The purpose of this editorial is to renew the call for assessing and documenting clinical significance/social validity in our intervention research. Clinical significance/social validity needs to be documented at three levels: (1) the goals of our intervention studies are relevant to interested parties, such as children, parents, referral sources, and third party payers; (2) the intervention procedures are acceptable to interested parties and are therefore more likely to be adopted if effective; and (3) the effects of our interventions are satisfactory to consumers and meet some standard of achieving clinically significant effects. This editorial describes methods for assessing clinical significance/social validity at all three levels.

A statistically significant treatment effect gives us confidence that an effect is not due to chance. However, a statistically significant effect does not inform us about the size, importance, or clinical significance of the effect. Even “effect sizes” (using Cohen’s convention of $r = .10$ as small, $r = .30$ as medium, and $r = .50$ as large; Cohen, 1988), do not determine if the threshold for clinical significance has been reached. For example, the effect size for low-dose aspirin used to prevent first heart attacks was $r = .02$, which is less than a “small” effect size according to Cohen. However in that study, the sample size was large (22,071) and the risk reduction was 44% ($p \leq .00001$). Consequently, the trial was prematurely stopped due to the obvious benefits of low-dose aspirin (Steering Committee of the Physicians’ Health Study Research Group, 1988).

There have been previous admonitions published in this journal about the need to document clinical significance/social validity in intervention research in pediatric psychology (Drotar, 1997, 2008, 2009; Drotar & Lemanek, 2001). The purpose of this editorial is to renew that call for assessing and documenting clinical significance/social validity in our intervention research. My objectives are to (1) define clinical significance and social validity, and (2) describe ways to assess clinical significance and social validity using examples from the literature.

Definitions of Clinical Significance and Social Validity

“The clinical significance of a treatment refers to its ability to meet standards of efficacy set by consumers, clinicians, and researchers” (Jacobson & Truax, 1991, p. 12). The term social validity comes out of the applied behavior analysis tradition and is a broader and more inclusive term than clinical significance. Social validity is assessed at three levels (Kazdin, 1977; Wolf, 1978):

- **Goals:** Are the specific goals of treatment what society wants? Do they focus on goals relevant to interested parties such as children, parents, referral sources, and third party payers?
- **Procedures:** Are the treatment procedures acceptable to consumers in terms of costs, ethics, and practicality?
- **Effects:** Are effects of an intervention satisfactory to consumers?

Assessing and Documenting Clinical Significance/Social Validity

**Treatment Goals**

One method for ensuring that treatment goals are relevant is to ask consumers to choose their own goals and then use goal attainment scaling (GAS) to rate changes in these goals after treatment. Goal attainment scaling is a statistically validated and reliable measure of valued goals chosen by consumers to represent their definition of improvement in their health (see Fisher, 2008 for an example with chronic pain patients).
Prior to developing interventions, we can also obtain information from consumers that will help us focus our interventions on relevant goals. For example, two surveys of parents whose children have juvenile arthritis found that adherence to therapeutic exercise was more problematic for the children than taking medications, suggesting that we focus adherence interventions on this regimen component (Hayford & Ross, 1988; Rapoff, Lindsley, & Christophersen, 1985). In addition, there is a growing body of literature assessing barriers to adherence among children with chronic illnesses and their parents (Modi & Quittner, 2006; Modi et al., 2009; Reikert & Drotar, 2002). Interventions can be targeted to specific barriers identified by children or parents (Rapoff, in press).

**Treatment Procedures**

We may have effective interventions but if children and parents do not find them acceptable, they will not use them. The social validity of treatment procedures can be assessed by asking parents and children to provide Likert-type ratings of acceptability and asking if they would recommend the intervention to others. This has been done by parents for a treatment program for bedtime problems (Burke, Kuhn, & Peterson, 2004), by adolescents with diabetes and their parents for a conflict resolution program (Wysocki et al., 1997), and by pediatricians, parents, and pediatric psychologists for a thumb sucking intervention (Friman & Leibowitz, 1990). In terms of specific techniques, such as positive reinforcement and time-out, those designed to increase positive behaviors are rated as more acceptable than those designed to reduce inappropriate behaviors (Calvert & Johnston, 1990). Treatments are also rated as more acceptable when they achieve rapid success, when they have few side-effects, and when they are not as complex and time intensive (Calvert & Johnston, 1990).

