extent to which the items in a given classification, such as a Domain, sub-Domain, or Total test score, cohere or hang together. In terms of the statistics involved, Alpha can be defined simply as the average of the intercorrelations among all the items in the designated classification. Cicchetti and Sparrow (1990) provided the following guidelines for grading levels of Alpha. It is important to realize that these guidelines assume that there are no floor or ceiling effects, either of which would attenuate or underestimate the alpha values to the extent to which these extraneous effects might occur. The guidelines, for α , are, as follows: <0.70 = Poor; 0.70 – 0.79 = Fair; 0.80 – 0.89 = Good; and ≥ 0.90 = Excellent. As an application, across ages ranging from birth to 90 years of age, Vineland α values, for Communication, Daily Living, and Socialization Domains ranged between 0.95 and 0.98 (Sparrow et al. 2008).

As one of the anonymous reviewers of this article stated, and quite correctly, Coefficient Alpha is influenced by the number of items in a given subdomain, domain, or total test score, such that, all other factors considered, an increase in the number of items will increase the size of the Alpha Coefficient.

ESs for *t*-Test Results: Comparison of Two Independent Groups

The ES is defined here as **d**, the standard difference or average difference between the two groups, divided by the within or common group standard deviation (SD). Guidelines are, as follows: **d** = 0 – 0.19 = Trivial; **d** = 0.20 – 0.49 = a Small ES; **d** = 0.50 – 0.79 = a Medium ES; and **d** > 0.80 reflects a Large ES (Borenstein et al. 2001; Cohen 1988).

In a recent application, it was demonstrated that, among toddlers with Autism Spectrum Disorders, those with functional language, compared to those without functional language, were at statistically higher levels (p between 0.01 and <0.0001) of: Visual Reception (**d** = 0.60, Medium ES); and Receptive Language (**d** = 1.3, Large ES; Paul et al. 2008).

It should be noted here that Kraemer, et al. (2003) showed that the Cohen (1988) **d** statistic, just discussed, is mathematically equivalent to three additional measures of ES. The article is entitled “Measures of Clinical Significance” and appeared in the *Journal of the American Academy of Child and Adolescent Psychiatry*.

The first of these measures, developed by Laupacis, et al. (1988) is the NNT or “the number of patients who must be treated to generate one more success or one less failure than would have resulted had all persons been given the comparison treatment” The second of these measures of clinical significance is the AUC (Area Under the Curve), and is based upon specific cut-points that derive from an intervally scaled clinical test score. The AUC has been defined as “the probability that a randomly selected subject in the treatment group has a better result than one in the comparison group” Here, the AUC refers to the area under a Receiver Operating Characteristic (ROC) curve.

The third of these clinical measures is called the Risk Difference, which is nothing more than the simple difference between the number of failures in the comparison and treatment groups (op cit, Kraemer, et al., pp. 1526–1528).

ESs for ANOVA/ANCOVA Results: \geq Three Groups

The ES formula here is **f** = SD Between Groups/SD Within Groups. The guidelines for **f** are: <0.10 = Trivial; 0.10 – 0.24 = Small; 0.25 – 0.39 = Medium; and ≥ 0.40 = Large.

These criteria would be applicable if, for example, there was interest in comparing, on variables of interest to an autism researcher, four groups of children, who were matched on age and gender, and, further, who met DSM criteria for Asperger Syndrome (AS); High Functioning Autism/Pervasive Developmental Disorder (HFA/PDD); and Pervasive Developmental Disorder Not-Otherwise-

Specified (PDD-NOS); with a comparison group comprised of children with Typical Development (TD).

As a final note, it is important to stress that in the paired comparisons between the groups comprising a given ANOVA or ANCOVA design, each resulting ES needs to be expressed as the aforementioned Cohen *d* statistic; and these comparisons need to follow from a statistically significant overall F ratio; and finally, one needs to control for the number of statistically significant findings that would be expected on the basis of chance alone.

The next section will focus upon the criteria that have been developed for deciding whether a particular inter-examiner reliability coefficient, such as Kappa (Cohen, 1960) for Nominally scaled data—e.g., a dichotomous classification of Autism defined by DSM criteria as either present or absent; Weighted Kappa (Cohen, 1968)—e.g., an ordinal classification of developmental psychopathology into one of three DSM-defined groups: Autistic; Non-Autistic (but on the Autism spectrum (ASD)); and a group of age and gender matched subjects of Typical Development (TD); or the Intraclass Correlation Coefficient (ICC; Bartko 1966, 1974) for variables deriving from an interval scale, such as extent of Pragmatic, Speech/Prosody, or Paralinguistic behaviors, as in Paul et al. (2010).

