1-1-2017

Management Issues: Large effect sizes do not mean most people get better - clinical significance and the importance of individual results

Scott A. Jensen
University of the Pacific, sjensen@pacific.edu

Samantha M. Corralejo
Utah State University

Recommended Citation
Jensen, S. A., & Corralejo, S. M. (2017). Management Issues: Large effect sizes do not mean most people get better - clinical significance and the importance of individual results. *Journal of Child and Adolescent Mental Health, 22*(3), 163–166. DOI: 10.1111/camh.12203
https://scholarlycommons.pacific.edu/cop-facarticles/621

Large Effect Sizes Do Not Mean Most People Get Better:

Clinical Significance and the Importance of Individual Results

Scott A. Jensen and Samantha M. Corralejo

University of the Pacific

Author Note

Scott A. Jensen, Department of Psychology, University of the Pacific; Samantha M. Corralejo, Department of Psychology, University of the Pacific.

Samantha M. Corralejo is now at Department of Psychology, Utah State University.

Correspondence concerning this article should be addressed to Scott A. Jensen, Department of Psychology, University of the Pacific, 3601 Pacific Ave., Stockton, CA, 95211. E-mail: sjensen@pacific.edu

Abstract

Background: This paper seeks to compare group statistical analysis with effect size, group measures of clinical significance (Reliable Change Index and Normative Comparison), and individual analysis of clinical significance. **Method:** Measures of variables important to parenting and child behavior improvement (Parenting Scale, Eyberg Child Behavior Inventory, and Parenting Stress Index) were administered pre and post for a nine to ten week group Behavioral Parent Training Intervention. Analysis compares traditional group statistical significance testing with group measure of clinical significance and individual analysis of clinical significance. **Results:** All three measures demonstrated statistically significant differences from pre to post, with large effect sizes. Group measures of clinical significance, however, demonstrated meaningful change only on the PSI, while individual analysis showed improvements of 54% of participants at best and 0% at worst. **Conclusions:** Individual analysis of clinical significance provides valuable information in treatment outcomes and should be included as a standard practice in outcomes research.

Keywords: clinical significance, statistical analysis, individual analysis, treatment outcomes, behavioral parent training

Large Effect Sizes Do Not Mean Most People Get Better:
Clinical Significance and the Importance of Individual Results

Most examinations of clinical intervention outcomes rely heavily on group statistical analysis to determine if the treatment group and one or more other groups are different in such a way that the difference was unlikely to have occurred by chance. Such analysis offers powerful tools that can provide important group information including effect size, which measures the strength of a statistically significant difference (Cohen, 1988). Group statistical analysis has been even further the norm with the increase of randomized control trials in evaluating potential empirically supported treatments (Chambless & Hollon, 1998).

Several researchers have noted that statistically significant differences between a treatment and non-treatment group, even when bolstered by effect size measures, are insufficient for the purposes of clinical outcomes research; group significance testing and effect size measures provide no information as to whether the treated individuals have returned to normal functioning or made clinically meaningful change (Barlow, 1981; Kazdin 1977, Yeaton & Sechrest, 1981). This concept of determining whether treated individuals return to normal functioning or make meaningful improvements has been labeled clinical significance. Clinical significance is not a new concept, but its adoption within outcome research remains variable (Jacobson, Follette, & Reventsdorf, 1984; Kendall & Grove, 1988). A major advantage to the concept of clinical significance is its clear focus on the treatment benefits rather than a simple reliance on the presence and/or strength of difference between treated and non-treated groups.

Measuring clinical significance is intended to answer two important questions "(a) Is the amount of change that has occurred, presumably because of treatment, large enough to be considered meaningful and (b) are treated individuals distinguishable from normal individuals

with respect to their primary complaints following treatment?" (Kendall et al., 1999, p. 285). In a special issue on the topic, Jacobson et al. (1999) and Kendall et al. (1999) laid out the two most common methods for answering these question of clinical significance: the Reliable Change Index (RCI) and Normative Comparisons.

The RCI, first proposed by Jacobson, Follette, and Revenstorf (1984) and since revised (Christensen & Mendoza, 1986; Jacobson & Traux, 1991), uses a statistical formula for calculating what can be considered a clinically meaningful change. It is most useful when a distribution and "normal" and "dysfunctional" groups are overlapping, such that it is difficult to determine if a specific individual is more like those in the "normal" or "dysfunctional" group. The formula calculates the amount of change pre to post treatment and compares it to the variability of the pretreatment groups taking into account the reliability of the measure used ($RC = \frac{x_2 - x_1}{S_{diff}}$; where X_2 is the post treatment score and X_1 is the pretreatment score; $S_{diff} = \sqrt{(2S_E)^2}$ is the standard error of the difference between the two scores; $S_E = sd\sqrt{(1-r)}$ where sd = standard deviation of the control group, and r = the test-retest reliability of the measure). Thus the calculation of Reliable Change helps determine if the post-treatment score is more similar to the normal population mean or the pretreatment group mean, thus allowing one to determine if meaningful improvement has been made (e.g. became more like the normal group than the dysfunctional group; See Jacobson & Traux, 1991 for a more detailed explanation).

