Running Head: CLINICAL SIGNIFICANCE

# "Clinical" Significance: "Clinical" Significance and "Practical" Significance are NOT the Same Things

Lisa S. Peterson

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 7, 2008.

## Abstract

Clinical significance is an important concept in research, particularly in education and the social sciences. The present article first compares clinical significance to other measures of "significance" in statistics. The major methods used to determine clinical significance are explained and the strengths and weaknesses of clinical significance quantification are examined. Finally, examples demonstrating the use and value of clinical significance in education and related fields are presented.

In research, the goal of any statistical analysis is to find results that are "significant". This term is problematic, because despite the implication research that is "significant" is not necessarily important (Thompson, 2006). As noted in various histories (cf. Hubbard & Ryan, 2000; Huberty, 1999, 2002), the concepts of "statistical significance" and "practical significance" have been intermingled and confused for several decades. In behavioral science, a third method, clinical significance (Campbell, 2005), has emerged as a way to further decide if something that is "significant" is actually important and valuable to researchers and their field. The present paper reviews the various types of significance and describes the methods used in establishing clinical significance.

## Statistical Significance

Statistical significance testing dates at least three hundred years, but reached the forefront of research in the early 1900s with the emergence of three methods: chi-square testing, *t* tests, and ANOVA (Thompson, 2002). Null hypothesis statistical significance testing, or NHSST, has grown further over the years, with numerous methods of deciding whether results of research are significant. In statistical significance, research is set up with a null hypothesis which states an assumption about a population (generally that all things are equal or that there is no change with a treatment). Statistical analyses are done that determine the probability ($p_{calculated}$) that the sample results could have come from a population described by this assumption, and given the sample size (Thompson, 2006).

The problem with statistical significance is that NHSST gives no indication of whether the results are important to the researchers or their field; it only tells us whether the results are likely given a certain assumption. Thompson (2002) noted that events that are

likely are often very important, but so can unlikely events also be important. As Thompson (1993) further observed with respect to NHSST p values, "If the computer package did not ask you your values prior to its analysis, it could not have considered your value system in calculating *p*'s, and so *p*'s cannot be blithely used to infer the value of research results" (p. 365). No human values are ascribed to NHSST results, making it harder to decide if the research is really important to clinicians, educators, or anyone hoping to use the knowledge gained in the "real world".

### *Practical Significance*

In order to obtain more "importance" from research, there has been a strong push to include indicators of practical significance in data analysis. The most important of these indicators is effect size. Effect size is a statistical method that quantifies the effect of a treatment or intervention in a research study by examining how much the statistics diverge from the null hypothesis (Thompson, 2006). There are many choices for effect sizes (Thompson, 2007), some that are "corrected" for individual differences that make replication of research more difficult, and some that are not; some that are concerned with mean differences, and some that are concerned with variability. The basic concept of all effect sizes is the same- did the treatment or intervention make a difference, and how much of a difference did it make?

Effect sizes are valuable because they give a much better idea of how "important" a study is. Instead of reading a study and simply being told that results are "significant", there is a quantifiable way of showing what that significance is. Effect sizes are seen as a critical component in research, with many journals now requiring effect sizes to be included in

studies. The American Psychological Association Task Force has also promoted effect sizes as being critical to research results (Wilkinson & APA Task Force on Statistical Inference, 1999).

## *Clinical Significance*

Most effect sizes are focused on group changes, with no indication of what happened on an individual level. A new movement, mostly out of psychology and the behavioral sciences, has added a third kind of significance to the research vernacular. In many fields, treatment is done to help people with a particular label, whether it is a mental disorder, a learning disability, or another diagnosis related to that field. In these instances, it is valuable to know not only if treatment is effective, but whether treatment affected the label or diagnosis. Is a certain therapy improving a client's depression enough to remove the depression diagnosis? Is a certain reading intervention improving a child's reading skills enough to move the student back into regular reading instruction?

Clinical significance (Campbell, 2005) methods attempt to answer these questions about research importance. The first clinical significance test was created in 1984 by Jacobson, Follette, and Revenstorf, a group of psychotherapists who felt a void in their field. They felt that knowing the mean results of a treatment did not give any real information about how many clients benefited from treatment, and how many clients moved from dysfunctional ranges to functional ranges (Jacobson, Follette, & Revenstorf, 1984). Their statistical method for determining clinical significance became the basis for the field of clinical significance. Since then, many variations on clinical significance have been developed by various researchers.

Clinical significance, then, brings a new determination of "importance" of research to fields in which individual improvements are at least as important as group improvements. It is a step forward from practical significance in fields where effect sizes are not enough to guide future work in the field.