Kazdin (1980) was the first to develop and validate a generic instrument (the Treatment Evaluation Inventory; TEI) for assessing treatment acceptability. The TEI is a 15-item questionnaire with items rated on a 7-point Likert scale. The TEI or modifications of it have been used in many studies to rate treatment acceptability (Kelley, Heffer, Gresham, & Elliott, 1989; Reimers, Wacker, & Koeppel, 1987; Spirrison, Noland, & Savoie, 1992; Witt & Martens, 1983).

**Treatment Outcomes**

The most ubiquitous assessment of clinical significance/social validity is at the level of treatment outcomes. The most straightforward criterion for determining significance is if treatment increased survival rates (e.g., greater adherence lead to lower mortality rates in the treatment of HIV/AIDS). Also if a treatment eliminates a problem, no one would argue that a clinical significant effect has been obtained. For example, Frank, Spirito, Stark, and Owens-Stively (1997) showed that sleep walking could be eliminated in three children using a scheduled awakening procedure. Generally, survival or cures are not reasonable goals for most of our interventions in pediatric psychology.

**Subjective Judgments or Observations**

We can ask relevant consumers to make retrospective judgments about whether they improved, stayed the same, or became worse following treatment (Norman, Sloan, & Wyrylicz, 2003). These can be as fine grained as researchers would like them to be, for example, using a 7-point Likert scale to rate the degree to which the person improved. We can also ask judges to rate observations for evidence of changes following intervention. For example, Finney, Rapoff, Hall, and Christophersen (1983) had teachers and graduate students rate randomly chosen videotapes at baseline and post-treatment for the presence of “distracting” behaviors of two adolescents treated for tic disorders. The ratings were much lower at post-treatment which was in agreement with objective coding of tic behaviors by trained observers.

**Reliable Improvements**

Before comparing treated clients to normative groups, some have argued that we need to first determine if the magnitude of change for individual patients treated is statistically reliable. The most popular index for doing this in the literature is the Reliable Change Index (RCI) proposed by Jacobson & Truax (1991). The formula for calculating the RCI is as follows:

\[
\text{RCI} = \frac{X_{\text{post}} - X_{\text{pre}}}{S_{\text{diff}}} 
\]

where \(X_{\text{post}}\) is the person’s post-test score, \(X_{\text{pre}}\) is the person’s pre-test score, and \(S_{\text{diff}}\) is the standard error of the difference between two scores.

The change is considered reliable or unlikely due to measurement error if the RCI is greater than 1.96 (Ogles, Lunnen, & Bonesteel, 2001).

Others have argued for using the Standard Error of Measurement (SEM) to determine if a reliable and meaningful change has occurred (Eisen, Ranganathan, Seal, & Spiro, 2007; Guyatt, 2000). The SEM is the standard deviation (SD) of an individual score which is calculated by multiplying the SD for a sample by the square root of one minus its reliability coefficient and an SEM of 1.
or higher is said to be indicative of a clinically important change (Eisen et al., 2007).

**Normative Comparisons**

Normative data have been collected on health-related quality of life measures for healthy and ill children (e.g., Varni, Seid, Knight, Uzark, & Szer, 2002; Varni, Seid, & Kurtin, 2001), for observational measures of social interaction for preschool children (Greenwood, Walker, Todd, & Hops, 1981), and observational measures of mealtime behaviors of families of children with cystic fibrosis and families of children with no chronic illness (Spieth et al., 2001). Also, widely used instruments for assessing emotional and behavioral functioning, such as the Child Behavior Checklist (CBCL), have predetermined cutoff scores based on extensive normative data to indicate clinically significant levels of problem behaviors. For example, March, Spence, and Donovan (2009) used a T score of <65 on the internalizing subscale of the CBCL to evaluate whether clinically significant change occurred following an internet-based cognitive-behavioral treatment for child anxiety disorders. They did an analysis with 41 of their 73 participants who had pretreatment CBCL internalizing T scores greater than 65 and found that 58% in the treatment group were less than 65 after treatment compared with 46% in the control group. Thus, following treatment we would want our patients to resemble their healthy or normal peers or maybe no longer meet criteria for a psychiatric disorder that prompted treatment (Kazdin, 2008).