The necessity for developing criteria for judging levels of strength of agreement or level of clinical significance of the size of a given reliability coefficient was motivated by the fact that, like the aforementioned R statistic, one of trivial importance, say, <0.10 , will reach a level of statistical significance at or beyond the 0.05 level, providing only that the sample size upon which it is based is sufficiently large.

Strength of Agreement (ESs) for Reliability Coefficients (Landis and Koch 1977)

The earliest contribution was developed and published by Landis and Koch (1977), and is applicable to Kappa, Weighted Kappa, and the ICC inter-examiner reliability coefficients, each of which corrects for the extent of Inter-Examiner agreement that would occur by chance alone (e.g. see Cicchetti et al. 2008, for descriptions of these reliability statistics and their known mathematical similarities).

The Landis and Koch criteria for a given reliability coefficient, are given as ordered levels of the strength of agreement: <0 = Poor; 0.00–0.19 = Slight; 0.20–0.40 = Fair; 0.41–0.60 = Moderate; 0.61–0.80 = Substantial; and 0.81–1.00 = Almost Perfect.

Cicchetti and Sparrow (1981); see also Cicchetti (1994) developed a shorter list of criteria that are conceptually similar to those of Landis and Koch, and are also applicable to Kappa, Weighted Kappa, and the ICC. Because of

their mathematical equivalence, Kappa, Weighted Kappa and the ICC have been long considered to be part of a family of interrelated reliability statistics (Fleiss 1975; Fleiss and Cohen 1973).

The Cicchetti and Sparrow criteria for levels of Clinical Significance (another proxy for ESs) are: <0.40 = Poor; 0.40–0.59 = Fair; 0.60–0.74 = Good; and ≥ 0.75 = Excellent.

As noted in Cicchetti, et al. (2006), the Landis and Koch (1977) criteria might be useful when the objective is to train clinical examiners to reach a level of acceptable reliability. This is because of the larger number of gradations of levels of strength of agreement. However, in most circumstances, the two sets of criteria can be used interchangeably because of their aforementioned conceptual similarities.

It was noted by Cicchetti (1988) that a wide range of levels, or Proportions of Observed inter-examiner agreement (PO) can reach the same level of Kappa, Weighted Kappa, or the ICC, depending upon the level of the Proportion of agreement expected on the basis of Chance alone (PC). These two elements form the basis of the formula for Kappa, or $(PO-PC)/(1-PC)$, the *difference* between the observed (PO) and expected levels of examiner agreement (PC), or $(PO-PC)$, as compared to the maximum difference that is possible, or $(1-PC)$. Therefore, criteria were suggested by Cicchetti (2001), for *combining* PO with Kappa, Weighted Kappa, or ICC, as follows; When $PO < 0.70$ and K, Kw, or the ICC < 0.40 , then the level of Clinical Significance is Poor; when $PO = 0.70-0.79$, and K, Kw, or the ICC = 0.40–0.59, then the level of Clinical Significance is considered Fair; when $PO = 0.80-0.89$ and K, Kw, or the ICC = 0.60–0.74, then the level of Clinical Significance is considered Good; and when $PO \geq 0.90$ and K, Kw, or ICC ≥ 0.75 , then the level of Clinical Significance is taken to be Excellent.

These criteria were applied recently to the assessment of the reliability of the ADI-R, across all items, when seven clinicians independently evaluated the same female toddler. PO values ranged between 0.94 and 0.96, with *corresponding* Weighted Kappa values between 0.80 and 0.88 (Cicchetti et al. 2008). Each of these values represents a level of Excellent inter-examiner agreement.

Earlier, Klin, et al. (2000), found that experienced clinical examiners agreed, on the diagnosis of Autism, with $PO = 0.98$ and Kappa at 0.94. These levels were both highly statistically significant ($p < .001$), that is, occurred well beyond chance expectations; *as well as* highly clinically significant (or demonstrating Excellent chance-corrected inter-examiner agreement).

At this point, the reader will realize that nothing has been said about the issue of placing confidence intervals (CIs) around a given statistical test, or around a measure of

ES. Such information is critical, especially when there is no opportunity to replicate a given research investigation, either because of the scarcity of the subject population (say, in the study of a very rare disorder or disease); or because of research funding limitations. While these important issues would demand additional articles in order to do them justice, we are fortunate, indeed, that this void has been filled quite adequately by the scholarly publication of Finch and Cumming (2009) that appeared in the *Journal of Pediatric Psychology*, entitled: “Putting Research in Context: Understanding Confidence Intervals from One or More Studies.” In addition to discussing CIs, as they apply to a host of statistical measures, the authors also provide an Appendix with mathematical formulae for calculating CIs for a number of standard research designs.

