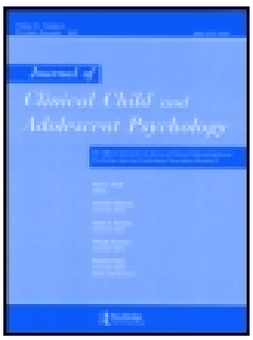

Routledge
Taylor & Francis Group

Journal of Clinical Child & Adolescent Psychology

ISSN: 1537-4416 (Print) 1537-4424 (Online) Journal homepage: https://www.tandfonline.com/loi/hcap20

Testing Robustness of Child STEPs Effects with Children and Adolescents: A Randomized Controlled Effectiveness Trial

John R. Weisz, Sarah Kate Bearman, Ana M. Ugueto, Jenny A. Herren, Spencer C. Evans, Daniel M. Cheron, Alisha R. Alleyne, Adam S. Weissman, J. Lindsey Tweed, Amie A. Pollack, David A. Langer, Michael A. Southam-Gerow, Karen C. Wells & Amanda Jensen-Doss

To cite this article: John R. Weisz, Sarah Kate Bearman, Ana M. Ugueto, Jenny A. Herren, Spencer C. Evans, Daniel M. Cheron, Alisha R. Alleyne, Adam S. Weissman, J. Lindsey Tweed, Amie A. Pollack, David A. Langer, Michael A. Southam-Gerow, Karen C. Wells & Amanda Jensen-Doss (2019): Testing Robustness of Child STEPs Effects with Children and Adolescents: A Randomized Controlled Effectiveness Trial, Journal of Clinical Child & Adolescent Psychology, DOI: 10.1080/15374416.2019.1655757

To link to this article: https://doi.org/10.1080/15374416.2019.1655757

Journal of Clinical Child & Adolescent Psychology, 00(00), 1–14, 2019

ISSN: 1537-4416 print/1537-4424 online
DOI: https://doi.org/10.1080/15374416.2019.1655757

Testing Robustness of Child STEPs Effects with Children and Adolescents: A Randomized Controlled Effectiveness Trial

John R. Weisz
Department of Psychology, Harvard University

Sarah Kate Bearman
Department of Educational Psychology, The University of Texas at Austin

Ana M. Ugueto
Department of Psychiatry and Behavioral Sciences, McGovern Medical School, The University of Texas Health Science Center at Houston

Jenny A. Herren
Department of Psychiatry and Human Behavior, Brown University

Spencer C. Evans
Department of Psychology, Harvard University

Daniel M. Cheron
Judge Baker Children's Center, Harvard Medical School

Alisha R. Alleyne
Instagram

Adam S. Weissman
The Child & Family Institute and Columbia University Teacher's College

J. Lindsey Tweed
Edmund N. Ervin Pediatric Center, Maine General Hospital

Amie A. Pollack
Vinacapital Foundation

David A. Langer
Department of Psychology, Suffolk University

Correspondence should be addressed to John R. Weisz, Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, MA 02183. E-mail: john_weisz@harvard.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hcap.

Michael A. Southam-Gerow
Department of Psychology, Virginia Commonwealth University

Karen C. Wells
Department of Psychiatry and Behavioral Sciences, Duke University Medical Center

Amanda Jensen-Doss
Department of Psychology, University of Miami

A critical task in psychotherapy research is identifying the conditions within which treatment benefits can be replicated and outside of which those benefits are reduced. We tested the robustness of beneficial effects found in two previous trials of the modular Child STEPs treatment program for youth anxiety, depression, trauma, and conduct problems. We conducted a randomized trial, with two significant methodological changes from previous trials: (a) shifting from cluster- to person-level randomization, and (b) shifting from individual to more clinically feasible group-based consultation with STEPs therapists. Fifty community clinicians from multiple outpatient clinics were randomly assigned to receive training and consultation in STEPs (n = 25) or to provide usual care (UC; n = 25). There were 156 referred youths—ages 6–16 (M = 10.52, SD = 2.53); 48.1% male; 79.5% Caucasian, 12.8% multiracial, 4.5% Black, 1.9% Latino, 1.3% Other—who were randomized to STEPs (n = 77) or UC (n = 79). Following previous STEPs trials, outcome measures included parent- and youth-reported internalizing, externalizing, total, and idiographic top problems, with repeated measures collected weekly during treatment and longer term over 2 years. Participants in both groups showed statistically significant improvement on all measures, leading to clinically meaningful problem reductions. However, in contrast to previous trials, STEPs was not superior to UC on *any* measure. As with virtually all treatments, the benefits of STEPs may depend on the conditions—for example, of study design and implementation support—in which it is tested. Identifying those conditions may help guide appropriate use of STEPs, and other treatments, in the future.

Virtually all mental health interventions that are found to be effective are effective only within a particular range of conditions and circumstances. Thus, a critical element of intervention science is identifying the conditions within which benefits of a particular treatment can be replicated and outside of which those benefits are reduced. Reproducability of findings is central to the scientific method more broadly (Staddon, 2017), and the need for tests of replication in psychology has been highlighted repeatedly in recent years (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012; Simons, 2014). Within clinical psychological science, in particular, tests of whether treatment findings are robust across trials and conditions are important because they have real-world clinical implications for those who seek help and those who provide it (Hengartner, 2018). A relevant concept, adapted from the field of economics, is the *winner's curse* (see, e.g., McMahon, 2014)—a tendency for highly positive initial findings to diminish in subsequent evaluations. This phenomenon may reflect a range of possible causes, including, for example, publication bias favoring initially exciting findings, the introduction of more rigorous study designs in later tests, cross-study differences in the study samples employed, or changes in methods of treatment implementation. At a minimum, tests that probe the impact of such variations can help identify the circumstances within which treatments are beneficial and beyond which effects may shrink.

This perspective may be especially relevant to new treatment approaches that disrupt traditional methods and for which early evidence appears promising. One example in research on psychotherapy with children and adolescents (herein "youths") is a multidiagnostic treatment approach represented by the Child STEPs model (Schoenwald, Kelleher, Weisz, & Research Network on Youth Mental Health, 2008). The STEPs model includes two components: (a) a modular, transdiagnostic treatment protocol and (b) a clinical feedback system that keeps clinicians informed of each youth's treatment response to help guide clinical decision-making throughout episodes of care. The treatment program is the Modular Approach to Therapy for Children with Anxiety, Depression, Trauma, or Conduct Problems (MATCH; Chorpita & Weisz, 2009). MATCH was developed in part to improve the synchrony between structured psychotherapies and characteristics of clinically referred youths (Weisz, Krumholz, Santucci, Thomassin, &

Ng, 2015). Most structured, manual-guided youth psychotherapies are focal and linear—that is, they focus on one type of disorder or problem (e.g., depression), and they entail a prescribed sequence of sessions with content presented in relatively fixed order. In contrast, youths referred for treatment often present with multiple disorders or problems, and their most pressing problems may fluctuate from session to session. To address these characteristics of referred youths, MATCH spans four broad problem domains (anxiety, depression, posttraumatic stress, and conduct problems), offering therapists a menu of 33 "modules"—that is, empirically supported treatment procedures for the four problem areas (Chorpita & Weisz, 2009). Flowcharts guide therapists in selecting and sequencing the modules to create a personalized treatment for each youth (Ng & Weisz, 2016), with those judgments further informed by a monitoring and feedback system (MFS), which gives clinicians frequent feedback on youth treatment response throughout episodes of care. This combination of multidiagnostic treatment (MATCH) plus MFS forms the STEPs model.