## Comparing the Types of Significance Testing

Because clinical significance is an emerging concept in research, it is often confused with the other methods of significance testing, and is often assumed to be just another way to describe practical significance. It is vital, then, that this paper is clear on how clinical significance is its own method that brings a new element to research.

### *Heuristic Example: Depression*

Suppose that you are a school psychologist looking for an effective way to work with a student at your school who is dealing with depression. You find two research articles describing randomized clinical trials involving depression therapy for children. Both seem like promising solutions for your student. The researchers in both cases did null hypothesis statistical significance testing, and obtained a $p_{calc}$ of 0.02, indicating that posttest results for the treatment group were different (and hopefully better) than that of the control group. The researchers were also concerned with the practical significance of their results, and calculated the effect size using Cohen's d. Both studies reported an effect size of 0.8, which is considered to be a very high effect size, indicating that the difference between the posttest scores of the treatment group and the control group was the same in both studies.

Most studies only give you, at best, these two significance testing results. But what

if the researchers had reported the clinical significance of their results? If they had, the two studies may not be as equal as they seem. As it turns out, if you had the clinical significance results, you would see a sharp contrast. The participants in the first study, despite the large effect size, still scored high enough on the posttest (a common depression rating scale) to need therapy after the treatment ended. In the second study, however, 75% of the participants dropped their posttest scores enough to no longer need therapy for depression. Knowing the clinical significance of these studies will clearly differentiate the two, and help you make your decision.

*Heuristic Example: Reading Fluency*

To further demonstrate the power of clinical significance, particularly in the face of other significance tests, a hypothetical study was created for this paper. In Mrs. Brown's third grade class, three children who are scoring below the grade level expectation on reading fluency are given an experimental intervention over a three week period. Three other children who are below the cutoff are not given any intervention. Both groups are given a pretest and posttest using progress monitoring passages from the Texas Primary Reading Inventory. At this point in the school year, the grade level expectation is 90 words per minute. Their scores are listed below, given in words per minute.

| Reading Fluency Intervention Results in a Third Grade Classroom | | |
|---|---|---|
| Intervention Group | Pretest | Posttest |
| Mary | 84 | 95 |
| Jose | 73 | 87 |
| Bobby | 87 | 91 |
| Control Group | Pretest | Posttest |
| Emma | 72 | 80 |
| Chris | 88 | 88 |
| Latoya | 84 | 81 |

Using statistical significance, we can use a *t* test to determine whether to reject the null hypothesis that the mean of the intervention group equals that of the control group. The test actually shows that there is no difference between the groups! This is most likely due to the small sample size; the effect of sample size on results is a major flaw of statistical significance testing (Thompson, 2006). Even if we had a large enough sample size to get a result from the test, would it tell us anything besides that the two groups were different? As a teacher or other educator trying to understand the results, just knowing that the intervention group did differently is not enough information to help make decisions.

Using practical significance, we can calculate the effect size using Glass's delta. This gives us an effect size of 1.35. So we know that our intervention has had a positive effect. This information is important, because now we see *how much* better the intervention group did than the control group. But we do not know how the students did individually, or if the intervention put them into the normal range.

To really look at what happened, we need to use clinical significance. Using the grade level expectation of 90 words per minute, we can quickly see that two of our intervention students, Mary and Bobby, not only improved, but are now performing on grade level. These students will no longer need supplemental instruction, and can return to the regular classroom for their entire reading curriculum.

Knowing the clinical significance of our intervention gives us a great advantage over just knowing that the intervention group is different than the control group, or that the intervention had a large effect size. We can, as educators, do a lot more with this study knowing that this intervention improved all students and put two out of their previous categorization.

## Methods for Evaluating Clinical Significance

While there are many methods for evaluating clinical significance, the five most commonly cited ones will be discussed in this overview. The basic statistical methods of each method will be discussed; for more information, see the article from which each method originated.

### *Jacobson-Truax (JT) Method*

The first major clinical significance method, JT was first developed by Jacobson, Follette, and Revenstrof (1984) and was later revised by Jacobson and Truax (1991). JT contains two steps- the definition of a cutoff point, and a comparison of scores. First, a cutoff point between dysfunctionality and functionality is established. This can be defined three ways, labeled "a"," b", and "c". Cutoff "a" says that the post-treatment score should fall outside the range of the dysfunctional population; cutoff "b" says that the post-treatment score should fall within the range of the functional population, and cutoff "c" says that the post-treatment score should fall closer to the mean of the functional population than the mean of the dysfunctional population. Once the researcher has chosen a cutoff, the second step is to determine how much change has occurred, which is determined using the Reliability Change Index (RCI). The RCI takes the difference in each participant's pretest and posttest scores and divides it by the standard error of the difference. Using the cutoff score and the RCI, participants are divided into four categories- recovered (RCI is positive and cutoff is met), improved (RCI is positive), unchanged (no change in RCI), or deteriorated (RCI is negative).

