Published in final edited form as:

J Consult Clin Psychol. 2012 December ; 80(6): 1103–1107. doi:10.1037/a0029991.

Agreement between therapists, parents, patients, and independent evaluators on clinical improvement in pediatric obsessive compulsive disorder

Adam B. Lewin, Ph.D., ABPP^{1,2,*}, Tara S. Peris, Ph.D.², Alessandro S. De Nadai, M.A.¹, James T. McCracken, M.D.², and John Piacentini, Ph.D., ABPP²

¹Departments of Pediatrics & Psychiatry, Rothman Center for Neuropsychiatry, University of South Florida College of Medicine, Saint Petersburg, FL

²Division of Child and Adolescent Psychiatry, UCLA Semel Institute for Neuroscience and Human Behavior, Los Angeles, CA

Abstract

Objective—Independent evaluators (IE) are employed widely in clinical trials to make unbiased determinations of treatment response. By virtue of being kept blind to treatment condition, however, IEs are also kept unaware of many pertinent clinical details that are relevant for decisions about clinical improvement. In this study, agreement among raters (children, parents, therapists, and IEs) about treatment response over the course of a 14-week clinical trial for pediatric obsessive compulsive disorder (OCD) was examined in order to determine the utility of non-blind clinician and patient ratings of treatment response.

Method—Participants were 71 youth (mean age= 12.2 years; 63.4% female) with a primary diagnosis of OCD and their parents participating in a psychotherapy trial. Independent evaluators provided response ratings (Clinician's Global Impressions-Improvement Scale; CGI-I) at weeks 4, week 8 and 14 and therapists, children and parents completed independent CGI-I ratings at weeks 2, 4, 8 and 14.

Results—Nonlinear mixed models revealed differences in rating parties, with therapists and IEs slower to rate treatment improvement compared to children and parents and growth curve models suggested that therapists and IEs produced generally consistent ratings. In addition, no evidence was found for an evaluator-by-treatment interaction, indicating that raters displayed these differences consistently across both active and placebo conditions.

Conclusions—Youth and parents may be able to provide accurate ratings of global improvement; non-blinded treating clinicians (with training in research-oriented assessment) can offer global improvement ratings commensurate with blinded IEs. Findings suggest that alternatives (or additions) to the blinded-IE model may be appropriate for assessing global improvement, especially with the growing emphasis on dissemination and effectiveness trials.

Correspondence concerning this article (and requests for an extended report of this study) should be addressed to Adam B. Lewin, Ph.D., ABPP, University of South Florida, Rothman Center for Neuropsychiatry, 880 6th Street South, Suite 460, Box 7523, Saint Petersburg, Florida 33701, USA, alewin@health.usf.edu, fax 727-767-7786.

*This research was conducted at the UCLA Semel Institute, 760 Westwood Plaza, Los Angeles, CA, 90024.

Keywords

OCD; Obsessive-Compulsive-Disorder; Independent Evaluator; Rater Agreement; Child; CBT; Cognitive Behavioral Therapy; Treatment Response

Randomized, double-blind controlled trials are the gold standard in treatment outcome research. In these trials, patients are assigned at random to a treatment condition, and both the researcher and the patient are kept “blind” (i.e., unaware) to which condition is assigned in order to ensure unbiased ratings of improvement at the end of treatment. Research on psychosocial interventions presents a unique challenge for implementing this design in that patients in these studies often cannot be kept blind to treatment condition and the treating clinician, by definition, never can. Whereas in pharmaceutical studies, medication containers can be stricken of labels in order to disguise active compounds from placebo, the presence or absence of therapy cannot be similarly concealed (Carroll, Rounsaville, & Nich, 1994; Wilkins, 1983). Consequently, evaluations can only be single-blind because therapists and patients typically are aware of the patient's assigned condition. The result is that ratings of improvement by treating clinicians are potentially subject to bias that arises from expectations about the treatment (Basoglu, Marks, Livanou, & Swinson, 1997; Seligman, 1995).