Two initial randomized controlled effectiveness trials (RCETs) of STEPs, one with a follow-up study examining longer term outcomes, have provided a positive picture to date. The first trial (Weisz et al., 2012), with youths ages 7–13, took place in 10 clinical service sites in two states. The sites included a mixture of free-standing and government-operated mental health service programs in clinics and schools, all serving the public, with an emphasis on relatively low-income families whose services were funded by Medicaid. In this trial, STEPs (using the original version of MATCH that addressed anxiety, depression, and conduct problems but not trauma) showed clinical outcomes on multiple measures that were superior to usual clinical care and superior to standard treatment protocols for anxiety, depression, and conduct problems. A subsequent report on that study (Chorpita et al., 2013) showed superiority of STEPs over usual care (UC) on symptom measures but not on functional impairment or use of auxiliary services, based on 2-year follow-up assessments. A more recent RCET (Chorpita et al., 2017) tested STEPs versus a form of UC that included county support for training and implementation of evidence-based practices, and incentives to use those practices, in three clinical service sites, with youngsters ages 5–15. Three community agencies were included, each providing services to children and adolescents, predominantly from low-income families, with support from county government-administered contracts. This study found that STEPs (using the current version of MATCH, designed for anxiety, depression, conduct problems, and trauma) produced outcomes on measures of symptoms, severity of youth- and caregiver-identified "top problems," and use of auxiliary services, that were superior to those of the comparison condition.

These initial findings were encouraging. However, as previously noted, early positive findings in intervention science warrant tests to assess the robustness of treatment benefit and to explore whether the benefit is limited by certain boundary conditions. The present study focused on two characteristics of the STEPs evidence base, to date, that warrant attention in assessing robustness: experimental design and implementation procedures.

Experimental Design

Both published trials of STEPs have used cluster randomization, in which groups of clinically referred youths (as opposed to individual youths) were assigned to conditions, with randomization at the cluster level. Cluster randomized designs may be used for a number of appropriate reasons (e.g., insufficient numbers to permit within-site randomization, or efforts to prevent cross-condition contamination). However, cluster randomization also raises significant methodological challenges (see, e.g., Campbell, Piaggio, Elbourne, & Altman, 2012; Donner & Klar, 2004; Esserman, Allore, & Travison, 2016); one challenge is that randomizing clusters (e.g., clinic sites or therapist clusters) rather than individual participants poses the risk that participants in the groups being compared may differ in unknown ways that are outcome relevant, impacting the validity of findings. As an example, in the initial STEPs trial (Weisz et al., 2012), some of the participating school-based mental health sites had only one clinician; in each such case, the clinician and thus the entire school were randomized to either STEPs or UC. Schools can differ markedly in their student populations, characteristics of the school and its personnel, which clinicians wish to work in them (as opposed to other schools), and even in the time and resources clinicians are given to treat students (vs. being assigned other school duties), and these variables may combine to foster good or poor treatment outcomes. The potential risks of cluster randomization suggest a need to test STEPs within an RCET design that involves random assignment at the individual participant level. In the present study we addressed this issue by arranging for individual random assignment of each youth, and each therapist, as well.

Implementation Procedures

Both published STEPs trials used implementation procedures that might be difficult to replicate under everyday clinical care conditions: In both trials, each MATCH therapist received weekly individual consultation from a MATCH expert in the study team, throughout the duration of the treatment phase of the study. In the present study,

MATCH expert consultants met with clinicians in groups (*M* group size = 3.82), a format that appears to be more representative of what clinics might do under normal conditions absent research funding (in part because consultant costs are not reimbursable via third-party payment). In addition to being more clinically representative, the group format might potentially generate peer interactions that could enhance clinician engagement in skill-building during consultation sessions, and the shared experience with peers might boost opportunities for collegial support and trouble-shooting during workdays between meetings with the expert consultant. On the other hand, given the relative complexity of the STEPs program (33 modules, four problem areas, multiple flowcharts, ongoing monitoring via an MFS, weekly clinical decision-making to personalize treatment for each youth), it is possible that clinicians might actually need individual consultation to achieve the outcomes seen in previous trials. Thus, our use of group consultation in the present study provided a look at whether the effectiveness of STEPs as seen in prior trials would be sustained (or improved or reduced) with a more practice-relevant form of clinician support than the individual consultation used previously.

Study Overview

In the present study, these experimental design and implementation features were included within an RCET based in three large outpatient community mental health clinics. Aside from the change in randomization procedure and the shift to group consultation, we sought to follow study procedures that corresponded to those of prior STEPs trials (see Method section). This effort was facilitated by overlap in study personnel: the study principal investigator (who is also a coauthor of the MATCH manual); the Project Director; three MATCH trainers, who led clinician trainings; two MATCH expert consultants, who worked with clinicians in the STEPS condition; two experienced experts on internalizing and externalizing, respectively, who provided weekly consultation to the consultants (see the upcoming text); and one study data analyst had all played the same roles in the initial STEPs trial. The study examined clinical outcomes for youths individually randomized to STEPs or UC, treated by clinicians who were also individually randomized to STEPs or to UC; UC clinicians agreed to use their current skills and knowledge to provide the treatment they thought best for each youth. We measured trajectories of change in multiple outcome measures encompassing mental health symptoms and the "top problems" identified as most important by youths and their caregivers. We conducted exploratory analyses of four candidate moderators because of their potential clinical relevance, although there was insufficient prior evidence to justify hypotheses. The candidate moderators included (a) problem severity at baseline, because identifying moderation by severity might guide future treatment assignment decisions in relation to STEPs; (b) medication use, to shed light on whether medication might usefully complement either STEPs or UC; (c) child welfare involvement, because STEPs includes intervention with caregivers, and involvement of child welfare is sometimes associated with caregiver difficulty or instability; and (d) initial focus of treatment (i.e., internalizing vs. externalizing problems), because a moderation effect might help inform future treatment assignment decisions. For our overall tests of STEPs versus UC, we reasoned that further supportive evidence would increase confidence that the initial STEPs findings are robust across significant variations in study procedures, and alternatively that a failure to replicate could add usefully to the picture of boundary conditions within which STEPs effectiveness is evident.

METHODS

Informed consent and assent were obtained from all caregivers and youths prior to study enrollment. All study procedures were approved by the Institutional Review Board.

Participants, Clinics, and Procedures

The final treatment sample included 156 youths ages 6–16 (M_{age} = 10.52, *SD* = 2.53) at baseline, the developmental range appropriate for MATCH and closely resembling the age range of previous STEPs trials. Some 48.1% were male; 79.5% were Caucasian, 12.8% multiracial, 4.5% Black, 1.9% Latino, and 1.3% identified as Other. Some 31.4% of caregivers reported annual family income of less than $20,000, 34.0% reported $20,000–$39,999, 24.4% reported $40,000–$79,999, and 5.8% reported $80,000 or more (4.5% did not provide that information). The sample included only youths referred to our partner clinics through normal community pathways. When families sought treatment at one of our partner clinics, they were told about the study opportunity by clinic staff, and families who consented were contacted by study staff for more study information, caregiver consent and youth assent, screening, and baseline assessments where appropriate; family members were compensated for their time as approved by the Institutional Review Board. Inclusion criteria were ages 6–15 at time of initial screening and a primary problem or disorder of anxiety, depression, traumatic stress, and/or conduct problems, with confirmation via borderline or clinical range scores on relevant subscales of the Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) or Youth Self-Report (YSR; Achenbach & Rescorla, 2001), or on the UCLA PTSD Reaction Index (Steinberg, Brymer,

Decker, & Pynoos, 2004). Exclusion criteria included caregiver reports of a suicide attempt within the previous year, schizophrenia spectrum disorder, autism or another pervasive developmental disorder, eating disorder, and intellectual disability. Youths with attention deficit/hyperactivity disorder were included if the primary focus was one of the four MATCH problem areas. The two conditions did not differ in age, $t(154) = 1.140$, $p = .256$; gender, $\chi^2(1) = 0.000$, $p = .995$; racial/ethnic minority representation, $\chi^2(1) = 0.007$, $p = .935$; percentage receiving medication while in the study (54% overall), $\chi^2(1) = 0.665$, $p = .415$; or family income bracket, $\chi^2(3) = 02.758$, $p = .431$.

The CONSORT diagram (Figure 1) summarizes participants' progression from referral through data analysis. Of 599 youths referred, 497 were screened, 308 completed baseline assessments, 235 were eligible and agreed to participate, and 179 attended at least one treatment session. Prior to analysis, some participants were removed for two methodological reasons. First, for cases in which two or more participants had the same caregiver, one youth was randomly excluded ($n = 21$) to eliminate dependencies in the data. Second, for cases in which the caregiver informant changed during treatment ($n = 18$), caregiver-reported data were excluded from analysis because of compromised reliability, but youth-report data were included when available ($n = 16$). This left a remaining total sample of 156 for analysis (77 STEPs, 79 UC), somewhat larger than in the initial STEPs trial (sample size: 62 STEPs, 53 UC; Weisz et al., 2012) and the second trial (78 STEPs, 60 UC; Chorpita et al., 2017).