As a valuable companion piece, appearing in the same issue of the same journal, Durlak (2009) discusses the importance, selection, and calculation and interpretation of ESs. He also provides a very useful Appendix for calculating various ESs; for transforming one ES into another; and the author also provides some valuable information on CIs in the context of ES calculations.

Finally, Durlak addresses an issue that is of extreme importance in furthering our understanding of the relationship between statistical significance and clinical significance, as, reflected, for example, in the size of a given ES. Thus, although it generally tends to be true that the larger the ES, the more clinically meaningful the result, there are situations in which this rule does not apply. Thus, one can cite examples in which a finding with virtually 100% chance probability of occurrence, and a corresponding ES of essentially 0, can nonetheless have a very high level of clinical or practical relevance. For example, in the grant awarding realm, a mere point or two can mean the difference between the funding or non-funding of a given research grant proposal. As a second striking example, a tiny fraction of a point can mean the difference between the awarding of an Olympian Silver rather than a Gold medal. And, finally, a candidate running for high political office can emerge victorious by a single vote more than her opponent, again, a result that represents close to 100% chance probability, with an ES of very close to 0 (see also, Durlak 2009 for more on this essential point).

If we now re-enter the world of clinical research, we can find analogous examples of the same phenomenon. Thus, Rosenthal (1991) provides examples of the results of important clinical trials in which a Trivial ES was nonetheless associated with such clinically meaningful outcomes that the trials were terminated on purely ethical grounds in the sense that the investigators refused to deny effective treatment to patients in the Control condition. In one such clinical trial, the investigators studied the effects of aspirin in preventing myocardial infarct, or heart attack;

and in a second trial, clinical researchers studied the effects of Propanolol in preventing death. In the aspirin trial the correlation, R , between treatment and outcome was a mere 0.03; and in the Propanolol trial, the analogous correlation between treatment and outcome was not much better, at $R = 0.04$. Each of these ESs was, when carried to two decimal places, 0.00!

In the next section of this report, we will focus upon how ES and the conceptually similar constructs we have discussed are related to the concept of marginal utility/diminishing returns.

ES, the jnd, Clinical Significance and Marginal Utility/Diminishing Returns: How Are They Related?

ES and conceptually similar terms are related to the concept of Marginal Utility, in this fundamental sense: Effect Size and Related Terms, Such As Clinical Significance, and Strength of Agreement, share in common that adding more units of a desired object increases satisfaction, *up to a point*.

How Does this Relate Conceptually to Power Analysis?

Perhaps the best way to begin here is to study the issue of what we hope to accomplish when we set out to undertake a Power Analysis. The goal, of course, is to determine that minimal sample size, for each group of interest that will produce both statistically significant and clinically relevant research outcomes. This may take the form of: an adequate ES; Good rater agreement; a clinically meaningful group difference; or, say, a sufficiently large R or multiple R , depending upon the hypotheses or research questions of interest.

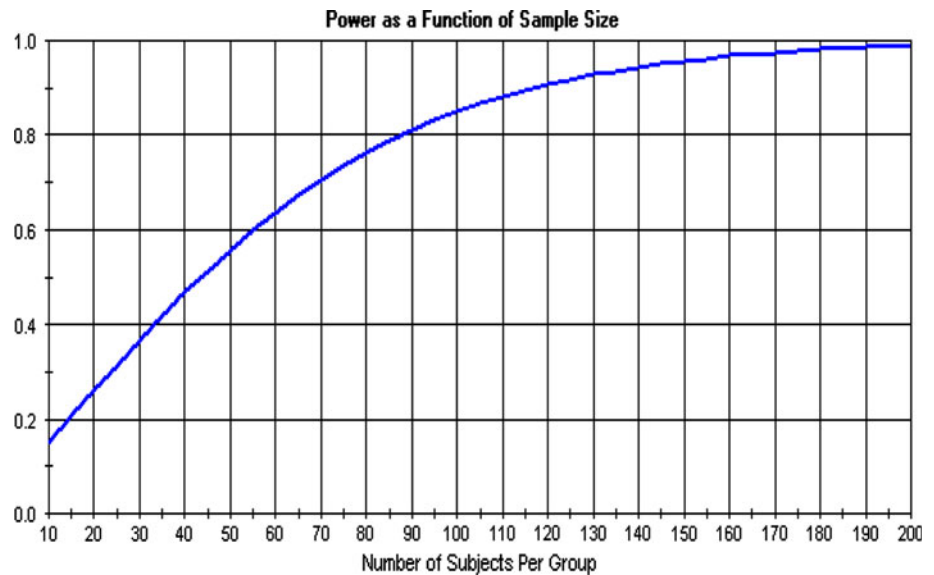
In determining a minimal sample size that will produce statistically and clinically meaningful results (adequate ES, Good agreement, large R), more is better than less, again, *up to a point*.