Accordingly, outcome research for psychological interventions has adopted the use of independent evaluators (IEs), who are clinicians trained to criterion in multi-informant and evidence-based assessment, kept blind to experimental condition, and employed to evaluate clinical outcomes. Across trials for various child mental health conditions, a critical role for IEs is to rate treatment response. In other words, did the patient get better and is the magnitude of change clinically meaningful? Making this determination can be challenging, as IEs must decide whether to base it on change in diagnosis, reduction in symptoms, change in functional impairment, or all three. One of the most widely-used tools for assessing clinical improvement is the Clinician Global Impressions of Improvement scale (CGI-I; Guy 1976). Considered by many to be the gold standard for designating responder status, the CGI allows for a single rating based on an integration of data in the context of clinical experience. Most notably, it is equally sensitive to clinical improvement compared to dimensional instruments (Kellner & Uhlenhuth, 1991) and has been used as the primary outcome measure in numerous youth clinical trials (e.g., POTS, 2004; TADS, 2007).

Notwithstanding the potential benefits of IE ratings, there are limitations of blind evaluators (Basoglu, Marks, Livanou, & Swinson, 1997,; Carroll et al., 1994; Roll et al., 2004). Specifically, the circumscribed role of IEs limits the amount of information they can use to make informed assessment decisions, as those charged with determining a CGI rating often have only limited access to relevant clinical details. For example, the IE cannot judge clinical improvement based on behavioral observations during treatment tasks or in-session information regarding compliance with therapy ‘homework’ assignments. This is particularly the case in exposure-based therapies (e.g., for obsessive-compulsive disorder [OCD] and other anxiety disorders) where observation of a child's habituation to anxious stimuli, avoidance, degree of ritualization, and other in-session behaviors provide valuable response-

status information. Indeed, research by Rabinowitz and colleagues suggests that for these reasons and others, independent clinical evaluators may not be able to detect the full extent of a patient's symptom relief (Rabinowitz et al., 2008). Moreover, relying on a single [IE] rating is a crucial limitation in assessment (De Los Reyes et al., 2011). Another consideration is that, despite a growing interest in understanding the *process* of change over the course of therapy (as opposed to simply examining pre-post change), requiring *additional* IE evaluations limits the feasibility of modern longitudinal techniques (e.g., multilevel modeling) given that the added time burden (and cost). In sum, the limited scope of an IE improvement rating may be enhanced by utilizing improvement ratings from other informants (e.g., the treating clinician, parent, and child).

Despite calls for integrating information from multiple respondents when determining response status, existing analyses of inter-rater agreement have concentrated on comparing diagnostic agreement (Grills & Ollendick, 2002) rather than agreement on improvement. This oversight is puzzling in the context of clinical trials research where *improvement* is often the primary outcome. Poor integration of information across informants means that a single rater may ultimately make the determination of treatment response, with or without complete information. Further, studies of cross-informant agreement on ratings of clinical improvement are virtually non-existent. To our knowledge, there are no empirical studies of child and parent proficiency in rating global mental health improvement, despite a growing National Institutes of Health initiative towards patient-rated outcomes (Cella et al., 2007). Finally, existing research examining inter-rater agreement has tended to focus on a single assessment time-point (Jensen-Doss & Weisz, 2008), in contrast to a methodologically-preferable longitudinal analysis of agreement (Singer & Willett, 2003). Thus, there remains a need for longitudinal research examining patient, therapist, and IE agreement on clinical improvement across the course of treatment.