The youth sample spanned a range of problems, including anxiety, depression, posttraumatic stress, and/or conduct problems, often with more than one of these problems co-occurring. Caregivers reported high levels and rates of CBCL Internalizing Problems ($M = 66.2$, $SD = 7.7$, with 83.5% being "elevated"; T scores ≥ 60; Achenbach & Rescorla, 2001) Externalizing Problems ($M = 67.2$, $SD = 8.3$, 82.7% elevated), and Total Problems ($M = 68.4$,

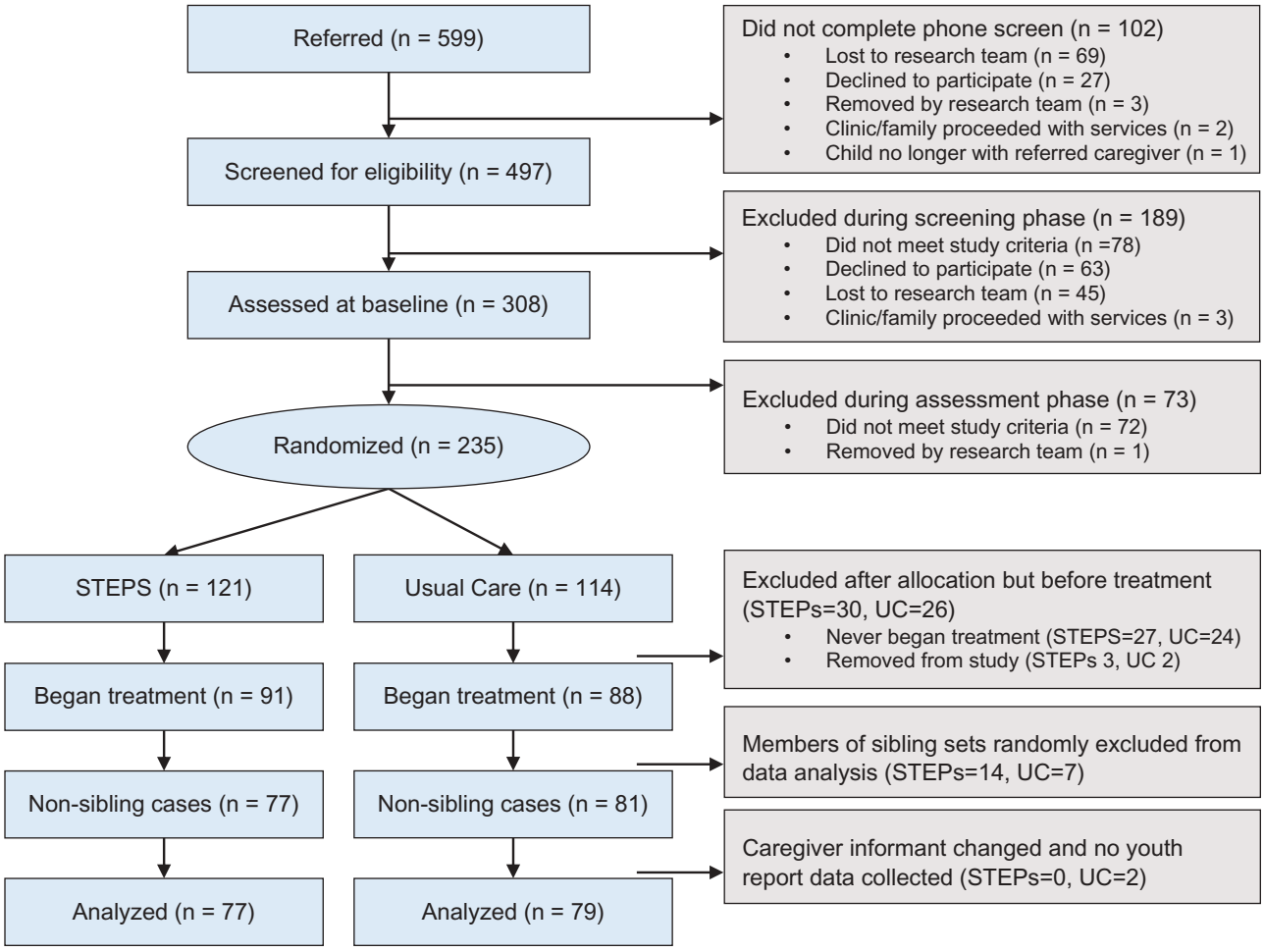

FIGURE 1 CONSORT diagram showing participant flow from referral to data analysis. *Note*: Prior to data analysis, cases were excluded because of sibling status (i.e., multiple youths with same caregiver informant), to eliminate dependencies in the data, and because of caregiver informant change, to reduce unreliability.

SD = 6.5, 89.2% elevated). Baseline YSR results showed lower but nontrivial levels and rates of YSR Internalizing (*M* = 53.6, *SD* = 10.9, 30.8% elevated), Externalizing (*M* = 52.4, *SD* = 9.8, 26.9% elevated) and Total Problems (*M* = 55.0, *SD* = 10.5, 36.9% elevated). Of note, 69.1% of these youths had co-occurring elevations in both internalizing and externalizing problems by parent report (16.2% by youth report). The STEPs and UC groups did not differ on any of these CBCL/YSR scale scores (*ps* = .245–.653) or problem frequencies (*p*s = .180–.799) at baseline.

Therapists and Clinics

Treatment was provided by 50 therapists—25 randomly allocated to STEPs, 25 to UC—employed by three large outpatient community mental health clinics, with branch offices encompassing urban, suburban, and rural areas. The clinics were nonprofits the services of which were reimbursed by third-party payers, primarily Medicaid. They provided care to youths referred from multiple sources, including families self-referring from the community, the state Department of Health and Human Services, Alternative Response Programs (responsible for reducing child maltreatment), and postadoption support programs. Seeking a representative sample, we included participants from all of these sources that met study criteria; the STEPs and UC conditions did not differ in referral source, χ^2 (1) = 0.001, p = .975. We sought data from therapists on their background and current practice, and 43 of the 50 responded. Therapists averaged 43.36 years of age (*SD* = 11.24); 33 were female; and 40 were White, 1 Hispanic, and 2 Other. All were licensed, and mean years of full-time clinical practice since professional training was 7.27 (*SD* = 6.83); 37 self-identified as social workers, 5 as counselors, and 1 as a psychologist; 41 reported having a master's degree (mainly MSW/LCSW), one a Ph.D., and one an Ed.D. Therapists reported a mean of 25.73 hr (*SD* = 23.27) in the previous 2 years attending workshops or other training programs for child/adolescent therapy. They reported a mean of 30.76 (*SD* = 10.02) active cases at a time and 1.38 hr (*SD* = 0.63) of supervision per week. The three participating clinics arranged training opportunities in empirically supported treatments for their clinicians, with content reportedly including trauma-focused cognitive-behavioral therapy (TF-CBT) and multisystemic therapy, and study clinicians reported a mean of 35.73 hr of training during the 2 years prior to the study. When we asked clinicians to identify their primary theoretical orientation, 13 reported CBT, 11 mixed/eclectic including CBT, 8 other mixed/eclectic, and 11 various singular orientations that excluded CBT (e.g., family systems). STEPs and UC therapists did not differ on any of the measured demographic, professional, or clinical service characteristics (*p*s > .20).

Experimental Design, Random Assignment, and Assessment Schedule

The RCET involved double randomization. Half the clinicians were randomized to STEPs and half to UC. When youths were referred for treatment, families who approved were contacted by project staff for initial screening; a baseline assessment followed, with eligible youths randomly assigned to STEPs or UC. The assessments included weekly measures of youths' response to treatment, displayed in the MFS for use by therapists and consultants during weekly case consultation (described next).