Normative Comparisons examine post treatment data in relation to normative samples using equivalency testing in combination with traditional statistical tests to determine if treated individuals look similar to those not considered to have the targeted problem (Kendall et al., 1999). Specifically, the mean of the treatment group is first examined using equivalency testing to determine if it is within the bounds established around the mean of the normative group based

on chosen delta values. A traditional statistical analysis comparing the treatment group post mean and the normative group mean then determines whether the groups are statistically significantly different. Based on the outcomes of the two analyses the results are categorized into one of four possibilities including a) statistically different/clinically equivalent (Cell I), b) clinically equivalent (Cell II - preferred outcome), c) different (Cell III), or d) equivocal findings (Cell IV; see Kendall et al., 1999, for a more detailed explanation).

Unfortunately, though RCI and Normative Comparisons can both be calculated individually (Baruch, Vrouva, & Wells, 2011), many still use them as group analyses (especially for Normative Comparisons). Interestingly, in our minds, the continued reliance on group statistics as the predominant means for determining clinical significance seems to undermine the greatest benefit to the concept of clinical significance: the opportunity to examine the usefulness of a specific treatment in actually helping individuals get better. However, the greater issue remains the lack of inclusion of any measure of Clinical Significance and the reliance on statistical significance and effect size measures alone.

The importance of the potential differences between group analysis of outcomes and individual analysis of outcomes should not be underestimated. As an example, consider two possible treatment outcomes: A) 20% of treated individuals make very large gains while 80% of the treated individuals and 100% of non-treated individuals make minimal gains; B) 80% of treated individuals make moderate gains and 20% of treated individuals and 100% of non-treated individuals make minimal gains. Most would agree that scenario "B" is a preferred outcome and would want to distinguish such outcomes from each other. It is quite conceivable however, that group analysis alone could result in very similar findings that do not distinguish between the two outcomes. Both scenarios could demonstrate statistical significance, large effect sizes, and even

group clinical significance. Only additional individual analysis would demonstrate that scenario "B" provided superior outcomes from far more people and thus would be preferred to that of scenario "A." While this is especially true in comparing traditional group analyses with individual clinical significance, it is even true in comparing group clinical significance with individual clinical significance, in that group comparisons for both scenarios could show generally good "average" clinical significance, but only individual comparisons would clearly show the 80% vs. 20% difference in improvement. The purpose of the present study is to serve as a comparison of traditional group statistical analysis, effect size, group clinical significance, and individual clinical significance to demonstrate the importance of individual analysis in outcomes research. The research was approved by an institutional review board and was conducted in concordance with the ethical standards of the American Psychological Association.

Method

Participants

Participants included self-referred parents who participated in a Group Behavioral Parent Training Program (BPT) at a university-based clinic between 2009 and 2014. To be included in the analysis, participants needed pre-scores above the clinical cutoff on at least one measure (three total), and at least one post-score. Of the 115 consecutive referrals, 59 (51%) dropped out before completing post measures. Of those remaining, 16 (14%) did not have scores above the clinical cutoffs on any of the three measures, leaving 38 with sufficient pre- and post-data for inclusion in this analysis. The sample was ethnically and financially diverse, giving a representative sample of the diverse community from which they came (see Table 1 for

demographic information). It should be noted that while the sample was diverse in most areas, the majority of participants were female.

Procedure

The researchers assessed progress by having participants complete the Parenting Scale (PS; *n* = 15; Arnold, O'Leary, Wolff, & Acker, 1993), Eyberg Child Behavior Inventory (ECBI; *n* = 22; Eyberg & Pincus, 1999), Parenting Stress Index (PSI; *n* = 24; Abidin, 1995), and demographic information at the beginning, middle, and end of the 9-10 week course. The analysis included only the first and last (either week 5 or 9-10) completion of the rating scales. These three measures give a well-rounded perspective on parent-perceived improvement by touching on stressful parent-child interactions, frequency and intensity of child problem behaviors, and parent responses to misbehaviors.