The aim of the present study is to address the lack of empirical basis in identifying optimal informants of outcome by examining agreement in ratings of clinical response by four different raters (IE, therapist, parent, and child) in the context of a 14-week randomized controlled psychotherapy trial. Associations between the IE-rated CGI-I along with therapist, parent and child CGI-I ratings are examined across time and randomly assigned psychotherapy conditions. To our knowledge, this is the first systematic examination of the relationship among IE, therapist, parent, and child outcome ratings in the context of a highly controlled treatment trial. While therapists have traditionally been considered undesirable reporters of outcome due to bias, the present clinical trial had strong standardization of procedures, which we believe may serve to calibrate therapist ratings to be more in tune with those made by IEs. Indeed, it has been asserted that through careful standardization, threats to validity such as allegiance bias can be alleviated (Wilson, Wilfley, Agras, & Bryson, 2011). Thus, we aim to test the hypothesis that under these carefully managed conditions, therapists may offer assessments of improvement that are concordant with independent raters. Additionally, patient reports of symptom levels have consistently displayed differences with those of trained professionals (Petkova, Quitkin, McGrath, Stewart, & Klein, 2000). Given their reduced experience with treatment-related change for OCD, we predict that parents and youths will report higher improvement rates, as they may interpret

any change as substantial given their frustrating experience with what was previously an intractable condition for their family.

Method

Participants and Procedure

Seventy-one treatment-seeking youth with OCD (ages 8 – 17 years, $M = 12.2$ years; $SD = 2.5$ years; 63.4% female) and their parents participated in a randomized controlled treatment outcome study for OCD conducted at a large university-based specialty program (Piacentini et al., 2011). Eligible youth (see Piacentini et al., 2011 for full inclusion criteria) were randomized to 12 sessions of either family based CBT (FCBT) or a comparison condition consisting of psychoeducation about OCD and systematic relaxation training (PRT). The Clinical Global Impression-Improvement (CGI-I; Guy, 1976) was completed by study IEs at week 4, week 8 and following the final therapy session at week 14. The CGI-I is a single-item rating of clinical improvement from baseline with scores ranging from 1 (very much improved) to 7 (very much worse). The CGI-I has been extensively validated across a wide variety of adult and child psychiatric disorders, including OCD. By convention, CGI-I ratings of 1 or 2 (much improved) are used to define a positive treatment response. The child's clinical improvement was rated by the parent and child using a scale identical to the CGI-I, except that directions were worded appropriately for each type of respondent (e.g., “How has your OCD changed since you started treatment?” for children, or “Please rate any change in your (the) child's OCD using the scale below” for parent and therapist, respectively). Therapists, children, and their parents completed global improvement ratings at weeks 2, 4, 8 and 14. Parents and children completed their rating forms independently and outside the presence of the therapist or IE. IEs did not have access to the therapist's or the parent or child improvement rating forms or to any other treatment-related materials. All IEs and study therapists had extensive experience in the assessment of OCD and completed a rigorous training and certification process (as outlined in Piacentini et al., 2011).

Data Analytic Plan

Differences among evaluators in response rates on the CGI were evaluated through PROC NLMIXED in SAS 9.2 using maximum likelihood estimation, where the unconditional means model, the unconditional growth model, and the conditional growth model were tested sequentially (Singer & Willett, 2003). This framework evaluates change over time via a model building process which sequentially adds predictors of interest. These models implement an intent-to-treat approach, using only the observed data provide information in creating patient change trajectories. Random effects were estimated for the overall intercept and time (with all observations being nested within participants), to evaluate if there were differences among evaluators in their ratings of treatment response (CGI – 2) over all time points. We evaluated overall differences among all four raters over time as well as a rater $\times$ time interaction, to see if rater trajectories varied over time. A treatment group $\times$ evaluator interaction was also evaluated to see if treatment group affected how evaluators rated patient improvement. Improvement of sequential model fit was tested with χ^2 difference tests of deviance values, where a significant χ^2 value reflects additive value in adding a predictor relative to the previous model. Pseudo- R^2 estimates for models were calculated where the

unconditional means model was used as a reference for all subsequent models fit (Nagelkerke, 1991). Post-hoc χ^2 tests of independence were considered in the case of a significant evaluator $\times$ time effect, in order to assess for differential reporting by evaluating party at each individual time point. If a χ^2 test of independence indicated a significant difference at a specific time point, this would indicate the evaluators differed in their ratings of improvement at that specific time point.