Measures

Trained assessors, kept naïve to treatment condition, administered measures to youths and caregivers following two longitudinal measurement schedules: (a) comprehensive standardized measures administered at baseline and 3, 6, 9, 12, 18, and 24 months thereafter, as well as posttreatment, and (b) brief rating scales administered weekly during treatment. Outcomes were examined longitudinally (i.e., using weekly and quarterly measures), as in the previous STEPs trials (Chorpita et al., 2017; Weisz et al., 2012).

CBCL and YSR

The CBCL is a parent-report measure with 113 youth problem items, each rated on a 0–1-2 scale (2 = very/often true). The YSR is a 112-item youth-report measure corresponding to the CBCL. From both the CBCL and the YSR, *T* scores, adjusted for age and gender, for the Internalizing, Externalizing, and Total Problems scales, were used as outcome measures. Evidence for CBCL/YSR validity and reliability is strong and extensive (Achenbach & Rescorla, 2001). The CBCL and YSR showed significant superiority of STEPs over UC in the first STEPs trial (Weisz et al., 2012).

Brief Problem Checklist

The 12-item Brief Problem Checklist (BPC; Chorpita et al., 2010) is a measure of internalizing (six items), externalizing (six items), and total problems, developed via application of item response theory and factor analysis to large CBCL and YSR data sets. In a psychometric study with 184 clinic-referred 8- to 13-year-olds (Chorpita et al., 2010), the 12-item total score showed alphas of .82 for parents and .76 for youths, and correlations between scores on corresponding BPC and CBCL/YSR scales were substantial (all > .57). The BPC was used for weekly tracking of problem trajectories. The BPC showed significant superiority of STEPs over UC in both previous trials (Chorpita et al., 2017; Weisz et al., 2012).

Top Problems Assessment

The Top Problems Assessment (TPA; Weisz et al., 2011) assesses youth and caregiver severity ratings for the top three problems the youth and caregiver independently identified as most important to them, in separate structured baseline interviews. Psychometric analyses with a sample of 178 clinic-referred youths supported the test–retest reliability (.69 to .91, across 5- to 21-day intervals), convergent and discriminant validity (in relation to standardized measures), and sensitivity to clinical change during treatment. The TPA showed significant superiority of STEPs over UC in both previous trials (Chorpita et al., 2017; Weisz et al., 2012). The top problems provided clinically useful information, but the problems were not required to match anxiety, depression, trauma, or conduct for youths to be included in the study (see inclusion criteria under the Participants, Clinics, and Procedures section).

Therapist Integrity in Evidence-based Interventions

As in both previous trials (Chorpita et al., 2017; Weisz et al., 2012), recordings of both MATCH and UC sessions were coded for presence/absence of evidence-based treatment procedures corresponding to MATCH. Sessions were coded in 5-min segments for presence/absence of 27 items reflecting that content. Competence can also be rated based on coders' global evaluation of how skillfully the therapist delivered any present item corresponding to MATCH content in a given session: range = 1 (*insufficient or superficial*) to 4 (*expert*). Coders were 10 bachelor's- and master's-level research assistants supervised by the primary Therapist Integrity in Evidence-based Interventions (Bearman, Herren, & Weisz, 2012; Bearman, Schneiderman, & Zoloth, 2017) developer. For MATCH content adherence, mean intercoder agreement (via independent coding of the same 54 randomly selected sessions) was intraclass correlation coefficient (ICC) [1, 1] = 0.87. For MATCH content competence, mean reliability across pairs was intraclass correlation coefficient (ICC) [1, 1] = 0.91. For this study, 390 sessions were sampled from a pool of 1,411 total audible recordings, with the following constraints: Sessions were randomly selected from the first, middle, and last thirds of the full treatment episode, after excluding initial sessions (these often had administrative content) and sessions of unrepresentative length (< 15 min or > 75 min). All coders were kept naïve to participant identity, characteristics, and study condition. Mean adherence percentage and competence rating were calculated for each youth's full treatment episode.

Treatment Conditions

STEPs Condition

Clinicians randomized to STEPs received training and group consultation in the use of MATCH and the MFS. Training replicated the original MATCH training (Weisz et al., 2012), totaling 6 days, broken into three 2-day units (each separated by 2–4 weeks): one emphasizing depression, one anxiety and traumatic stress, and one conduct problems. Trainings included didactic presentations, case examples, trainer modeling, and therapist role-plays with feedback. After training, therapists treated study participants, guided by consultation from MATCH experts on the study team. The weekly consultations included review of cases informed by the MFS dashboard (which documented the modules used and any changes in the weekly youth and caregiver ratings), identification of treatment challenges, suggestions from the consultant regarding application of MATCH skills, group problem-solving, and plans for upcoming sessions. When non-MATCH issues (e.g., impulsive behavior, irritability, stressful life events) surfaced in sessions, deviations from MATCH were possible, but consultation provided an opportunity to clarify relevance to MATCH problems (e.g., stressful events might trigger anxiety or depression symptoms that could be addressed by MATCH) and to MATCH strategies (e.g., a new life stressor may be addressed via the MATCH problem solving module). Mean number of clinicians per group was 3.82 (SD = 1.47). Consultants had their own weekly consultations, 1 hr per week, with one experienced expert in treatment of youth internalizing problems and another experienced expert in treatment of youth externalizing problems. These consultations lasted throughout the treatment phase of the study.

UC Condition

Therapists randomized to UC agreed to use their current skills and knowledge to provide the treatment they thought best for each youth they treated, as they would do ordinarily in the absence of a study. Supervision of their work followed the usual practices in their clinic, and therapy for each case continued until a normal end of treatment.

Duration of Treatment in the Two Conditions

Overall, treatment duration was 190.5 (SD = 134.1) days, 16.9 (SD = 14.0) sessions. There were no group differences in days (M_{STEPS} = 175.6, M_{UC} = 204.7), $t(153) = 1.36$, $p = .177$, or number of sessions (M_{STEPS} = 17.7, M_{UC} = 16.1), $t(153) = 0.69$, $p = .361$.

Session Content: Adherence and Competence Coding for the Two Conditions

The 390 randomly selected session recordings included 209 for STEPs and 181 for UC. MATCH sessions included more of the evidence-based content found in MATCH (M = 67.0%) than UC sessions did (M = 26.6%), $F(1, 388) = 3.67$, $p < .001$, and mean therapist competence for

that content in the sessions was also significantly higher in the MATCH condition (*M* session competence = 2.15) than in UC (*M* session competence = 1.15), $F(1, 388) = 60.47$, $p < .001$.

Treatment Attendance and Termination in the Two Conditions

Client attendance was similar in the two conditions, with no significant condition differences in the number of cancellations ($p = .88$) or client "no shows" ($p = .58$). Assessing treatment completion versus "dropout" is complicated for a treatment protocol that has no prescribed number of sessions, as was the case with MATCH and UC. However, data were available on whether cases had a "routine termination," were "lost to therapist" (i.e., the therapist lost contact with the family, and the case was closed), or fell into "other" or "unknown" categories. "Routine termination" was more common in STEPs than UC cases (60.4% vs. 30.7%), $\chi^2(1) = 14.78$, $p < .001$, and cases "lost to therapist" were marginally more common in UC than STEPs (26.1% vs. 15.4%), $\chi^2(1) = 2.78$, $p = .095$.

