American Psychological Association

Journal of Consulting and Clinical Psychology

0022-006X/18/$12.00

2018, Vol. 86, No. 9, 726–737
http://dx.doi.org/10.1037/ccp0000331

When the Torch Is Passed, Does the Flame Still Burn? Testing a "Train the Supervisor" Model for the Child STEPs Treatment Program

John R. Weisz
Harvard University

Ana M. Ugueto
The University of Texas Health Science Center at Houston

Jenny Herren
Warren Alpert Medical School of Brown University

Lauren K. Marchette
Cambridge Health Alliance, Harvard Medical School

Sarah Kate Bearman
The University of Texas at Austin

Erica H. Lee
Boston Children's Hospital, Harvard Medical School

Kristel Thomassin
University of Guelph

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

J. Lindsey Tweed
Maine General Health, Augusta, Maine

Jacqueline Hersh
Appalachian State University

Jacquelyn N. Raftery-Helmer
Worcester State University

Adam S. Weissman
The Child and Family Institute, New York, New York, and Teacher's College, Columbia University

Amanda Jensen-Doss
University of Miami

Objective: We assessed sustainability of an empirically supported, transdiagnostic youth psychotherapy program when therapist supervision was shifted from external experts to internal clinic staff. ***Method:*** One hundred sixty-eight youths, aged 6–15 years, 59.5% male, 85.1% Caucasian, were treated for anxiety, depression, traumatic stress, or conduct problems by clinicians employed in community mental health clinics. In Phase 1 (2.7 years), 1 group of clinicians, the Sustain group, received training in Child STEPs (a modular transdiagnostic treatment + weekly feedback on youth response) and treated clinic-referred youths, guided by weekly supervision from external STEPs experts. In Phase 2 (2.9 years), Sustain clinicians treated additional youths but with supervision by clinic staff who had been trained to supervise STEPs. Also in Phase 2, a new group, External Supervision clinicians, received training and

John R. Weisz, Department of Psychology, Harvard University; Ana M. Ugueto, Department of Psychiatry and Behavioral Sciences, McGovern Medical School, The University of Texas Health Science Center at Houston; Jenny Herren, Department of Psychiatry and Human Behavior, Warren Alpert Medical School of Brown University; Lauren K. Marchette, Department of Psychiatry, Cambridge Health Alliance, Harvard Medical School; Sarah Kate Bearman, Department of Educational Psychology, The University of Texas at Austin; Erica H. Lee, Department of Psychology, Boston Children's Hospital, Harvard Medical School; Kristel Thomassin, Department of Psychology, University of Guelph; Alisha Alleyne and Daniel M. Cheron, Judge Baker Children's Center, Harvard Medical School; J. Lindsey Tweed, Edmund M. Ervin Pediatric Center, Maine General Health, Augusta, Maine; Jacqueline Hersh, Department of Psychology, Appalachian State University; Jacquelyn N. Raftery-Helmer, Department of Psychology, Worcester State University; Adam S. Weissman, The Child and Family Institute, New York, New York, and Department of Psychology, Teacher's College, Columbia University; Amanda Jensen-Doss, Department of Psychology, University of Miami.

This research was funded by grants (to John R. Weisz) from the Annie E. Casey Foundation, the Norlien Foundation, and the John D. and Catherine T. MacArthur Foundation. We thank them for their support but acknowledge that the findings and conclusions presented in this report are those of the authors alone and do not necessarily reflect the opinions of these Foundations.

We are grateful to all the participating clinic leaders, clinicians, children, and caregivers, and we extend special thanks to Sylvie Demers, Alice Dunworth, Blanca Gurrola, Karen Mosher, and Barbara Piotti.

Author John R. Weisz receives royalties from sales of *Modular Approach to Therapy for Children with Anxiety, Depression, Trauma, or Conduct Problems.*

Correspondence concerning this article should be addressed to John R. Weisz, Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, MA 02138. E-mail: john_weisz@harvard.edu

supervision from external STEPs experts and treated referred youths. Phase 2 youths were randomized to Sustain or External Supervision clinicians. Groups were compared on 3 therapist fidelity measures and 14 clinical outcome measures. ***Results:*** Sustain clinicians maintained their previous levels of fidelity and youth outcomes after switching from external to internal supervision; and in Phase 2, the Sustain and External Supervision groups also did not differ on fidelity or youth outcomes. Whereas all 34 group comparisons were nonsignificant, trends with the largest effect sizes showed better clinical outcomes for internal than external supervision. ***Conclusions:*** Implementation of empirically supported transdiagnostic treatment may be sustained when supervision is transferred from external experts to trained clinic staff, potentially enhancing cost-effectiveness and staying power in clinical practice.