Results

In the unconditional means model, significant effects were found for the test of the fixed ($t(70) = -6.67, p < .001$) and random intercept ($t(70) = 3.71, p < .001$), indicating that overall scores differed from zero, and that there was heterogeneity of residuals around the overall mean that could be explained by other variables (e.g., change over time). The unconditional growth model was an improvement over the unconditional means model ($\chi^2(2) = 259.30, p < .05$, pseudo- $R^2 = .37$) and identified both significant fixed ($t(69) = 7.48, p < .001$) and random effects for time ($t(69) = 2.77, p < .01$). This indicated that, on average, patients were rated as improved over time, and that there was some heterogeneity among patients' response rates over time that could be predicted by other variables (e.g., type of evaluator). Entering type of evaluator into the conditional growth model as a predictor improved model fit and predicted some of the previously unclaimed residual variance ($\chi^2(3) = 44.00, p < .05$, pseudo- $R^2 = .42$), including the emergence of a significant effect for type of evaluator ($F(3,69) = 12.64, p < .001$). Adding an evaluator $\times$ time interaction in an additional model resulted in a further improvement in model fit ($\chi^2(3) = 10.00, p < .05$, pseudo- $R^2 = .44$). In this model, both the effect of evaluator ($F(3, 69) = 8.39, p < .001$), and the effect of evaluator $\times$ time were significant ($F(3,69) = 3.11, p = .03$).

Parents and children reported consistently higher rates of treatment response over time, whereas the therapist and IE were more conservative in their initial evaluations of global improvement, but "caught up" over time with later ratings showing some convergence to those for parents and children (Figure 1). Adding treatment group and a treatment group $\times$ evaluator effect to the model did not result in an appreciable improvement in model fit ($\chi^2(4) = 7.01, p > .05$, pseudo- $R^2 = .44$), suggesting that treatment group assignment did not differentially affect the observed differences among evaluators. Post-hoc χ^2 tests found differences among evaluators in rating treatment response at weeks 2 ($\chi^2(2) = 16.11, p < .001$), 4 ($\chi^2(3) = 13.86, p < .01$), and 8 ($\chi^2(3) = 9.79, p < .05$), but not week 14 (e.g., post-treatment) ($\chi^2(3) = 5.16, p = .16$), providing further support for the convergence of the different rater types over time.

Discussion

The present investigation examined cross-informant agreement on ratings of global clinical improvement as measured via the CGI-I. In keeping with our hypotheses, therapists appeared to provide ratings that were consistent with IE ratings. In addition, both therapists and IEs were more conservative in their improvement ratings compared to parents and children, designating responder status more slowly over the course of the study.

The finding that therapists and IEs produce similar ratings of improvement has several important implications. From a research standpoint, it suggests that the treating clinician may be capable of providing objective improvement ratings despite knowledge of (and investment in) the treatments being delivered. Although one might expect the treating clinician to rate improvement more favorably compared to IEs due to their knowledge of treatment condition, our data suggest that, in cases where ratings diverge, therapists may actually be more conservative than IEs in their CGI-I ratings. This may be because their direct ongoing contact with patients provides a more complete clinical picture compared to the brief snapshots of information obtained during IE visits (Sharp & Power, 1998).

Comparable ratings of global improvement between therapists and IEs suggest that in cases where the use of blind evaluators is not possible, therapist ratings may be a reasonable alternative. This is particularly important for effectiveness research, where emphasis is placed on testing interventions in community settings and blind IE ratings can be extremely difficult to obtain (Basoglu et al., 1997; Carroll et al., 1994; Roll et al., 2004). While the use of IEs is certainly desirable, it can be difficult to have the IE complete weekly ratings; should clinicians be able to accurately rate symptoms, statistical modeling of patient change over time could be conducted without the costs/constraints in collecting multiple time points using IEs. In addition, providing training in evidence-based assessment may increase the ability for clinician-evaluators to provide meaningful ratings, consistent with the National Institutes of Health (NIH) roadmap to facilitate and enhance collection of health outcome data (Cella et al., 2007).