Both group and individual analyses were used to assess whether participants improved according to PSI, ECBI, or PS scores. The group analyses consisted of standard statistical analysis using t-tests comparing group means, effect size calculations, calculation of group RCIs, and calculation of group Normative Comparisons for each measure. Two individual analyses were examined: an individual calculation of RCI and an individual calculation of Normative Comparisons. We calculated individual RCI as recommended by Jacobson & Traux (1991; $\mathrm{RC} = \frac{x_2 - x_1}{S_{\mathrm{diff}}}$). For interpretive analysis, an RC score greater than 1.96 is considered clinically significant. We calculated individual normative comparisons using the same formulas recommended by Kendall et al. (1999), but substituted the individual score in place of the sample mean of the clinical group. For interpretive analysis, see the cell categorizations listed above. For both RCI and Normative comparisons, the normative group information was obtained from

the normative sample information provided by the publishers or other normative data for each of the three measures

** Include Table 1 here **

Results

Statistical group analysis of pre-post scores on each measure revealed a significant difference and large effect sizes for the PS, ECBI and PSI (see Table 2). These results suggest that following BPT there was a significant decrease in scores across the three measures.

As recommended by Jacobson et al. (1999) and Kendall et al. (1999), we conducted further analysis to determine group clinical significance using RCI and Normative Comparisons. To go beyond group analysis, we examined improvement within individuals using RCI and Normative Comparisons as noted above. The results for all group and individual analyses are found in Table 2.

Of note, though all three group statistical analyses resulted in significant improvements with large effect sizes, only one measure demonstrated a reliable change (PSI), and none of the three measures were equivalent to the normative comparison groups based on group Normative Comparison. Individual analysis demonstrated that on only one measure did over half of the participants make a reliable change (PSI = 54%); very few participants were statistically equivalent post treatment to the normative comparison groups (highest percentage was ECBI = 18%). The fact that individual analysis using RCI and Normative Comparisons yielded results that differed in varying ways from traditional statistical analysis is notable. Higher effect size measures or even RCI calculations did not consistently equate to a higher percentage of individuals that improved.

** Include Table 2 here **

Discussion

The purpose of the present study was to compare the results of group statistical analysis with those of group RCI and Normative Comparisons, as well as to demonstrate the added utility of individual analysis of results based on RCI and Normative Comparisons. Our findings first corroborate the concern that simple statistical analysis of group treatment outcomes research does not adequately convey the nature of outcomes, even when effect sizes are included. Across all three measures, group analysis found large improvements based on effect size. Analysis of clinical significance, however, showed that only one of the three measures demonstrated a reliable change, and none of the three measures demonstrated a return of the treated individuals to the normal range based on the normative samples. Thus while the group statistical analysis and effect size measures would lead one to conclude a strong positive effect from treatment, analysis of clinical significance suggests that more caution should be used in drawing conclusions. Treated individuals made meaningful improvements in decreasing parental stress, but improvements in child behavior and parenting skill were more moderate and did not meet the cutoff for meaningful change. None of the measures suggested treated individuals would have returned to being within the normative range on these measures following treatment.

Even these more accurate group findings of reliable change and normative comparison still do not adequately tell the story of how many or what percentage of individuals improved. When examining treatment outcomes, it is ultimately individuals and their improvement that determine the worth of a specific treatment. The current findings demonstrate that large changes in some individuals often give a skewed conclusion of real outcomes when only traditional group analysis is used. Specifically, on the PSI, RCI analysis demonstrated a reliable decrease in scores. However, individual analysis of meaningful change on the PSI showed that 54% of

treated individuals made a reliable change. In our minds, a meaningful improvement by slightly over half of participants leads to a different interpretation of results than would a significant finding for a group RCI improvement. In this case, large improvements by some individuals masked little to no improvement in others. Of further concern when comparing group vs**.** individual analysis using RCI is that differences are variable across measures. While the group RCI for the ECBI approached the 1.96 cutoff (RCI = 1.85), only 27% of participants showed meaningful improvement. On the other hand, for the PS, though the group RCI is lower (RCI = 1.51), just under twice as many participants made meaningful improvements (47%). We conclude that the individual analysis of RCI and Normative Comparisons provide the most accurate and useful information when drawing conclusions about he effectiveness of a treatment. Individual comparisons highlight the variable impact of treatments on individuals that are often masked in group analysis.

Because the criteria are more stringent for Normative Comparisons (return to normal functioning, as opposed to make meaningful improvements), we see even lower numbers for the individual normative comparisons. Only 18% of participants returned to the normative sample range on the ECBI, 8% on the PSI, and zero individuals on the PS.

Limitations to the present study include the high attrition rate in the program, which prevented the analysis of the majority of the participants. The study also includes only parent-report data, which are subject to self-report bias, though the real purpose of the present analysis is more to demonstrate how data are analyzed and presented than to establish actual findings on the efficacy of BPT. The relatively low sample size may exacerbate the differences noted between group and individual results; however, such differences are possible even with very large samples and the sample size in the present study are similar to those of other BPT research.