What is the public health significance of this article?
The study suggests that clinical service programs for youths may be able to sustain the implementation of an empirically supported transdiagnostic treatment without unending dependence on external expert supervision from the treatment developer team. Internal clinic staff were trained to supervise a specific empirically supported transdiagnostic treatment; the clinicians they supervised achieved treatment fidelity and youth clinical outcomes that did not differ significantly from those of clinicians supervised by external experts.

Keywords: empirically supported, psychotherapy, children, adolescents, sustainability

Over the past 2 decades, the identification of empirically supported treatments (ESTs) has markedly influenced psychotherapy research, and to some extent, clinical care. There has been an array of efforts to enrich and expand applications of empirically supported practices and make them accessible and effective in clinical practice. Examples include the development of methods for blending protocol fidelity with flexibility (Kendall & Beidas, 2007; Nock, Goldman, Wang, Albano, & Jellinek, 2004), for deciding which EST to use when the first one tried has not worked (Lei, Nahum-Shani, Lynch, Oslin, & Murphy, 2012), and for rebooting psychotherapy via an array of delivery models to make it more widely accessible (Kazdin & Blase, 2011).

Another approach—especially evident in youth psychotherapy, to date—has involved using a modular design to produce transdiagnostic treatment—that is, interventions with moveable components that can be used to create personalized treatment capable of addressing multiple problems and disorders (see, e.g., Chorpita & Weisz, 2009; Institute of Medicine, National Academy of Sciences, 2015; Weisz & Chorpita, 2011). This approach has involved identifying separable components or elements of ESTs (see Chorpita, Daleiden, & Weisz, 2005), creating separate modules for these (i.e., brief summaries of each included treatment component), and compiling a menu of modules from which clinicians may select those deemed most appropriate for a particular client. To the extent that the modules included are drawn from treatments for different disorders and problems, the modular approach can be used to address multiple individual clinical conditions, multiple forms of comorbidity, and changes that may occur over time in a youth's most pressing problems and intervention needs during a treatment episode.

One such transdiagnostic treatment model is Child STEPs (Schoenwald, Kelleher, & Weisz & the Research Network on Youth Mental Health, 2008). The STEPs model includes a modular treatment program combined with a monitoring and feedback system (MFS) used to guide clinician decision making during episodes of care. The treatment program is Modular Approach to Therapy for Children with Anxiety, Depression, Trauma, or Conduct Problems (MATCH; Chorpita et al., 2009). MATCH is a synthesis of multiple components of ESTs for the four problem clusters reflected in the name of the program. The treatment manual includes 33 modules (3- to 4-page descriptions of specific treatment components (e.g., practicing—graduated exposure for anxiety), plus handouts, treatment aids, and flowcharts to guide clinician decision making (e.g., regarding which modules to use and in which order for a particular youth and how to adjust the module sequence, depending on youth response during treatment). Clinician judgments during treatment are informed by a web-based MFS that uses brief youth- and caregiver-report measures to provide weekly updates on youth treatment response.

Published research to date has shown beneficial effects of Child STEPs. The first published study (Weisz et al., 2012; using a version of MATCH that addressed anxiety-, depression-, and conduct-related problems) was a randomized controlled effectiveness trial in 10 clinical service sites in two states, with youths aged 7–13 years. In this trial, STEPs showed clinical outcomes that were superior to usual clinical care and to standard treatment protocols for anxiety, depression, and conduct problems. A 2-year follow-up report on the same study (Chorpita et al., 2013) showed continued superiority of STEPs over usual clinical care. In a more recent randomized controlled effectiveness trial conducted in three clinical service sites with youths aged 5–15 years, Chorpita et al. (2017) found that Child STEPs (using the most current form of MATCH, which addresses anxiety, depression, trauma, and conduct problems) produced outcomes superior to those of clinicians trained in multiple standard evidence-based practices. Thus, studies to date suggest that the Child STEPs approach may be quite beneficial in everyday practice, at least when used with the external support provided through research trials. An important next step is investigating strategies for sustaining the model when that external research support fades.

