# Supplementary Online Content

Weisz JR, Chorpita BF, Palinkas LA, Schoenwald SK, Miranda J, Bearman SK, Daleiden EL, Ugueto AM, Ho A, Martin J, Gray J, Alleyne A, Langer DA, Southam-Gerow MA, Gibbons RD; Research Network on Youth Mental Health. Testing standard and modular designs for psychotherapy treating depression, anxiety, and conduct problems in youth: a randomized effectiveness trial. *Arch Gen Psychiatry*. Published online November 7, 2011. doi:10.1001/archgenpsychiatry.2011.147.

**eAppendix.** Expansion of Two Elements of the Methods Section
**eReferences**

This supplementary material has been provided by the authors to give readers additional information about their work.

**eAppendix: Expansion of Two Elements of the Methods Section**

**1. Diagnostic Interviewer Training and Reliability of Diagnoses**

The ChIPS[1] and PChIPS[2] were administered separately to youths and their caregivers by interviewers who had completed 32 hours of training, observed multiple expert interviews, conducted ten individual practice interviews, and completed a series of test interviews in three of which they were required to achieve perfect agreement with an expert rater before being certified as study interviewers. Study diagnoses were assessed for inter-interviewer reliability by having 40 randomly-selected interviews (split pre/post and youth/caregiver) coded by a new coder trained to the same standards as study interviewers. Kappa was .94 for overall reliability of child-based diagnosis, .82 for caregiver-based diagnosis. Reliability was also assessed for diagnoses in the categories of disruptive behavior disorders (child, $\kappa = 1.00$, $p < .001$; parent, $\kappa = .74$, $p < .001$), anxiety disorders (child, $\kappa = 1.00$, $p < .001$; parent, $\kappa = .90$, $p < .001$), and depressive disorders (child, $\kappa = 1.00$, $p < .001$; parent, $\kappa = .73$, $p = .001$). All kappas were at or above Landis-Koch[3] standards for substantial agreement.

Combined diagnoses were generated from youth and caregiver report. Because evidence suggests that caregivers may be more accurate reporters of their children's externalizing problems, and children may be better reporters of their own internalizing problems,[4] diagnoses generated by both interviews were accepted, diagnoses generated only by child-report were accepted if the disorders were internalizing (e.g., anxiety or depressive disorders), and diagnoses generated only by parent report were accepted if the disorders were externalizing (e.g., conduct disorder, oppositional defiant disorder), per Silverman-Nelles[5] guidelines.

To assess reliability of the combined diagnoses, two doctoral-level psychologists blind to participant identity independently reviewed 20 randomly selected pairs of ChIPS and PChIPS, generating composite diagnoses for all the pairs. The mean kappa was .84. Reliability was also assessed for composite diagnoses in the three study diagnostic categories of disruptive behavior disorders ($\kappa = .88$, $p < .001$), anxiety disorders ($\kappa = 1.00$, $p < .001$), and depressive disorders ($\kappa = 1.00$, $p < .001$).

**2. Adherence to Condition: UC, Standard, and Modular**

Treatment sessions in all three conditions were coded for *adherence to study condition*—i.e, whether MATCH procedures were well-represented in the modular condition, standard EBT procedures were well-represented in the standard condition, and the UC condition included low levels of MATCH and standard EBT procedures. For the standard condition, we used a microanalytic adherence measure to code the presence or absence of each of the model-specific elements (e.g., relaxation training) of the treatments used in the standard condition (*Coping Cat,*[6,7] *PASCET,*[8-10] *Defiant Children[11]*) in five minute segments throughout the full session. An additional code, "other," indicated a treatment element not specific to the three standard treatment manuals. Because element categories were not mutually exclusive, more than one element could be coded within each increment. For the modular condition, a similar measure was used to record the presence or absence of each of the 31 MATCH modules for anxiety, depression, or disruptive behaviors. An additional item allowed for the coding of "other," indicating a treatment element not found in a MATCH module. For UC, we used randomly embedded UC tapes in both the standard and modular coding samples, and used either the standard or the modular coding systems to assess for the presence/absence of their component elements and to examine the levels of potential overlap in treatment content among conditions (e.g., low instance of standard or modular elements in UC would demonstrate treatment differentiation). Coding was done by two doctoral level coders, six psychology doctoral students, and four post-BA research assistants.