Planned Analyses

Following the analytic strategy used in prior STEPs trials, we examined trajectories of change in (a) brief parent- and youth-reported measures of internalizing and externalizing (Chorpita et al., 2010) and "top" problems (Weisz et al., 2011), administered weekly throughout treatment (Weisz et al., 2012), and (b) full-length measures of parent- and youth-reported internalizing and externalizing problems (Achenbach & Rescorla, 2001), administered quarterly through 2 years postbaseline (see Chorpita et al., 2013). This trajectory analysis approach is used in RCET comparisons in which neither condition has a fixed duration or number of sessions, and in which findings of endpoint/posttreatment comparisons would thus be difficult to interpret (see Weisz et al., 2012). Analyses compared youths in STEPs versus UC in their trajectories of change on weekly (BPC, TPA) and quarterly (CBCL/YSR) measures per parent and youth report. Mixed models were specified to include fixed effects for intercept, condition, time, condition × time, and site (dummy coded), with intercept and time treated as random effects. Consistent with prior trials (Chorpita et al., 2013; Weisz et al., 2012), time was modeled as the natural logarithm of days+1 since baseline. Of primary interest was the condition × time interaction term, which tests whether one group exhibited faster improvement than the other group. Magnitude was measured in a standardized effect size (ES) consistent with Cohen's *d* benchmarks of 0.2, 0.5, and 0.8 for small, medium, and large effects. Again following previous trials, ES was computed as the condition × time estimate divided by the square root of the overall time slope variance. To provide a clinically meaningful picture, we also report model-implied 1-year and 2-year score change estimates, and whether these outcomes could be considered clinically reliable change (Jacobson & Truax, 1991). Reliable change thresholds were calculated as $s\sqrt{(1-r)}*1.96$, using standard deviation (*s*) from this sample and reliability (*r*) from psychometric publications for the CBCL/YSR (Achenbach & Rescorla, 2001), BPC (Chorpita et al., 2010), and Top Problems (Weisz et al., 2011).

To adjust for multiple comparisons, planned family-wise Bonferroni corrections were employed to reduce the risk of chance findings. Given the three distinct outcomes of interest collected from each informant in the weekly assessments (TPA, BPC Internalizing, BPC Externalizing [BPC Total excluded due to redundancy]) and quarterly assessments (CBCL/YSR Internalizing, Externalizing, Total), the significance threshold for our primary results was set at 0.05/3 = 0.0167 for our main results. Because exploratory moderator analyses increase the likelihood of Type I errors multiplicatively, we extended this same logic (four different moderator variables applied within each set of three outcomes) to four sets of post hoc exploratory analyses, for a significance threshold of 0.0167/4 = 0.0042. All models used full maximum likelihood estimation and robust standard errors, which can handle moderate departures from normality and patterns of data missing at random.

Availability of specific measures did not differ by condition for any parent-report outcomes, but the UC condition had a greater proportion of cases with data on the YSR (90% vs. 78%), $\chi^2(1) = 4.139$, $p = .042$, and the BPC/TPA (89% vs. 77%), $\chi^2(1) = 3.913$, $p = .048$. In the quarterly assessment schedule, participants completed a mean of 6.2 (*SD* = 2.2; 78%) parent- and youth-report assessments out of the eight possible occasions (including pre- and post-treatment), with no differences between groups (*p*s > .32). Similarly, participants completed a mean of 73% (*SD* = 18%) of their expected weekly progress-monitoring assessments during treatment per parent report, and *M* = 70% (*SD* = 19%) per youth report, again with no differences between groups (*p*s > .09). Both the quarterly and weekly response rates followed a similar pattern of gradual linear decrease in response rates over time, ranging from 99% at baseline to 64% at the 24-month quarterly, for both parent- and youth-report.

We used Optimal Design Version 3.01 to estimate power for multilevel models to detect main effects of treatment condition on slopes of change over time (measured via six or more repeated observations). Based on prior research, there was adequate power a priori ($1–\beta = 0.8$) to detect medium effects (ES = 0.5) at standard thresholds ($\alpha = 0.05$). Post hoc power analyses, using the observed median cluster dependencies (ICC = 0.02) among therapists (*N* = 50), confirmed that our sample (*N* = 156) offered

power (1–β = 0.82) to detect the modal ES found in prior research (ES = 0.5; the majority of previous effects on comparable measures in STEPs trials fell between 0.45 and 0.59; see Table 3). Sufficient power (1–β > 0.8) was maintained for informant-specific outcomes across varying levels of effective sample size, as noted ealier.

TABLE 1
Coefficient Estimates for Condition (STEPS vs. UC) by Time (Log-Day) on Weekly and Long-Term Outcomes by Youth and Caregiver Report

	Youth Report			*Caregiver Report*		
Outcome	*Estimate*	p	*ES*	*Estimate*	p	*ES*
Weekly Measures						
BPC Total	−0.084	.518	0.12	0.132	.355	0.20
BPC Internalizing	−0.082	.297	0.20	0.150	.049	0.40
BPC Externalizing	−0.002	.984	0.00	−0.018	.857	0.04
Top Problems	−0.0004	.998	0.00	0.093	.207	0.24
Quarterly Measures						
CBCL/YSR Internalizing	−0.137	.582	0.15	0.217	.324	0.27
CBCL/YSR Externalizing	0.021	.934	0.04	0.093	.634	0.14
CBCL/YSR Total	−0.067	.783	0.10	0.162	.421	0.22

Note: Negative estimates indicate that youths in the STEPS condition evidenced a faster rate of reduction in symptoms over time than those in the usual care (UC) condition. Following Weisz et al. (2012), effect size was calculated as the ratio of the difference in rates of change between the two conditions divided by the square root of the overall time slope variance; it expresses the absolute value of the standardized magnitude of the effect. When a Bonferroni correction was applied, none of the condition × time estimates were significant. BPC = Brief Problem Checklist; CBCL = Child Behavior Checklist; YSR = Youth Self-Report.

RESULTS

Results comparing trajectories of change for STEPs versus UC are presented in Table 1, with slopes and model-implied score changes presented in Table 2. Across nearly all outcome measures, between-group comparisons were consistently nonsignificant with negligible to small effect sizes ($ps > .2$, ESs < 0.3). The singular exception was that UC produced faster improvement on parent-reported BPC Internalizing scores compared to STEPs ($p = .049$, ES = 0.40); however, this result did not survive correction for multiple comparisons. Figure 2 illustrates the overall pattern of results across all measures, plotting the outcomes for BPC Total, Top Problem severity, and CBCL/YSR Total scores per both informants. As shown, trajectories of change were extremely similar across conditions.

Despite the absence of between-group differences, there was a general pattern of within-group improvement on nearly all outcomes (YSR Externalizing was the sole exception). As shown in Table 2, these changes were both statistically significant and clinically meaningful. For

TABLE 2
Slopes and 1-Year and 2-Year Change Estimates by Condition

	STEPs				*Usual Care*			
Outcome	*Slope*	p	*1-Year Change*	*2-Year Change*	*Slope*	p	*1-Year Change*	*2-Year Change*
Youth Report								
Weekly Measures								
BPC Internalizing	−0.240	< .001	−1.42	—	−0.158	< .001	−0.93	—
BPC Externalizing	−0.241	< .001	−1.42	—	−0.240	< .001	−1.41	—
BPC Total	−0.480	< .001	−2.83	—	−0.396	< .001	−2.34	—
Top Problems	−0.644	< .001	−3.82[rc]	—	−0.643	< .001	−3.80[rc]	—
Quarterly Measures								
YSR Internalizing	−1.278	< .001	−7.54[rc]	−8.43[rc]	−1.141	< .001	−6.73	−7.52[rc]
YSR Externalizing	−0.723	< .001	−4.27	−4.77	−0.744	< .001	−4.39	−4.51
YSR Total	−1.156	< .001	−6.82[rc]	−7.62[rc]	−1.090	< .001	−6.43[rc]	−7.18[rc]
Caregiver Report								
Weekly Measures								
BPC Internalizing	−0.303	< .001	−1.79	—	−0.453	< .001	−2.67[rc]	—
BPC Externalizing	−0.449	< .001	−2.65	—	−0.431	< .001	−2.54	—
BPC Total	−0.753	< .001	−4.44[rc]	—	−0.885	< .001	−5.22[rc]	—
Top Problems	−0.546	< .001	−3.22[rc]	—	−0.639	< .001	−3.77[rc]	
Quarterly Measures								
CBCL Internalizing	−1.310	< .001	−7.73[rc]	−8.63[rc]	−1.527	< .001	−9.01[rc]	−10.07[rc]
CBCL Externalizing	−1.085	< .001	−6.40[rc]	−7.15[rc]	−1.178	< .001	−6.95[rc]	−7.77[rc]
CBCL Total	−1.158	< .001	−6.83[rc]	−7.64[rc]	−1.320	< .001	−7.79[rc]	−8.70[rc]

Note: Slopes reflect the estimate of change in scale score per log-day, and the 1- and 2-year figures estimate the scale score change following that amount of time since the baseline assessment. BPC = Brief Problem Checklist; YSR = Youth Self-Report; CBCL = Child Behavior Checklist.