This sustainability challenge is now a central focus of dissemination-implementation science (see, e.g., Aarons, Hurlburt, & Horwitz, 2011; Glasgow, Klesges, Dzewaltowski, Estabrooks, & Vogt, 2006; Simpson & Flynn, 2007; Weisz, Ng, & Bearman,

2014). Prominent leaders in the field have stressed that ". . . we have little systemic knowledge about what factors facilitate or limit sustainment of an EBP [i.e., evidence-based practice] in a service setting . . ." (Aarons et al., 2011, p. 15). To fill this gap, investigators have launched valuable pioneering work aimed at building and evaluating models for sustainability (see, e.g., Chamberlain et al., 2012; Greif, Becker, & Hildebrandt, 2015). A number of these models and the related research have involved train-the-trainer approaches to sustaining practices, including addressing such interesting questions as whether there is a scale-up penalty (i.e., a reduction in fidelity or outcomes) associated with efforts to sustain a practice (e.g., Forgatch & DeGarmo, 2011; Tommeraas & Ogden, 2017). Much of the sustainability research to date has focused on interventions that target one primary problem domain—for example, conduct problems, anxiety. As a complement to that important work, it will be useful to also address the potentially complex challenge of sustaining transdiagnostic, multicomponent modular treatment approaches. Such treatments may require that clinicians learn the equivalent of multiple separate treatments for a range of different problem domains and then learn (a) how to select the most appropriate combination of components of those treatments for the particular youth being treated, (b) how to shift the focus and navigate across different problem areas as new information arises or as new treatment needs become prominent, and (c) how to integrate information from MFSs with information from the treatment manual to inform session-by-session judgments about treatment procedures.

To use Child STEPS, for example, clinicians must learn the 33 modules spanning four protocols (i.e., anxiety, depression, trauma, conduct), learn how to integrate assessment data with decision flowcharts, learn to navigate across problem areas as new problems are identified, and young clients' most pressing problems and treatment needs shift. In addition, learning Child STEPs requires learning how to interpret the weekly feedback on youth treatment response and adapt treatment appropriately in response. To date, guidance through all these interlinked processes has been provided via ongoing supervision/consultation from STEPs experts. Indeed, the beneficial outcomes achieved in the Child STEPs trials to date have all been found with study procedures involving intensive (i.e., 1 hour) weekly clinical supervision from Child STEPs experts affiliated with one of the treatment developers. In these studies, the external supervisors have monitored the progress of each youth in the Child STEPs treatment condition, week by week, and provided feedback and guidance to the treating clinician on flowchart use, module selection, and protocol shifts, from the beginning to the end of treatment, for each youth.

Such external expert supervision is likely to be unsustainable in everyday clinical practice. One obstacle is the funding required to pay expert consultants. This can be a major challenge for clinics, which often work with extremely lean budgets (Schoenwald et al., 2008). Even if funding were not an obstacle, the availability of consultants who have the relevant specialized expertise is limited and certainly not sufficient to meet the needs of very many clinical service programs. Thus, it is understandable that one question frequently asked by participating clinic personnel and administrators is whether the use of external consultants may be phased out, with internal clinic personnel providing the supervision. The question is relevant to the sustainability of Child STEPs, in particular, and to the dissemination of complex and transdiagnostic treatments more generally.

The general question of whether supervision of a particular treatment program can be successfully transferred from external to internal personnel entails at least two specific questions: (a) what effect will such a transfer have on fidelity to the program by the clinicians who implement it and (b) what effect will the transfer have on the clinical outcomes of youths treated with the program? In the present study, we addressed both questions with particular reference to the Child STEPs program. We carried out two phases of research—each phase lasting 2.5–3 years—in three large community mental health clinics. In the first phase, a subset of clinicians from each clinic received training and weekly supervision from external STEPs experts, and those clinicians implemented STEPs with their young clients. In the second phase, the random assignment phase, supervision of the Sustain group was transferred to internal clinic staff who had been trained and supervised in STEPs and in providing STEPs supervision. Also during this second phase, a second group of clinicians in the same clinics—the External Supervision (ES) group—received STEPs training and supervision only from the external STEPs experts and implemented STEPs with their young clients. Throughout Phase 2, youths were randomly assigned to Sustain or ES clinicians.