To validate this argument, let us assume that we are applying for an Autism research grant and we wish to determine what minimal sample size in each of our study and control groups can be expected to produce differences that are both statistically and clinically meaningful. Suppose now that we determine that samples of 90 children in both our study and control group will produce a Medium ES, with a Power of 80% to detect a statistically significant group difference on a treatment variable of interest.

Let us say that we further determine that we can increase our Power estimate from 80 to 90% by increasing our group sample sizes from 90 to 120 children per group. This can be shown graphically, as illustrated in Fig. 1.

However, in considering the added personnel and time required to recruit the additional 30 subjects per group

Fig. 1 Power as a function of sample size, using hypothetical data: see text for further details



increase in sample size, we discover that this stratagem is both *cost and time-inefficient*, so we choose to stay with the original plan of 90 children per group with its associated Power Estimate of 80%.

This same type of logic would apply to the issue of say, developing a Screener for determining whether a child is diagnosed to be on the Autism Spectrum (AS). We would again be struggling with striking the necessary balance between Power and number of items required to develop a screener that has good psychometric properties. The late and renowned psychometrician Nunnally (1978) spoke eloquently more than three decades ago, on this very issue, with reference to providing “standards of reliability” for developing new clinical tests:

“...for basic research, it can be argued that increasing reliabilities much beyond 0.80 is often wasteful of time and funds. At that level correlations are attenuated very little by measurement error. To obtain a higher level of reliability, say of 0.90, strenuous efforts at standardization in addition to increasing the number of items might be required. Thus the more reliable test might be excessively time consuming to construct, administer, and score” (p. 245).

But, the reader may ask, how does this psychometric argument relate to the economic concept of Marginal Utility/Diminishing Returns? The following reasoning applies:

When too many units are added of a desired consumer product, a point is reached in which the output (satisfaction) begins to decrease. As previously noted, economists named this now very familiar inter-connected dual concept Marginal Utility/Diminishing Returns. It now becomes quite obvious that this is the same phenomenon that applies to ES, Power Analysis, and the Nunnally (1978) expressed cost of

increasing test reliability beyond a certain point. And, finally, the same logic applies to the psychophysical concept of the jnd. Once the jnd is defined for a given sense modality, it makes little sense to speak in terms of 2, 3, or more jnds!

Discussion

Broad Conclusions and Lessons Learned

Although the roles of the Bayesians, later Biostatisticians, and the Economist Friedrich von Wieser in the development of criteria to define levels of ES and similar concepts, have been widely disseminated, it is nonetheless unfortunate that the thrust of most research reports is on levels of statistical significance, all too often to the neglect of clinical significance and related concepts. As has been demonstrated, adding the dimension of ES, Clinical Significance, or Practical Significance, Strength of a research finding, or the concept of the jnd, provides a richness of understanding that is not possible when statistical significance alone is used to understand the meaning of Autism or a biobehavioral result, more generally.

In a broader sense, and as far as we are aware, this article represents the first attempt to link together historically, theoretical constructs that chronicle the evolution of the concept of clinical significance, as it has been expressed, across several disciplines, and to demonstrate their relevance for research in Autism and beyond.

References

- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19, 3–11.