[rc]Surpasses reliable change index.

TABLE 3
Benchmarking Comparison of the Present Study with Previous STEPs Trials

	Present Study		*Weisz et al. (2012)*[a]		*Chorpita et al. (2017)*	
Sample Characteristics	$N = 156$		$N = 174$ (STEPs vs. UC, $N = 115$)		$N = 138$	
Age, *M* (*SD*), Range	10.5 (2.5), 6–16		10.6 (1.8), 7–13		9.3 (2.8), 5–17	
Male, %	48		70		55	
Race/Ethnicity						
White/Caucasian, %	80		45		4	
Latino/a, %	2		6		78	
African American, %	5		9		10	
Multiethnic, %	13		32		8	
Received Any Psychotropic Medication, %	54		27		17	
Family Income < $40K, %	65		55		91	
Therapist Experience/Cases	$N = 50$		$N = 84$ (STEPs vs. UC, $N = 55$)		$N = 50$	
Years of Experience, *M*	7.3		7.6		3.3	
No. of Study Cases, *M*	3.1		2.1		2.8	
Baseline Characteristics[b]	STEPs	UC	STEPs	UC	STEPs	CIT
Caregiver Internalizing, %	87	80	81	81	73	65
Caregiver Externalizing, %	86	80	79	77	80	75
Youth Internalizing, %	32	30	43	49	26	24
Youth Externalizing, %	24	30	21	26	32	42
Treatment Characteristics	STEPs	UC	STEPs	UC	STEPs	CIT
MATCH Training, days	6	—	6	—	5	—
MATCH/EBT Content, %	67	27	83	8	78	7
No. of Sessions, *M*	17.7	16.1	16.2	—	21.7	30.2
Treatment Duration, *M* Days	175.6	204.7	210.2	275.5	191.8	267.0

	Comparison		Slopes		Comparison		Slopes		Comparison	
Primary Outcomes[c]	Est	ES	STEPs	UC	Est	ES	STEPs	UC	Est	ES
Caregiver Total	0.13	0.20	−0.75	−0.64	−0.44*	0.54	−0.94	−0.50	−0.007**[,d]	0.51[d]
Caregiver Internalizing	0.15	0.40	−0.30	−0.45	−0.21[†]	0.38	−0.50	−0.29	−0.004**[,d]	0.46[d]
Caregiver Externalizing	−0.02	0.04	−0.45	−0.43	−0.23*	0.50	−0.45	−0.21	−0.003*[,d]	0.38
Caregiver TPA	0.09	0.24	−0.55	−0.64	−0.33**	0.72	−0.65	−0.32	−0.005**[,d]	0.56
Youth Total	−0.08	0.12	−0.48	−0.40	−0.24	0.32	−0.81	−0.47	—	—
Youth Internalizing	−0.08	0.20	−0.24	−0.16	−0.15	0.34	−0.39	−0.24	—	—
Youth Externalizing	0.00	0.20	−0.24	−0.24	−0.09	0.24	−0.37	−0.21	—	0.50
Youth TPA	0.00	0.20	−0.64	−0.64	−0.14	0.28	−0.61	−0.47	—	0.38

	Comparison		Slopes		Comparison		Slopes		Comparison	
Long-Term Outcomes[c]	Est	ES	STEPs	UC	Est	ES	STEPs	UC	Est	ES
CBCL Total	0.16	0.22	−1.16	−1.32	−0.70*	0.59	−2.08	−1.38	—	—
CBCL Internalizing	0.22	0.27	−1.31	−1.53	−0.53[†]	0.45	−2.04	−1.52	—	—
CBCL Externalizing	0.09	0.14	−1.09	−1.18	−0.62*	0.55	−1.68	−1.06	—	—
YSR Total	−0.07	0.10	−1.16	−1.09	−0.59[†]	0.45	−2.38	−1.79	—	—
YSR Internalizing	−0.14	0.15	−1.28	−1.14	−0.68*	0.53	−2.69	−2.01	—	—
YSR Externalizing	0.02	0.04	−0.72	−0.74	−0.39	0.29	−1.49	−1.10	—	—

Note. Dashes indicate that the value was not available, not reported, or not applicable. Some estimates (Est) were rounded for clarity of presentation across studies. UC = usual care; CIT = community-implemented treatment, representing UC in a service context where EBT usage was supported and considered mandatory; MATCH = Modular Approach to Therapy for Children with Anxiety, Depression, Trauma, or Conduct Problems; EBT = evidence-based treatment; TPA = Top Problems Assessment; CBCL = Child Behavior Checklist; YSR = Youth Self-Report.

[a]Weisz et al. (2012) results are pooled from the cited study and its long-term follow-up report by Chorpita et al. (2013); note that this study included a third condition (standard manualized treatments) that is not of interest for benchmarking purposes; here we report values specific to MATCH versus UC where indicated.

[b]Percentages represent proportion of the sample falling above published cutoffs—on the Strengths and Difficulties Questionnaire Emotional and Conduct Problems scales in Chorpita et al. (2017) and on the CBCL/YSR Internalizing and Externalizing scales in both the present study and in the Weisz et al. (2012) study.

[c]Estimates represent the effect of Treatment × Time, modeled as days since baseline (log-days for Weisz et al. and the present study). Effect sizes (ESs) were calculated the same across all studies.

[d]Caregiver and child-report estimates were aggregated into composite variables; ESs were reported separately by informant only when they differed.

[†]$p < .10$. *$p < .05$. **$p < .01$.

example, across both informants and conditions, average Top Problem severity dropped by 3–4 points over 1 year, exceeding the thresholds for reliable change. Similarly, BPC Total Problems in both groups dropped by about 4–5 points on average per parent-report and 2–3 points per youth-report, and both conditions saw similar one-year reductions in BPC Internalizing and Externalizing scores —by about 2–3 points per parent-report and 1 point per youth-report. Results for long-term quarterly measures followed a similar pattern, with all parent-report and most youth-report outcomes surpassing thresholds for reliable change. At 2 years postbaseline, CBCL/YSR Total Problem *T* scores had dropped by about 6–8 points in both conditions, whereas Internalizing and Externalizing *T* scores had dropped by about 7–9 and 4–7 points, respectively. Again, Figure 2 summarizes this overall pattern of improvement that was seen in both groups.

Further analyses were conducted to investigate initial severity (CBCL Total Problems for parent-reported outcomes and YSR Total Problems for youth-reported outcomes), primary problem area (50.0% internalizing, 50.0% externalizing), any child welfare involvement (59.6% yes, 40.4% no), or inclusion of medication in treatment (53.8% yes, 46.2% no) as moderators and predictors of treatment response. With the Bonferroni correction applied, two of these variables showed significant main effects, and one produced a significant moderator effect. First, having a primary externalizing problem predicted significantly greater improvement in externalizing problems per parent-report on both the BPC and CBCL ($ps < .005$). Second, higher initial severity predicted overall worse outcomes on CBCL Total ($p < .001$) and Internalizing ($p = .002$) scores. Finally, for the YSR Internalizing score alone, there was a significant moderator effect for involvement in the child welfare system ($p = .002$). Dismantling the effect into its components showed (a) no significant subgroup differences between treatment conditions within the child welfare subsample or within the nonchild welfare subsample, and (b) no significant differences between treatment conditions within the nonchild welfare subsample, but a trend toward STEPs outperforming UC within the child welfare subsample ($p = .005$, ES = −1.68), although this difference did not survive the Bonferroni correction. No effects were found for medication. In sum, the null results just reported were not moderated by the youth's primary presenting problem, involvement with the child welfare system, or medication status; and these variables showed relatively few and inconsistent main effects on overall treatment outcomes.