This two-phase design of this hybrid effectiveness-implementation study made it possible to test two questions related to sustainability: (a) Are clinicians who are trained and experienced in STEPs (i.e., the Sustain group) able to maintain their previous levels of fidelity and youth outcomes after external STEPs supervision is withdrawn and replaced by internal clinic supervision; and (b) when trained and experienced STEPs clinicians (i.e., the Sustain group) shift to internal staff supervision, how do their fidelity and youth outcomes compare with the fidelity and youth outcomes achieved concurrently by clinicians who are using STEPs with external expert supervision (i.e., the ES group)?

Method

Informed consent/assent was obtained from all caregivers/youths prior to their participation, and all study procedures were institutional review board (IRB) reviewed and approved.

Overview of Study Procedures

The study spanned two phases in two successive time periods. In Phase 1, 26 clinicians from three outpatient community mental health clinics used STEPs (described below) to treat youths who had been referred to their clinics through normal community pathways (i.e., we did not advertise or recruit youths). These clinicians received STEPs training (in three 2-day workshops, separated by 3–5 weeks) and supervision (1 hr per week throughout Phase 1) from external STEPs experts affiliated with the study team, and they treated referred youths as described. Also during Phase 1, the external STEPs experts trained and supervised a group of six clinical staff from the three clinics to prepare them to provide STEPs supervision in Phase 2. Phase 1 (including training, intake assessments, active treatment, and posttreatment assessments) lasted 2 years 8 months and 1 day.

We then transitioned to Phase 2, the random assignment phase (see CONSORT diagram in Figure 1). The 14 clinicians from

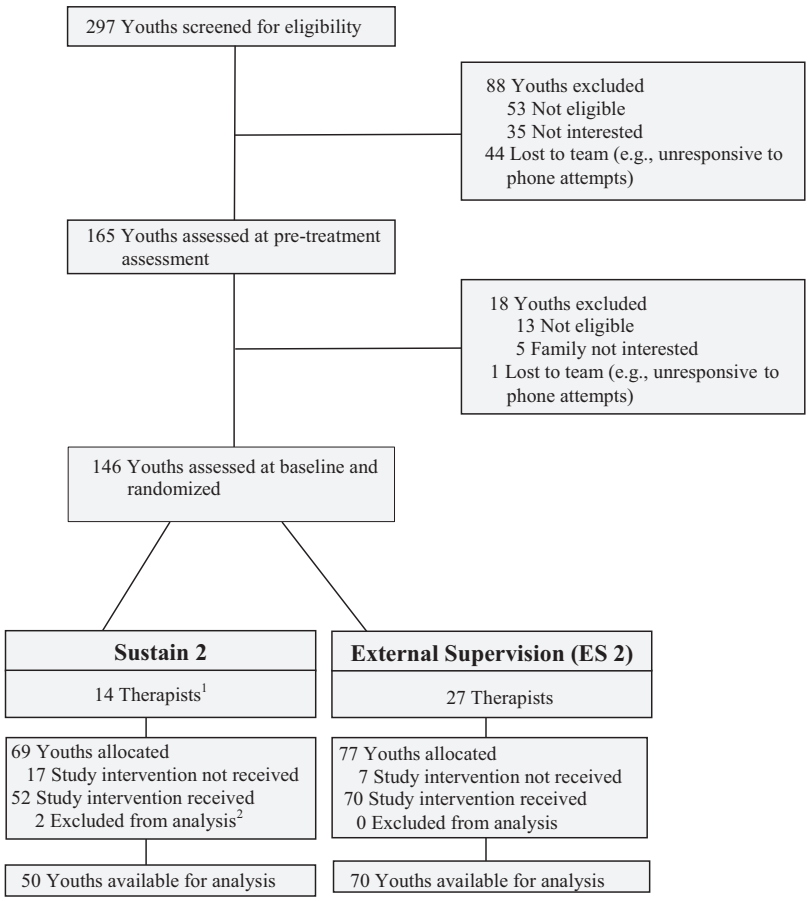

Figure 1. Flow diagram (CONSORT) showing sampling and randomization processes for the Phase 2 randomized trial comparing the Sustain 2 and External Supervision groups. [1]The 14 Sustain 2 therapists had been trained in MATCH during Phase 1. [2]One youth was found to have a sibling in the study, so one sibling was randomly excluded; one youth was excluded because of a caregiver/reporter change during the study.