While some research on BPT examines individual outcomes (See those reviewed by Eyberg, Nelson, & Boggs, 2008), most studies use simple statistical significance to determine outcomes. Across other areas, the use of Clinical significant continues to increase, with some journals requiring its inclusion for consideration (La Greca, 2005). Still intervention studies continue to be published with no mention of clinical significance or individual findings. We believe that intervention treatment outcome research would benefit from consistent individual analysis. Individual analysis requires additional work beyond group analysis, but is fairly straightforward and does not require additional collection of data. Such analysis will help clarify the true clinical benefit that results from treatment. Group analysis is too sensitive to the influence of large individual improvements by a subset of participants. Beyond better clarifying the true nature of benefit from treatment on an individual basis, individual analysis also better opens avenues for determining what leads to benefit for some individuals and not others. We strongly recommend the use of individual analysis such as the calculation of individual RCI scores as well as individual calculation of Normative Comparisons in all treatment outcomes research.

WORD COUNT: 3053 (including title, abstract, references, and tables)

Acknowledgements: No funding was provided for the current research. The authors have declared that they have no competing or potential conflicts of interest. The authors share responsibility for the content of the manuscript.

Table 1

Demographics for Included Participants

	Total
***n* (Parents)**	25
Gender (%)	
Male	20
Female	80
Parent ethnicity (%)	
Hispanic	44.7
White	26.3
Asian/pacific islander	5.2
African-American	7.9
Other	5.3
Married/partner	52.7
Parent education (%)	
College graduate	28.9
Some college	36.8
Income (%)	
$40,000 and above	45.2
Below $40,000	54.8

Table 2

Group and Individual Analyses across the ECBI, PSI, and PS

		Group				Individual	
Measure	*n*	*t*	*d*	RCI	Normative Classification	RCI % above 1.96	Normative % Clinically Equivalent
ECBI	21	3.68**	.80	1.85	Cell III	27.3	18.2
PSI	23	4.13**	.89	2.93	Cell I	54.2	8.3
PS	14	3.67**	.95	1.51	Cell I	46.7	0

References

Abidin, R. (1995). Parenting Stress Index Manual (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Arnold, D. S., O'Leary, S. G., Wolff, L. S., & Acker, M. M. (1993). The parenting scale: A measure of dysfunctional parenting in discipline situations. *Psychological Assessment, 5*, 137-144. doi:10.1037/1040-3590.5.2.137

Barlow, D. H. (1981). On the relation of clinical research to clinical practice: Clinical issues, new directions. *Journal of Consulting and Clinical Psychology, 49*, 147-155. doi:10.1037/0022-006X.49.2.147

Baruch, G., Vrouva, I., & Wells, C. (2011). Outcome findings from a parent training programme for young people with conduct problems. *Child and Adolescent Mental Health, 16*, 47-54. doi:10.1111/j.1475-3588.2010.00574.x

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*, 7-18. http://dx.doi.org/10.1037/0022-006X.66.1.7

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy, 17*, 305-308. doi:10.1016/S0005-7894(86)80060-0

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

Eyberg, S. M., Nelson, M. M., & Boggs, S. R. (2008). Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. *Journal of Clinical Child and Adolescent Psychology, 37*, 215-237. doi:10.1080/15374410701820117

Eyberg, S. M., & Pincus, D. (1999). *Eyberg Child Behavior Inventory and Sutter–Eyberg Student Behavior Inventory—Revised: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Jacobson, N. S., Follette, W. C, Revenstorf, D., Baucom, D. H., Hahlweg, K., & Margolin, G. (1984). Variability in outcome and clinical significance of behavioral martial therapy: A reanalysis of data. *Journal of Consulting and Clinical Psychology, 52,* 497-504. http://dx.doi.org/10.1037/0022-006X.52.4.497

Jacobsen, N. S., Roberts, L. S., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application and alternatives. *Journal of Consulting and Clinical Psychology, 67,* 300-307. http://dx.doi.org/10.1037/0022-006X.67.3.300

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19. http://dx.doi.org/10.1037/0022-006X.59.1.12

Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification, 1,* 427-453. doi:10.1177/014544557714001

Kendall, P. C., & Grove, W. M. (1988). Normative comparisons in therapy outcome. *Behavioral Assessment, 10,* 147-158.

Kendall, P. C., Mars-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 285-299. doi:10.1037//0022-006X.67.3.285

Le Greca, A.M. (2005). Editorial. *Journal of Consulting and Clinical Psychology, 73,* 3-5. doi: 10.1037/0022-006X.73.1.3

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49*, 156-167. http://dx.doi.org/10.1037/0022-006X.49.2.156