DISCUSSION

We conducted a randomized controlled trial of Child STEPs in a multiproblem sample of 156 youths ages 6 through 16, majority Caucasian, from low- to middle-income families. The study resembled previous trials comparing STEPS to UC in design, procedures, and study team, but there were two primary differences in study methods: (a) individual randomization of youths instead of cluster randomizing, and (b) group consultation for STEPs therapists instead of individual therapist consultation. We found a marked difference in outcome, relative to the previously favorable findings: We did not find STEPs superior to UC on *any* of the study outcome measures. Table 3 compares features and findings of the three studies, showing a number of similarities across the studies as well as a few apparent differences—for example, the current youth sample was more gender balanced, less ethnically diverse, and more likely to receive medication than samples in the two prior studies. It is possible that cross-study differences in findings could have been influenced by such differences or by differences not measured (e.g., therapist's belief in evidence-based practice). Other possible reasons for the difference in findings may relate to planned

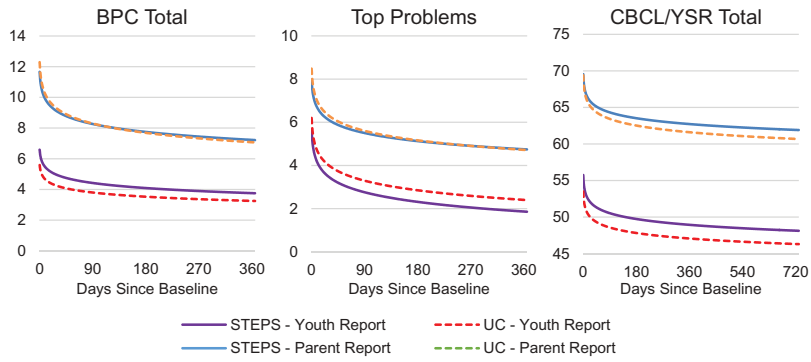

FIGURE 2 Trajectories of change on composite measures by informant and treatment condition. *Note*: Trajectories are based on models reported in Tables 2 and 3. Brief Problem Checklist (BPC) Total are raw sum scores, Top Problems are raw mean scores, and CBCL/YSR are *T* scores. Trajectories are plotted out to 1 year for weekly measures and to 2 years for quarterly measures because of the differences in longitudinal measurement schedules.

differences between the present study and prior studies, to which we now turn.

Our change to individual randomization of youths may have produced a somewhat more rigorous comparison than cluster randomizing might have done. As noted by Donner and Klar (2004) and Campbell et al. (2012), cluster randomized designs raise several methodological concerns, one of which is that randomizing clusters increases the risk that participant groups being compared may differ in unknown ways that are outcome relevant, impacting the validity of findings. As an example, in the initial trial (Weisz et al., 2012), with school-based sites each having one clinician, it is possible that schools with STEPs clinicians and those with UC clinicians differed in their youth populations in outcome relevant ways. Of course, there are often good reasons for using cluster randomized designs, as in the prior STEPs trials, but the methodological concerns suggest the value of using individual randomization where possible, as was the case in the present study.

Our shift to group consultation may also help account for the findings. In the previous STEPs trials, each clinician had individual consultation to discuss cases and make plans with her or his MATCH expert consultant. In the current study, with consultation groups averaging 3.82 clinicians per group, the amount of time available to each clinician, and for discussion of each case, was more limited than in the previous trials—and perhaps insufficient for a full review of the challenges of the cases, or for the kind of fully articulated planning that could enhance outcomes. It is also possible that the group process, with multiple clinicians' opinions aired per case, reduced clarity for some clinicians regarding how best to apply MATCH to their cases, or what next steps to follow in treatment. Although it might appear that peer support would enhance the value of group consultation (as discussed in the introduction), study consultants noted that the limited time and the group social context (including clinician discomfort at being observed by peers) made it difficult for clinicians to carry out the modeling and role-plays that research suggests may be important for effective implementation of evidence-based practices (Bearman et al., 2017, 2013). Of course, the impact of shifting from individual to group consultation can be assessed directly only via a trial specifically contrasting the two approaches, but the present data suggest the testable hypothesis that shifting to less individualized implementation support may weaken fidelity. Consistent with that hypothesis, only 67% of the coded sessions in the present study corresponded to the evidence-based procedures of MATCH; in the initial trial by Weisz et al. (2012) and the second trial by Chorpita et al. (2017), the figures were 78% and 83%, respectively.

To provide a comparative context, we have summarized in Table 3 the features and findings of the present study together with those of prior STEPs trials. The three study samples were clinically similar at baseline, with rates of above-threshold internalizing and externalizing problems falling within a 65%–87% range by parent report and a 21%–49% range by youth report. The most pronounced cross-study difference was in race/ethnicity, with the highest percentage of White youths in the present sample, the sample in Weisz et al. (2012) the most diverse, and the southern California sample of Chorpita et al. (2017) predominantly Latino/a. An interesting pattern shown in the table is that UC youths in the present study showed steeper slopes of improvement on most weekly measures than UC youths in previous research, suggesting that they may have been a more challenging comparison group. The three participating clinics had leaders who had arranged for training opportunities in empirically supported treatments for their clinicians, and our study clinicians reported a mean of 35.73 hr of training during the previous 2 years. This may have elevated the quality of treatment provided by UC therapists, thus reducing prospects for MATCH to outperform UC. This possibility is consistent with our finding that 27% of UC session content corresponded to MATCH's empirically supported content. In Weisz et al. (2012) and Chorpita et al. (2017), the corresponding figures were 8% and 7%, respectively. Comparison to the Chorpita et al. findings is intriguing because clinicians in the control condition of that study had had substantial exposure to evidence-based practices just prior to the study; they had recently received training and consultation in a mean of 2.14 evidence-based therapies, and they were working under contracts that required the use of evidence-based practices for youth mental health service reimbursement. As Chorpita et al. (2017) put it, "All participating agencies and their therapists were subject to considerable external pressure to implement [evidence-based treatments] with their entire caseloads" (p. 18). The contrast between their relatively low levels of evidence-based practice, despite this external pressure, and the substantially higher level by UC clinicians in the present study highlights our need for a deep dive into research on usual clinical care to clarify the experiences and conditions that are actually associated with implementation of empirically supported practices in everyday care.

It is useful to consider these findings together with those of a companion study of STEPs sustainability (Weisz et al., in press). In that study, 14 of the therapists from the present study participated in a subsequent follow-up assessment of whether STEPs fidelity and youth outcomes would be sustained after the therapists' consultation was shifted from external STEPs experts to STEPs-trained internal clinic supervisors. That assessment showed similar levels of fidelity and outcomes for 48 youths those 14 therapists had treated when guided by STEPs expert consultants (i.e., a subset of the present sample) versus 50 youths those

therapists subsequently treated when guided by internal clinic supervisors. These and other findings of that study (including comparisons with another sample of youths and therapists guided by expert STEPs consultants) suggested that STEPs fidelity and clinical outcomes may be sustained when consultation is shifted to STEPs-trained clinic supervisors. The findings of the present study suggest that how fidelity and outcomes that are sustained under those conditions will compare to those of usual clinical care may depend on a number of additional factors, including the nature and effectiveness of the UC that is practiced in the settings where the research is being carried out.

Limitations of our study include the substantial number of youths who were referred and randomized but never appeared for treatment (see Figure 1), a common problem in community mental health centers and one that limits generalizability of findings. Another limitation is the absence of a third study condition—that is, individual consultation—that could have provided a direct test of individual versus group support; that test remains a useful task for future research. An additional limitation was that, although study therapists in the STEPs condition committed to not sharing any MATCH information and appeared to take that commitment quite seriously, we had no way to monitor compliance, so we cannot definitively rule out the possibility that some of evidence-based content found in the UC group reflected cross-condition communication. Finally, although our sample for group comparisons was larger than those of the previous trials, and we had substantial power to detect medium effects, power was diminished for smaller effects. That said, it should be noted that five of our eight main effects were either zero or negative, showing better outcomes for UC than STEPs, and the mean effect size was −.06 showing UC > STEPs.

Taken together, the findings suggest, at a minimum, that the performance of a particular treatment protocol in randomized trials may differ rather markedly from one context to another. Factors that may account for these variations cannot be established with certainty here, in the absence of separate trials testing their impact. However, our study suggests some candidate factors to be tested in future research, including study design variations (e.g., randomization procedures), variations in implementation support (e.g., individual vs. group consultation), and even variations in the nature and quality of the condition to which the candidate treatment is compared (e.g., differences in the nature or quality of UC).