Phase 1 who were still working in one of the clinics and available for the study continued to treat new cases using STEPs, but for Phase 2 their supervision was transferred to the clinic staff who had prepared to be STEPs supervisors. These 14 clinicians, referred to here as the Sustain group, did not differ significantly from the 12 who had departed their agencies on any demographic or professional characteristics or on STEPs fidelity measures (see *Assessing Fidelity*, below). Also in Phase 2, a new group of 27 clinicians from the same three clinics, who had not previously been trained in STEPs, received the same training and supervision from external STEPs experts in the study team that had been provided to the Sustain group in Phase 1. This new group, referred to here as the ES group, also treated referred youths. All referred youths in Phase 2 were randomly assigned to either Sustain or ES clinicians. Phase 2 (including training, intake assessments, active treatment, and posttreatment assessments) lasted 2 years 11 months and 9 days.

Clinicians in Phase 1 attended a mean of 42.91 hr of training; clinicians in Phase 2 attended a mean of 43.39 hr. Whereas training had a standard duration, supervision hours varied across clinicians according to their length of involvement in the study, the number of study youths each clinician treated, and duration of treatment episodes. Mean hours of supervision per clinician was 57.21 in Phase 1 and 70.87 in Phase 2, values that are difficult to interpret, given variation in the factors that influenced supervision time per clinician.

The primary study questions were addressed by examining MATCH fidelity and youth treatment outcomes in two comparisons. Comparison 1 involved Sustain clinicians in Phase 1 versus Phase 2—that is, during versus after external expert STEPs supervision was provided. Comparison 2 involved Sustain clinicians versus ES clinicians in Phase 2—that is, when there was concurrent treatment by different clinician groups, one with and one without external expert STEPs supervision.

Therapist Participants and Service Settings

Therapist participants were 41 clinicians (14 Sustain and 27 ES) used in the participating clinics, who routinely provided psychotherapy to youths in those clinics, who volunteered for the study (clinics did not require participation), and who gave written consent following university-approved IRB procedures. Compensation was provided from project funds for all research-related activities, based on insurance reimbursement rates, to ensure that

project participation would not result in loss of income. Table 1 presents the demographic and professional characteristics of the therapist participants. The therapist samples resembled those of published STEPs trials in mean age (45.4 and 45.05 in the Sustain and ES groups, 40.6 in Weisz et al. [2012], not reported in Chorpita et al., [2017]) and gender (93%, 82%, 80%, and 96% female, respectively) but had less minority group representation (7%, 30%, 44%, not reported in Chorpita et al., [2017]) and more years of professional experience (13.2, 10.7, 7.6, 3.3); the modal professional specialty in each sample was social work, although the percentages differed across studies (79%, 59%, 40%, 40%). To place our sample in a broader context, we compared our clinician characteristics with those reported in a nationally representative survey of the attitudes of 1,112 clinicians from 100 clinics in 26 U.S. states (Aarons et al., 2012); this sample averaged 28.26 years of age, was 76.4% female and 29% minority, and reported 10.76 years of professional experience (professional specialty not reported). Because Phase 2 involved randomizing youths to Sustain or ES clinicians, we compared those two therapist groups in Phase 2 on demographic and professional characteristics; there were no significant group differences (see Table 1).

The clinics were recruited because (a) they were among the largest in their geographic areas, together providing a large proportion of the youth mental health care in the region, and (b) their leaders and clinicians expressed sufficient interest in participating that they appeared to be committed, reliable partners. The clinics, which had not participated in the previously noted STEPs studies by Weisz et al. (2012) or Chorpita et al. (2013, 2017), were free-standing independent nonprofit entities funded through reimbursement for services, not supported by large county- or state-funded training initiatives, and using clinicians with a variety of backgrounds and training. Services for the clientele of the clinics were paid by insurance, primarily Medicaid.