Notably, we found no evidence that the observed pattern of differences between raters was affected by treatment group (given the lack of a significant evaluator type-by-treatment interaction). This is important given that one might expect clinicians to favor the active treatment in their CGI-I ratings and downgrade responses to the control condition. Instead, it appears that their trajectory of ratings roughly parallels that of the IEs who were blind to treatment condition. This provides some additional indication that clinicians involved in treatment may be able to accurately gauge progress despite their role in patient care.

The present findings must be interpreted in light of limitations. First, all clinicians had extensive expertise in evidence-based assessment and treatment of OCD and anxiety. Thus, it is possible that the backgrounds and prior experience of some or all of the therapists may have enhanced their ability to maintain reliability with IEs. Additionally, the extent to which our findings are generalizable to broader range of child psychopathology or other age groups is unclear. Symptom reduction over treatment course may impact the convergence of ratings. These limitations notwithstanding, the present study provides valuable information on how parents, children, therapists, and blind evaluators view progress over the course of treatment. Consistent ratings across therapists and IEs suggest that therapists are able to provide relatively objective assessments of treatment progress. Further, parents and youth agreed upon global improvement status across time points and their rating were consistent with the IE at post-treatment.

Our data suggest that blindness may not be essential for making accurate improvement ratings, provided that raters are well-trained and efforts are made to minimize bias. There

may allowance be a more sizable role for patient and clinician raters of global outcome in clinical trials, an important implication as NIH initiatives shift toward dissemination studies (National Institute of Health, 2009) and patient-reported outcome measures (Cella et al., 2007). Replication of the current study and future research assessing the utility of trained, non-blinded raters in studies of different diagnostic groups, other treatment modalities, and from community-clinician samples is needed to extend these findings.

Acknowledgments

This study was supported by NIMH R01MH 58459 (Piacentini), NIMH K23 MH085058 (Peris), and NARSAD Young Investigator Awards (Lewin, Peris). The authors would like to thank Elysse Arnold with for her assistance in preparing this manuscript.

References

- Basoglu M, Marks I, Livanou M, Swinson R. Double-blindness procedures, rater blindness, and ratings of outcome. Observations from a controlled trial. *Archives of General Psychiatry*. 1997; 54:744–748. [PubMed: 9283510]
- Carroll KM, Rounsaville BJ, Nich C. Blind man's bluff: Effectiveness and significance of psychotherapy and pharmacotherapy blinding procedures in a clinical trial. *Journal of Consulting and Clinical Psychology*. 1994; 62:276–280. [PubMed: 8201064]
- Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Rose M. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. *Medical Care*. 2007; 45:S3–S11. 10.1097/01.mlr.0000258615.42478.55 00005650-200705001-00002 [pii]. [PubMed: 17443116]
- De Los Reyes A, Youngstrom EA, Swan AJ, Youngstrom JK, Feeny NC, Findling RL. Informant discrepancies in clinical reports of youths and interviewers' impressions of the reliability of informants. *Journal of Child and Adolescent Psychopharmacology*. 2011; 21:417–424. 10.1089/cap.2011.0011. [PubMed: 22040187]
- Grills AE, Ollendick TH. Issues in parent-child agreement: The case of structured diagnostic interviews. *Clinical Child and Family Psychology Review*. 2002; 5:57–83. [PubMed: 11993545]
- Jensen-Doss A, Weisz JR. Diagnostic agreement predicts treatment process and outcomes in youth mental health clinics. *Journal of Consulting and Clinical Psychology*. 2008; 76:711–722. 2008-13625-001 [pii] 10.1037/0022-006X.76.5.711. [PubMed: 18837589]
- Nagelkerke N. A note on a general definition of the coefficient of determination. *Biometrika*. 1991; 78(3):691–692.
- National Institute of Health. The National Institute of Health Strategic Plan. 2009. Retrieved from <http://www.nimh.nih.gov/about/strategic-planning-reports/index.shtml>
- Petkova E, Quitkin FM, McGrath PJ, Stewart JW, Klein DF. A method to quantify rater bias in antidepressant trials. *Neuropsychopharmacology*. 2000; 22:559–565. S0893133X99001542 [pii] 10.1016/S0893-133X(99)00154-2. [PubMed: 10788756]
- Piacentini J, Bergman RL, Chang S, Langley A, Peris T, Wood JJ, McCracken J. Controlled comparison of family cognitive behavioral therapy and psychoeducation/relaxation training for child obsessive-compulsive disorder. *Journal of the American Academy of Child Adolescent Psychiatry*. 2011; 50:1149–1161. S0890-8567(11)00690-3 [pii] 10.1016/j.jaac.2011.08.003. [PubMed: 22024003]
- Pediatric OCD Treatment Study Team. Cognitive-behavior therapy, sertraline, and their combination for children and adolescents with obsessive-compulsive disorder. *Journal of the American Medical Association*. 2004; 292:1969–1976. 20069-00110.1001 /jama.292.16.1969. [PubMed: 15507582]
- Rabinowitz J, Levine SZ, Medori R, Oosthuizen P, Koen L, Emsley R. Concordance of patient and clinical ratings of symptom severity and change of psychotic illness. *Schizophrenia Research*. 2008; 100:359–360. S0920-9964(07)00463-X [pii] 10.1016/j.schres.2007.10.021. [PubMed: 18077138]