**Coding training**. Coders were trained using the following procedure: After familiarizing themselves with the treatment manuals, coders participated in 12 hours of didactic presentations led by a doctoral-level staff member who was expert in the manual content; in addition, exemplar videos were used to illustrate treatment modules/components. As part of the training process, sessions were viewed, coded, and discussed as a group with the training leader (who also served as the expert coder and reliability standard-see below). Next, all coders, working independently, coded a set of 13 treatment sessions randomly selected from the modular and standard conditions, then met to compare codes and discuss discrepancies. Each coder was then assigned the same set of seven randomly selected treatment sessions and coded independently without discussion. The sum of five-minute increments that were positive for the presence of a given modular or standard treatment element was calculated for each coder on each session tape, resulting in a set of sums reflecting the frequency of use of each individual modular or standard treatment element. Given the limited number of tapes in the reliability sample, the sums were converted to dichotomous present/absent scores denoting whether a treatment element was present in each recording. Inter-coder reliability was calculated for the double-coded tapes using

kappa. All coders needed to demonstrate acceptable agreement with an expert rater ($\kappa > .70$; M $\kappa = 1.00$) in order to meet training standards and proceed to coding study sessions. All coders met this standard.

**Coding reliability and evidence of adherence to condition**. To assess reliability of the coding after the training was complete, we randomly selected 20 session tapes (10 modular, 10 standard) from the total coded sample and had these double-coded using the expert rater. The mean kappa was .88, and the median kappa was 1.0, indicating that the coding system was reliable across coders. To assess adherence, we used stratified random sampling of session recordings for coding, to ensure coverage of all study sites and initial treatment focus on each of the three problem areas (anxiety, depression, and conduct). Some 309 treatment sessions were coded, encompassing the standard, modular, and UC conditions. We compared the mean percentage of total session content (i.e., of 5-minute coded segments) that was model-specific (i.e., standard manual content and modular manual content), within the three study conditions. We also calculated the percentage of session content not described in either the standard or the modular manuals. In the standard condition, 92.75% of session content was comprised of the model-specific treatment elements used in Coping Cat, PASCET, and Defiant Children. In the modular condition, 82.95% of session content was consistent with that specified in the MATCH protocol. In UC, only 8.47% of session content was consistent with that described in either the standard or modular manuals. Instead, 91.53% of UC session content was "Other."

References

1. Weller EB, Weller, R.A., Rooney, M.T., & Fristad, M.A. *CHIPS-Children's Interview for Psychiatric Syndroms*. Washington, DC: American Psychiatric Press; 1999.

2. Weller EB, Weller RA, Rooney MT, Fristad MA. *Children's Interview for Psychiatric Syndomes-Parent Version (P-ChIPS)*. Washington, DC US: American Psychiatric Association; 1999.

3. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics.* 1977;33(1):159-174.

4. Kazdin AE. Psychiatric diagnosis, dimensions of dysfunction, and child behavior therapy. *Behav Ther.* 1983;14(1):73-99.

5. Silverman WK, Nelles WB. The Anxiety Disorders Interview Schedule for Children. *J Am Acad Child Adolesc Psychiatry.* 1988;27(6):772-778.

6. Kendall PC. Treating anxiety disorders in children: Results of a randomized clinical trial. *J Consult Clin Psychol.* 1994;62(1):100-110.

7. Kendall PC, Kane, M., Howard, B, & Siqueland, L. *Cognitive-behavioral treatment of anxious children: Treatment manual*. Ardmore, PA: Workbook Publishing; 1990.

8. Weisz JR, Moore PS, Southam-Gerow MA, Weersing VR, Valeri SM, McCarty CA. *Therapist's Manual PASCET: Primary and secondary control enhancement training program*. 3 ed. Los Angeles: University of California; 2005.

9. Weisz JR, Southam-Gerow MA, Gordis EB, et al. Cognitive-behavioral therapy versus usual clinical care for youth depression: An initial test of transportability to community clinics and clinicians. *J Consult Clin Psychol.* 2009;77(3):383-396.

10. Weisz JR, Thurber CA, Sweeney L, Proffitt VD, LeGagnoux GL. Brief treatment of mild-to-moderate child depression using primary and secondary control enhancement training. *J Consult Clin Psychol.* 1997;65(4):703-707.

11. Barkley RA. *Defiant children: A clinician's manual for assessment and parent training*. 2 ed. New York: The Guilford Press; 1997.