Table 1
Clinician Demographic and Professional Characteristics

Characteristics	Sustain $N = 14$	ES $N = 27$
Demographic characteristics		
M (*SD*) age in years	45.4 (10.9)	45.05 (13.6)
Gender		
Female, %	92.9%	81.5%
Race/ethnicity		
Caucasian, %	92.9%	70.4%
Other, %	7.1%	3.7%
Professional characteristics		
M (*SD*) years of professional experience	13.2 (7.0)	10.7 (6.8)
Licensure		
Licensed, %	100.0%	77.8%
Professional discipline		
Social worker, %	78.6%	59.3%
Other (e.g., licensed professional counselor), %	21.4%	22.2%

Note. ES = External Supervision; Figures are based on clinician-reported data. Per institutional review board consent procedures, five clinicians chose not to report discipline or licensure, and seven chose not to report race/ethnicity.

Youth Participants

Some 168 youth participants and their caregivers gave informed assent/consent to study participation following IRB-approved procedures. In both phases, the youths were 6–15 years old, had been referred through normal community channels, and presented with clinical problems including anxiety, depression, posttraumatic stress, and/or conduct problems. Following MATCH requirements, youths were excluded if there was evidence of intellectual disability, pervasive developmental disorder, psychotic symptoms, bipolar disorder, or if their top ranked clinical concern involved inattention or hyperactivity. The sample was 59.5% male, with 85.1% Caucasian/Euro, 2.4% African American, 1.8% Latino, 0.6% Asian American, and 10.1% mixed and other. The study involved comparisons among three groups of these youths: (a) Sustain 1, that is, 48 youths treated by Sustain clinicians in Phase 1 who were being supervised during that phase by external STEPs experts; (b) Sustain 2, that is, 50 youths treated by Sustain clinicians in Phase 2 after the clinicians had transitioned from external supervision to internal supervision provided by their clinic's staff; and (c) ES 2, that is, 70 youths treated by ES clinicians in Phase 2 using STEPs with supervision from external experts.

Table 2 presents demographic and clinical characteristics of the three youth samples. The table shows three significant differences between the Sustain 1 and ES 2 groups—groups that were not compared in any of the main study analyses; those two groups differed on Youth Self-Report (YSR; see *Measures* below) Total Problems, Brief Problem Checklist (BPC; see *Measures* below) total score, and the Top Problems Assessment (TPA; see *Measures* below). The groups that were to be compared in the planned study analyses did not differ on any of the measures.

Measures

Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001). The CBCL is a widely used 118-item caregiver-report measure of youth emotional and behavioral problems. Items are rated as 0 (*not true*), 1 (*somewhat or sometimes true*), or 2 (*very true or often true*). We used the Total Problems scale and the two broad-band Internalizing and Externalizing problem scales. Validity and reliability of these scales are well documented in previous research (see Achenbach & Rescorla, 2001). We administered the CBCL at 3, 6, 9, 12, 18, and 24 months.

Youth Self-Report (Achenbach & Rescorla, 2001). The YSR is a 118-item youth self-report counterpart to the CBCL. We used the Total Problems scale and the Internalizing and Externalizing broadband scales, all supported by extensive validity and reliability evidence (Achenbach & Rescorla, 2001). The YSR was originally designed for youths aged 11–18 years, but subsequent research has indicated that that the Total Problems, Internalizing, and Externalizing scales are reliable and valid for children as young as age 7 years (Ebesutani, Bernstein, Martinez, Chorpita, & Weisz, 2011; Yeh & Weisz, 2001). We administered the YSR at 3, 6, 9, 12, 18, and 24 months.