- Bartko, J. J. (1974). Corrective note to “the intraclass correlation coefficient as a measure of reliability”. *Psychological Reports*, *34*, 418.
- Bayes, T. (1763). “An essay, by the late Reverend Mr. Bayes, F.R.S. communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S. Philosophical Transactions, giving some account of the present undertakings, studies and labours of the ingenious in many considerable parts of the world, vol 53, 370–418.
- Bolanowski, S. J., Jr., & Gescheider, G. A. (Eds.). (1991). *Ratio scaling of psychological magnitude: In honor of the memory of S.S. Stevens*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Borenstein, M. (1998). The shift from significance testing to effect size estimation. In A. S. Bellak & M. Hershen (Series Eds.) & N. Schooler (Vol. Ed.), *Research and methods: Comprehensive clinical psychology* (Vol. 3, pp. 313–349). New York, NY: Pergamon.
- Borenstein, M., Rothstein, H., & Cohen, J. (2001). *Power and precision: A computer program for statistical power analysis and confidence intervals*. Englewood, NJ: Biostat, Inc.
- Cicchetti, D. V. (1988). When diagnostic agreement is high, but reliability is low: Some paradoxes occurring in joint independent neuropsychology assessments. *Journal of Clinical and Experimental Neuropsychology*, *10*, 605–622.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.
- Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology*, *23*, 695–700.
- Cicchetti, D. V. (2008). From Bayes to the just noticeable difference to effect sizes: A note to understanding the clinical and statistical significance of oenologic research findings. *Journal of Wine Economics*, *3*, 185–193.
- Cicchetti, D. V., Bronen, R., Spencer, S., Haut, S., Berg, A., Oliver, P., et al. (2006). Rating scales, scales of measurement, issues of reliability: Resolving some critical issues for clinicians and researchers. *Journal of Nervous and Mental Disease*, *194*, 557–564.
- Cicchetti, D. V., Lord, C., Koenig, K., Klin, A., & Volkmar, F. (2008). Reliability of the ADI-R: Multiple examiners evaluate a single case. *Journal of Autism and Developmental Disorders*, *38*, 764–770.
- Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*, 127–137.
- Cicchetti, D. V., & Sparrow, S. S. (1990). Assessment of adaptive behavior in young children. In J. J. Johnson & J. Goldman (Eds.), *Developmental assessment in clinical child psychology: A handbook* (chap. 8 (pp. 173–196). New York: Pergamon.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *23*, 37–46.
- Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.). *Handbook of clinical psychology*.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for partial credit. *Psychological Bulletin*, *70*, 213–220.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Glendale, NJ: Lawrence Erlbaum, Associates.
- Cronbach, L. J. (1950). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
- Durlak, J. A. (2009). How to select, calculate and interpret Effect Sizes. *Journal of Pediatric Psychology*, *34*, 917–928.
- Fechner, G. (1907). *Elemente der Psychophysik I u. II* Leipsig. Germany: Breitkopf & Hartel.
- Finch, S., & Cumming, G. (2009). Putting research in context: Understanding, confidence intervals from one or more studies. *Journal of Pediatric Psychology*, *34*, 903–916.
- Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, *31*, 651–659.
- Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619.
- Klin, A., Lang, J., Cicchetti, D. V., & Volkmar, F. (2000). Inter-rater reliability of clinical diagnosis and DSM-IV criteria for autistic disorder: Results of the DSM-IV autism field trial. *Journal of Autism and Developmental Disorders*, *30*, 163–167.
- Kraemer, H. C., Morgan, G. H., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*, 1524–1529.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.
- Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine*, *318*, 1728–1733.
- Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, *20A*, 175–240. and 263–294.
- Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Transactions of the Royal Society of London Series A*, *231*, 289–337.
- Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.
- Paul, R., Chawarska, K., Cicchetti, D., & Volkmar, F. (2008). Language outcomes of toddlers with autism spectrum disorders: A two year follow-up. *Autism Research*, *1*(2), 97–107.
- Paul, R., Miles-Orlovsky, S., Marcinko, H. C., & Volkmar, F. (2010). Conversational behaviors in youth with high-functioning ASD and Asperger Syndrome. *Journal of Autism and Developmental Disorders*, *39*, 115–125.
- Rosenthal, R. (1991). Meta-analytic procedures for social research. *Applied Social Research Methods Series*, *6*, 1–155.
- Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland II: A revision of the vineland adaptive behavior scales: I. Survey/caregiver form* (2nd edn). Circle Pines, Minnesota: American Guidance Service.
- Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2008). *Vineland II: A revision of the vineland adaptive behavior scales: II. Expanded form* (2nd edn). Circle Pines, Minnesota: American Guidance Service.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *10*, 677–680.
- Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In Stevens, S. S. (Ed.). *Handbook of experimental psychology*, chap. 1 (pp. 1–49). New York, NY: Wiley.
- Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, *161*, 849–856.
- Stone, H., & Sidel, J.I. (Eds.). (1993). *Sensory evaluation practices* (2nd ed.). New York, NY: Academic Press.
- Von Wieser, F. (1893). *Natural value* (English ed ed.). New York, NY: MacMillan.