To demonstrate how JT works, the previous data on reading fluency will be reviewed. Mary and Bobby raised their posttest scores, and would therefore have a positive RCI. They also met the cutoff score of 90 words per minute, so they would be considered "recovered". Emma and Jose raised their posttest scores, giving them a positive RCI also, but they did not meet the cutoff score, so they are "improved". Chris's posttest score was identical to his pretest score, so he has an RCI of zero and is "unchanged". Latoya's posttest score was lower than his pretest score, so she has a negative RCI and is "deteriorated".

*Gulliksen-Lord-Novick (GLN) Method*

GLN was developed as an attempt to fix what Hsu (1999) considered to be a fault in JT. He felt that JT's use of the pretest and posttest difference in determining RCI did not consider regression to the mean. To solve this problem Hsu created GLN, which borrows statistical methods from the work of Gulliksen in 1950 and Lord and Novick in 1969. In GLN, the pretest score and posttest score are subtracted from the hypothesized population mean in order to factor in regression to the mean. In addition, instead of dividing by the standard error GLN divides by the standard deviation of the population.

*Edwards-Nunnally (EN) Method*

Speer (1992) was critical of JT for the same reasons as Hsu, and developed his own method to deal with regression to the mean, based on ideas put forth by Edwards in 1978 and Nunnally in 1965. EN factors in the reliability of the scores being used, which brings pretest scores more toward the mean (i.e., the score is smaller after this adjustment). Then, this estimated score is placed on a confidence interval. The participant's change between

pretest and posttest is then determined by this confidence interval instead of the score. Because of the use of a confidence interval, a greater change is necessary to have clinical significance than with JT.

*Hageman-Arrindell (HA) Method*

Hageman and Arrindell (1999) felt that two changes needed to be made to JT. First, they wanted to see a more defined distinction between individual change and group change, which they felt required different statistical methods. Second, they, like Hsu and Speer, felt that true scores that factored in regression to the mean were more accurate than observed scores. To solve these issues, HA introduces two new indices, which were developed by Cronbach and Gleser (1959; cited in Hageman & Arrindell, 1999). The reliability of change, $RC_{indiv}$, computes the client's classification with at least 95% accuracy. The $RC_{indiv}$ can be interpreted quickly- a score greater than 1.65 means the participant has deteriorated, a score between 1.65 and -1.65 shows no reliable change, and a score below 1.65 indicated improvement. The clinical significance of change, $CS_{indiv}$, modifies JT's cutoff score calculations by determining true scores and reliability coefficients. $CS_{indiv}$ can be used to classify the participant into four categories: deteriorated, not reliably changed, improved but not recovered, and recovered. HA addresses the concern about looking at individuals versus groups by creating the indices $RC_{group}$ and $CS_{group}$, which have their own calculations that reflect group characteristics.

*Hierarchical Linear Method (HLM)*

Speer and Greenbaum (1995) proposed a method for clinical significance that does

not use pretest and posttest scores. HLM is based on growth curve models, and requires at least three data points from the individual. Bayes estimates are used to determine the changes in the individual. The details of these calculations are beyond the scope of the present introductory paper, and are usually determined by computer software. Proponents of HLM believe that it allows greater flexibility than the classic models and more valuable information.

| Common Methods for Determining Clinical Significance | | | |
|---|---|---|---|
| Method | Developed by (year) | Formula | Description |
| Jacobson-Truax Method (JT) | Jacobson, Follette, & Revenstorf (1984), revised by Jacobson & Truax (1991) | $\frac{(X_{post} - X_{pre})}{(2[S_{pre}(1-r_{xx})^{0.5}])^{0.5}}$ | Determines cutoff points and Reliability Change Index (RCI) |
| Gulliksen-Lord-Novick Method (GLN) | Hsu (1999) | $\frac{[X_{post} - M_{pop}] - r_{xx}[X_{pre} - M_{pop}]}{S_{pop}(1 - r^2_{xx})^{0.5}}$ | Alters JT by factoring in hypothesized group means |
| Edward-Nunnally Method (EN) | Speer (1992) | $[r_{xx}(X_{pre} - M_{pre}) + M_{pre}] \pm 2S_{pre}(1 - r_{xx})^{0.5}$ | Alters JT by placing true score on a confidence interval |
| Hageman-Arrindell Method (HA) | Hageman & Arrindell (1999) | $\frac{(X_{post} - X_{pre})r_{dd} + (M_{post} - M_{pre})(1-r_{dd})}{((r_{dd})^{0.5})((2S^2_E)^{0.5})}$ | Alters JT by calculating clinical significance index and reliability of change index; can calculate individual or group change |
| Hierarchical Linear Model (HLM) | Speer and Greenbaum (1995) | $B^*/V^{*1/2}$ | Uses growth curve models to determine clinical change |