Brief Problem Checklist (Chorpita et al., 2010). To measure outcome trajectories over time with a standardized measure, we used the BPC, a 12-item measure of internalizing, externalizing, and total problems, developed through application of item response theory and factor analysis to data from the CBCL and YSR. Factor analysis of the BPC yielded two factors—that is, internal-

Table 2
Youth Demographic and Clinical Characteristics

Characteristics	Sustain 1 $N = 48$	Sustain 2 $N = 50$	ES 2 $N = 70$	Three-group comparison
Demographic characteristics				
M (*SD*) age in years	9.7 (2.8)	9.3 (2.6)	9.6 (2.6)	$F(2, 165) = .4, p = .68$
Male, %	52.1%	60.0%	64.3%	$\chi^2(2) = 1.8, p = .41$
Caucasian, %	81.3%	86.0%	87.1%	$\chi^2(2) = .8, p = .66$
Annual family income				$\chi^2(10) = 9.1, p = .52$
$0–19,000	37.5%	24.0%	31.4%	
$20,000–39,000	35.4%	42.0%	31.4%	
$40,000–59,000	16.7%	14.0%	14.3%	
$60,000–79,000	8.3%	4.0%	7.1%	
$80,000–99,000	.0%	10.0%	8.6%	
$100,000 or more	2.1%	6.0%	7.1%	
Clinical characteristics				
Receiving medications, %	45.8%	48.0%	37.1%	$\chi^2(2) = 1.6, p = .44$
M (*SD*) T1 CBCL total	67.3 (6.7)	69.0 (5.9)	68.5 (6.7)	$F(2, 162) = .9, p = .39$
M (*SD*) T1 YSR total	54.6 (9.9)[a]	58.6 (9.5)	59.8 (9.4)[b]	$F(2, 121) = 3.1, p = .048$
M (*SD*) T1 BPC parent total	9.8 (4.5)	10.6 (4.3)	10.5 (4.1)	$F(2, 162) = .5, p = .60$
M (*SD*) T1 BPC youth total	5.4 (4.0)[a]	6.9 (4.4)	7.8 (4.1)[b]	$F(2, 121) = 3.4, p = .037$
M (*SD*) T1 TPA parent mean	6.7 (2.1)	7.1 (2.5)	6.8 (2.1)	$F(2, 162) = .5, p = .62$
M (*SD*) T1 TPA youth mean	4.6 (2.4)[a]	5.8 (2.6)	6.3 (2.4)[b]	$F(2, 115) = 4.7, p = .012$

Note. CBCL = Child Behavior Checklist; YSR = Youth Self-Report; BPC = Brief Problem Checklist; TPA = Top Problems Assessment. In the case of significant omnibus tests of group differences, post hoc tests (pairwise χ^2 tests or Tukey's tests) were conducted. Groups with different superscripts in the same row differed significantly from one another.

izing (six items; scores can range from 0 to 12) and externalizing (six items; scores can range from 0 to 12). Psychometric analyses have shown strong internal consistency and test-retest reliability as well as large significant correlations between BPC scales and the corresponding CBCL and YSR scales. Longitudinal data have shown that the BPC scales significantly predict change on related measures of youth symptoms during treatment (Chorpita et al., 2010). We administered the BPC weekly throughout treatment.

Top Problems Assessment (Weisz et al., 2011). To measure outcome trajectories over time using an idiographic, consumer-driven approach, we used the TPA. The TPA procedure involves having youths and caregivers identify, separately at the beginning of treatment, the top problems each considers most important to work on in treatment. Subsequently, each week throughout treatment, youths are asked to rate the severity of the problems they identified, and caregivers the problems they identified (each using a 0–10 scale). Research on the TPA with clinically referred 7- to 13-year-olds and their caregivers (Weisz et al., 2011) has shown strong evidence of test-retest reliability, convergent and discriminant validity, sensitivity to change during treatment, slope reliability, and the association of TPA slopes during treatment with standardized measure slopes. We administered the TPA weekly throughout treatment.

STEPs Treatment, Clinician Decision-Making, Monitoring, and Feedback System

Clinicians in the study were trained and supervised in the use of MATCH (Chorpita & Weisz, 2009) and the web-based MFS that displayed BPC and TPA scores from youths and caregivers together with a record of which modules had been used and when, all within an electronic dashboard for each youth. Clinicians discussed their cases with their supervisor each week, informed by the dashboard.

STEPs Supervisors and Supervision Model

The study involved two sets of STEPs supervisors: external STEPs supervisors and internal clinic supervisors. Supervision for the Sustain clinicians in Phase 1 and the ES clinicians in Phase 2 was provided by eight external STEPs supervisors, all trained and experienced in STEPs and all part of the study team working with one of the treatment developers. All held PhDs in clinical psychology; 75% were female, and their mean age at study baseline was 30.38 years (*SD* = 2.45 years), mean years of clinical experience was 7.06 (*SD* = 1.90 years), and race/ethnicity was reported as 87.5% Caucasian only and 12.5% Hispanic and Caucasian.