Implications for clinical practice warrant attention as well. If it were true that our findings result from a "watering down" of consultation procedures, one implication might be that for a particularly novel and rather complex treatment—for example, one with four problem area protocols, 33 modules, multiple decision flowcharts, and weekly MFS monitoring needed to guide decision-making—a certain threshold and form of clinician support may be required for successful implementation. Consultation procedures that fall below that threshold, or outside that form, may risk providing insufficient support for individual case review, problem-solving, and treatment planning by clinicians who are just learning the treatment and thus may not provide the scaffolding needed to maximize mastery and effectiveness. That said, it may be unrealistic to set a standard that requires extensive and expensive individual expert consultation for each clinician. Efficiency and scalability are vital for real-world mental health providers, given their financial vulnerability and need for close fiscal management (Schoenwald et al., 2008), so a more realistic goal may be development of low-cost supports for effective implementation by clinicians (e.g., online guidance, virtual consultation, or learning collaborative methods). Indeed, developing and testing low-cost scalable supports for effective implementation seems a worthy goal for a broad array of empirically supported interventions.

ACKNOWLEDGMENTS

We are grateful to the youths, caregivers, and clinicians who participated in this study; to the clinic administrators and state officials whose leadership facilitated and supported the project; and to Shannon Dorsey for her wise guidance.

DISCLOSURE STATEMENT

John Weisz is a coauthor of the treatment manual used in this study (i.e., *Modular Approach to Therapy for Children with Anxiety, Depression, Trauma, or Conduct Problems*) and receives royalties for sales of the manual. The other authors report no conflicts of interest related to this study.

FUNDING

The study was supported by funding to John R. Weisz from the John D. and Catherine T. MacArthur Foundation (Grant 07-90231-000-HCD), Casey Family Programs (Grant 02200), and the Annie E. Casey Foundation (Grant 211.0004).

ORCID

Spencer C. Evans http://orcid.org/0000-0001-8644-817X

REFERENCES

Achenbach, T. M., & Rescorla, L. (2001). *ASEBA school-age forms & profiles*. Burlington, VT: Aseba.

Bearman, S. K., Herren, J., & Weisz, J. R. (2012). *Therapy integrity in evidence based interventions: Observational coding system, coding manual*. Austin: University of Texas at Austin.

Bearman, S. K., Schneiderman, R. L., & Zoloth, E. (2017). Building an evidence base for effective supervision practices: An analogue experiment of supervision to increase EBT fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *44*(2), 293–307. doi:10.1007/s10488-016-0723-8

Bearman, S. K., Weisz, J. R., Chorpita, B. F., Hoagwood, K., Ward, A., Ugueto, A. M., & Bernstein, A. (2013). More practice, less preach? The role of supervision processes and therapist characteristics in EBP implementation. *Administration and Policy in Mental Health and Mental Health Services Research*, *40*, 518–529. doi:10.1007/s10488-013-0485-5

Campbell, M. K., Piaggio, G., Elbourne, D. R., & Altman, D. G. (2012). Consort 2010 statement: Extension to cluster randomised trials. *BMJ*, *345*(e5661), 1–21. doi:10.1136/bmj.e5661

Chorpita, B. F., Daleiden, E. L., Park, A. L., Ward, A. M., Levy, M. C., Cromley, T., … Krull, J. L. (2017). Child STEPs in California: A cluster randomized effectiveness trial comparing modular treatment with community implemented treatment for youth with anxiety, depression, conduct problems, or traumatic stress. *Journal of Consulting and Clinical Psychology*, *85*(1), 13–25. doi:10.1037/ccp0000133

Chorpita, B. F., Reise, S., Weisz, J. R., Grubbs, K., Becker, K. D., & Krull, J. L. (2010). Evaluation of the brief problem checklist: Child and caregiver interviews to measure clinical progress. *Journal of Consulting and Clinical Psychology*, *78*(4), 526–536. doi:10.1037/a0019602

Chorpita, B. F., & Weisz, J. R. (2009). *MATCH-ADTC: Modular approach to therapy for children with anxiety, depression, trauma, or conduct problems*. Satellite Beach, FL: PracticeWise.

Chorpita, B. F., Weisz, J. R., Daleiden, E. L., Schoenwald, S. K., Palinkas, L. A., Miranda, J., … Research Network on Youth Mental Health. (2013). Long-term outcomes for the child STEPs randomized effectiveness trial: A comparison of modular and standard treatment designs with usual care. *Journal of Consulting and Clinical Psychology*, *81*(6), 999–1009. doi:10.1037/a0034200

Donner, A., & Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health*, *94*(3), 416–422. doi:10.2105/AJPH.94.3.416

Esserman, D., Allore, H. G., & Travison, T. G. (2016). The method of randomization for cluster-randomized trials: Challenges of including patients with multiple chronic conditions. *International Journal of Statistics in Medical Research*, *5*(1), 2–7. doi:10.6000/1929-6029.2016.05.01.1

Hengartner, M. (2018). Raising awareness for the replication crisis in clinical psychology by focusing on inconsistencies in psychotherapy research: How much can we rely on published findings from efficacy trials? *Frontiers in Psychology*, *9*(256), 1–5. doi:10.3389/fpsyg.2018.00256

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.

McMahon, F. J. (2014). Prediction of treatment outcomes in psychiatry–where do we stand? *Dialogues in Clinical Neuroscience*, *16*(4), 455–464.

Ng, M. Y., & Weisz, J. R. (2016). Building a science of personalized intervention for youth mental health. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *57*(3), 216–236. doi:10.1111/jcpp.12470

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi:10.1126/science.aac4716

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. doi:10.1177/1745691612465253

Schoenwald, S. K., Kelleher, K., Weisz, J. R., & Research Network on Youth Mental Health (2008). Building bridges to evidence-based practice: The MacArthur foundation child system and treatment enhancement projects (child STEPs). *Administration and Policy in Mental Health*, *35*(1–2), 66–72. doi:10.1007/s10488-007-0160-9

Simons, D. J. (2014). the value of direct replication. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *9*(1), 76–80. doi:10.1177/1745691613514755

Staddon, J. (2017). *Scientific method: How science works, fails to work, or pretends to work*. New York, NY: Taylor and Francis.

Steinberg, A. M., Brymer, M. J., Decker, K. B., & Pynoos, R. S. (2004). The University of California at Los Angeles post-traumatic stress disorder reaction index. *Current Psychiatry Reports*, *6*(2), 96–100. doi:10.1007/s11920-004-0048-2

Weisz, J. R., Chorpita, B. F., Frye, A., Ng, M. Y., Lau, N., Bearman, S. K., … The Research Network on Youth Mental Health. (2011). Youth top problems: Using idiographic, consumer-guided assessment to identify treatment needs and to track change during psychotherapy. *Journal of Consulting and Clinical Psychology*, *79*(3), 369–380. doi:10.1037/a0023307

Weisz, J. R., Chorpita, B. F., Palinkas, L. A., Schoenwald, S. K., Miranda, J., Bearman, S. K., … Research Network on Youth Mental Health. (2012). Testing standard and modular designs for psychotherapy treating depression, anxiety, and conduct problems in youth: A randomized effectiveness trial. *Archives of General Psychiatry*, *69*(3), 274–282. doi:10.1001/archgenpsychiatry.2011.147

Weisz, J. R., Krumholz, L. S., Santucci, L., Thomassin, K., & Ng, M. Y. (2015). Shrinking the gap between research and practice: Tailoring and testing youth psychotherapies in clinical care contexts. *Annual Review of Clinical Psychology*, *11*, 139–163. doi:10.1146/annurev-clinpsy-032814-112820

Weisz, J. R., Ugueto, A. M., Herren, J., Marchette, L. K., Bearman, S. K., Lee, E. H., … Jensen-Doss, A. (in press). When the torch is passed, does the flame still burn? Testing a "train the supervisor" model for the child STEPs treatment program. *Journal of Consulting and Clinical Psychology.*