- Roll D, Ray SE, Marcus SM, Passarelli V, Money R, Barlow DH, Gorman JM. Independent evaluator knowledge of treatment in a multicenter comparative treatment study of panic disorder. *Neuropsychopharmacology*. 2004; 29:612–618. 10.1038/sj.npp.1300373 1300373 [pii]. [PubMed: 14702024]
- Seligman ME. The effectiveness of psychotherapy. The Consumer Reports study. *American Psychologist*. 1995; 50:965–974. [PubMed: 8561380]
- Sharp DM, Power KG. Psychologist, patient and general practitioner ratings of outcome of pharmacological and psychological treatments for panic disorder and agoraphobia in primary care. *Behavioural and Cognitive Psychotherapy*. 1998; 26:13–27.
- Singer, JD.; Willett, JB. *Applied longitudinal data analysis. Modeling change and event occurrence*. Oxford University Press; Oxford, United Kingdom: 2003.
- The TADS Team. The Treatment of Adolescent Depression Study (TADS). *Archives of General Psychiatry*. 2007; 2007; 64:1132–1144. [PubMed: 17909125]
- Wilkins W. Failure of placebo groups to control for nonspecific events in therapy outcome research. *Psychotherapy Theory, Research, and Practice*. 1983; 20:31–37.
- Wilson GT, Wilfley DE, Agras WS, Bryson SW. Psychological treatments of binge eating disorder. *Archives of General Psychiatry*. 2011; 67:94–101. 67/1/94 [pii] 10.1001/archgenpsychiatry.2009.170. [PubMed: 20048227]

Figure 1.

Response rate (percent of cases rated as 'Much Improved' or 'Very Much Improved' on the CGI-I) reported by evaluators (parent, child, therapist, and IE) across time.

Table 1

Correlations between Global Improvement Ratings by Evaluator Type

	IE-Therapist	IE-Child	IE-Parent	Therapist-Child	Therapist-Parent	Parent-Child
Week 2	.18	.25	.58	.01	.02	.14
Week 4	.18	.25	.58	.31	.40	.40
Week 8	.24	.25	.37	.48	.74	.40
Week 14	.65	.59	.76	.55	.75	.62

Note. IE = Independent Evaluator; § = IE Clinician Global Impression – Improvement ratings not available until Week 4;

* p < .05;

** p < .01;

*** p < .001