The problem with this approach is that normative data and standardized cutoff scores may not exist or may be undesirable. Take the example of adherence to regimens for chronic illness. I calculated the mean level of adherence, SD, and range (R) for studies reviewed in the first chapter of my book on three different diseases (Rapoff, in press, Table 1.2). Across 18 studies examining adherence to medications for asthma the mean rate was 50% (SD = 22%, R = 10–85%); across 17 studies examining adherence to a gluten-free diet for celiac disease the mean rate was 69% (SD = 19%, R = 30–95%); and across 7 studies examining adherence to antiretroviral drugs in the treatment of HIV/AIDS the mean rate was 75% (SD = 14%, R = 54–93%). What are we to make of these data? Is this an adequate level of adherence considering that the average rate of adherence across all regimens for pediatric chronic diseases is estimated to be between 50 and 55% (Rapoff, in press)? The answer is that for most diseases and regimens we do not know the minimum level of adherence necessary to achieve an acceptable therapeutic response (Gordis, 1979; Haskard, DiMatteo, & Williams, 2009). The standard in the literature has been an adherence rate of 80% or better but this was based on early data on adult hypertensive patients where that level of adherence to antihypertensive medications resulted in adequate blood pressure control. For diseases like HIV/AIDS, an adherence level of 90 or 95% may be necessary to achieve adequate suppression of the virus (Rapoff, in press).

**Standards Determined by Expert Consensus**

In the pain literature, extensive studies and reviews have been done to determine clinically significant changes in measures of pain intensity. The consensus from many studies by different investigators is that a 2-point change on a 0–10 pain rating scale is a meaningful change in pain intensity (Farrar, Portenoy, Berlin, Kinman, & Strom, 2000; Farrar, Young, LaMoreaux, Werth, & Poole, 2001). Others have suggested a 50% reduction in pain scores from baseline to post-treatment for individual patients (Hicks, von Baeyer, & McGrath, 2006). Other intervention studies used expert opinion to establish standards for meaningful clinical change, such as pedodontists recommending plaque levels less than or equal to 20% for children (Dahlquist et al., 1985) and an NIH consensus panel recommending optimal calcium intake levels of 1500 Ca per day for children (Stark et al., 2005).

**Use Single Subject Designs**

Although randomized, between-groups, controlled clinical trials are considered the “gold standard” for experimentally evaluating the efficacy of treatments, there is a long tradition in medicine and psychology of investigating the effects of interventions at the individual level using single-subject designs (Barlow, Nock, & Hersen, 2009). Single-subject designs offer a number of advantages over traditional group designs: (1) they provide flexibility in the choice of independent variables and allowance for changes in these over the course of a study (if something is not working, an intervention can be modified and introduced as a new condition); (2) they accommodate small sample sizes (appropriate for studying rare conditions or smaller sample sizes available at any one site); (3) they are better at exposing individual variability in outcome measures and if an individual patient has made a clinically meaningful change; (4) they produce results that are more easily understood by clinicians (who work at the level of individual patients); and (5) they are recognized as legitimate designs that can help to establish empirically validated treatments and evidence-based practices (Barlow et al., 2009; Rapoff & Stark, 2008). The most common designs are the reversal and multiple baseline designs that can be
used with one patient or more (see Barlow et al., 2009 for the authoritative book on single-subject designs).

For example, we did a study with a 14-year-old male with juvenile rheumatoid arthritis who was non-adherent to a non-steroidal anti-inflammatory medication (Rapoff, Purviance, & Lindsley, 1988). Table I shows the results of the study by phase using a reversal design in terms of the effect on adherence and a clinical outcome (active joint count). It can clearly be seen that the token system intervention improved adherence. In addition, improvements in adherence were also associated with a change in a meaningful clinical parameter. At the level of the individual, single-subject designs allow us to determine if a noticeable change has occurred as the result of interventions.

Assess Changes in Clinically Relevant Collateral Measures
As illustrated in Table I, we should not only assess primary outcomes (such as measures of adherence) but other clinically relevant measures to see if they change in the expected direction and in tandem with changes in the target measures. For example, in adherence intervention studies, we should assess clinical outcome measures, such as pulmonary function for children with asthma, and quality of life measures (Rapoff, in press).

Conclusions
In designing and then submitting intervention studies to the Journal of Pediatric Psychology, I would strongly recommend that investigators include a measure of clinical significance/social validity at least at the level of treatment outcomes. Documenting the clinical significance of intervention studies is important to the growth of our field and is a priority for the Journal of Pediatric Psychology.

Hopefully, suggestions in this editorial will be helpful in this endeavor. We need to start somewhere and as Alan Kazdin (2008) has said: “Let us not begin with what cannot be done” (p. 153).

Conflict of interest: None declared.

Received and revisions received October 10, 2009; accepted October 13, 2009

References


Varni, J. W., Seid, M., & Kurtin, P. S. (2001). The PedsQL™ 4.0: Reliability and validity of the Pediatric Quality of Life Inventory™ version 4.0 generic core scales in health and patient populations. Medical Care, 39, 800–812.