*Comparison of Methods*

Researchers have attempted to determine which of the methods available for determining clinical change is optimal. In one study (McGlinchey, Atkins, & Jacobson, 2002), data on participants with major depressive disorder was analyzed using JT, GLN,

EN, HA, and HLM. The researchers concluded that, despite the supposed improvements made to JT by GLN, EN, and HA, there was no evidence that any of these methods were better than JT. HA showed the most deviation from the other methods, but this was expected given the amount of change needed to change categories in this method.

Another study, done by Atkins, Bedics, McGlinchey, and Beauchaine (2005), compared JT, GLN, EN, and HA in a simulation study. The results showed that JT and GLN were almost identical, while HA was significantly more conservative. EN was more "certain": it had more recovered and deteriorated cases, and less unchanged. Again, this study did not show any evidence that would move researchers away from using JT. The authors of this study, as well as of the previous one, recognized that further research is needed to determine if there is one method that best captures the changes during treatment.

## Discussion

### *Applications in Education*

While clinical significance is based in psychology, it is clearly valuable in education as well. In any research involving educational services, an intervention is being put in place with the hopes of improving educational outcomes. Clinical significance could be used anytime an intervention is designed to affect labels or diagnoses. In special education, for example, clinical significance could show whether strategies affect a child's diagnosis with a disability. In an early study (Webster-Stratton, Hollinsworth, & Kolpacoff, 1989), evaluating training programs for families with children with conduct problems, clinical significance methods were used to determine if the child's score on the Child Behavior Checklist was moved into the normal range. Another study, this time in speech therapy

(Finn, 2003), used clinical significance to evaluate children's progress in interventions for stuttering. Other areas where clinical significance could be used, but has yet to be explored, are reading interventions (particularly with the advent of Response to Intervention, where children are moved through "tiers" as their interventions become more intensive), bilingual education and second language instruction (whether a child moves from Spanish to English dominant, for example), and performance on standardized testing.

*Conclusions*

Clinical significance adds extra information to research in the social sciences and education. It helps to determine change at the individual level, both how a participant has responded to a treatment or intervention, and whether this response affects a label or diagnosis. There are many choices as to which statistical method to use to determine statistical significance, and while research on which is most effective is still in the early stages, almost all of these methods are based on the same theoretical concepts. Adding clinical significance information to research would be valuable to anyone who wants to really understand the effect of a new treatment, and should be considered in the future as a new and promising way to interpret data when recognized diagnostic criteria exist for the phenomena being studied.

References

Atkins, D. C., Bedics, J. C., McGlinchey, J. B., & Beauchaine, T. P. (2005). Assessing clinical significance: does It matter which method we use? *Journal of Counseling and Clinical Psychology, 73*, 982-989.

Campbell, T. C. (2005). An introduction to clinical significance: An alternative index of intervention effect for group experimental design. *Journal of Early Intervention, 27*, 210-227.

Finn, P. (2003). Evidence-based treatment of stuttering: II. Clinical significance of behavioral stuttering treatments. *Journal of Fluency Disorders, 28*, 209-218.

Hageman, W. J., & Arrindell, W. A. (1999). Establishing clinically significant change: increment of precision and the distinction between individual and group level of analysis. *Behaviour Research and Therapy, 37*, 1169-1193.

Hsu, L. M. (1999). A comparison of three methods of identifying reliable and clinically significant client changes: commentary on Hageman and Arrindell. *Behaviour Research and Therapy, 37*, 1195-1202.

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

Huberty, C. J. (1999). On some history regarding statistical testing. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 1-23). Stamford, CT: JAI Press.

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

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*, 336-352.

McGlinchey, J. B., Atkins, D. C., & Jacobson, N. S. (2002). Clinical significance methods: which one to use and how useful are they? *Behavior Therapy, 33*, 529-550.

Speer, D. C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology, 60*, 402-408.

Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology, 63*, 1044-1048.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*, 361-377.

Thompson, B. (2002). "Statistical", "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-80.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: The Guilford Press.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423-432.

Webster-Stratton, C., Hollinsworth, T., & Kolpacoff, M. (1989). The long-term effectiveness and clinical significance of three cost-effective training programs for families with conduct-problem children. *Journal of Consulting and Clinical Psychology, 57*, 550-553.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.