Supervision for the Sustain clinicians in Phase 2 was provided by six internal clinic staff STEPs supervisors; all were clinic employees, and none was part of the study team working with the treatment developer. Eligibility required completion of a previous STEPs training program plus previous experience delivering MATCH (for a minimum of four treated youths, encompassing a minimum of three of the primary problem areas—anxiety, depression, posttraumatic stress, and conduct), with supervision from external expert STEPs supervisors. To be eligible, clinicians were not required to be current supervisors in the clinics, but all were required to have a mental health-related professional degree and to be employed in the clinic in a child/adolescent psychotherapy role. Given eligibility, the internal supervisors were identified through collaboration between MATCH expert consultants (who evaluated their knowledge and use of MATCH skills and STEPs procedures) and clinic leaders (who evaluated their appropriateness and eligibility from a clinic administrative and personnel perspective). All

the internal supervisors identified had Master of Social Work degrees, and five were Licensed Clinical Social Workers. Some 83% were female, mean age at study baseline was 46.67 years (SD = 10.86 years), mean years of clinical experience was 12.83 years (SD = 5.08 years), and race/ethnicity was reported as 100% Caucasian. Preparation for their supervision roles occurred prior to Phase 2 and included (a) a 7-hr training in STEPs supervision, conducted by external expert supervisors, (b) 6-months observing weekly supervision of clinicians by an external STEPs supervisor, and (c) a step-down sequence in which the internal supervisor was joined by a collaborating external STEPs supervisor weekly for 1 month, biweekly for 1 month, and then once in the third month. After these steps had been completed, the internal supervisors independently supervised their own clinic's clinicians in the use of STEPs with youths referred to their clinics.

The supervision model, used with all supervisors across study conditions and phases, was designed to support supervisors' efforts to guide clinicians in effective use of MATCH. Before meeting with clinicians, the supervisor contacted all clinicians in the group via e-mail or fax to (a) provide summary updates on all active cases [for example, session attendance, modules covered, whether MFS dashboard scores were improving, stable, or growing worse], (b) prioritize cases for discussion [new cases, recent and looming crises, worst MFS dashboards, and clinician requests given priority], and (c) provide a draft agenda for the upcoming meeting]. During the meeting, the steps included (a) review of draft agenda with revision if needed, (b) brief check-in regarding all active cases [for example, did session occur, what was covered and with which module(s), and any new issue needing discussion] (c) discussion of prioritized cases, including trouble-shooting problems arising during sessions, and planning for next sessions, (d) MATCH skill-building exercises with supervisor modeling, clinician practice and role-play, and feedback—procedures resembling but less standardized than the behavioral rehearsal approach described by Edmunds et al. (2013) and Dorsey et al. (2017), (e) discussion of common themes arising during the meeting, and (e) summary of action steps and soliciting of clinician input regarding the agenda for the next meeting.

All supervisors, across conditions and phases, participated in our standard weekly group peer supervision meetings—internal supervisors with one another, external MATCH supervisors with one another—to review challenging cases, problem solve any difficulties faced by clinicians, and provide feedback to one another on treatment plans. To support thoughtful discussion of the clinical problems being treated, half the meetings of each group included a treatment researcher with expertise in internalizing problems, and half included a treatment researcher with expertise in externalizing problems. These individuals provided perspectives on the two broad forms of psychopathology being addressed in treatment, but neither had experience or expertise in MATCH or STEPs. Thus, the peer supervision for internal supervisors included no input from MATCH experts, the peer supervision for external supervisors included all MATCH expert supervisors, and both groups had support from an internalizing and externalizing problem consultant.

Assessing Fidelity: Coding Treatment for MATCH Adherence and Competence

To assess fidelity to MATCH, trained coders who were blind to study condition coded 493 digital recordings of therapy sessions (28% of the total of 1,788 available) randomly selected from Sustain 1 (N = 209 recordings sampled from N = 870 available), Sustain 2 (N = 92 recordings sampled from N = 302 available), and ES 2 (N = 192 recordings sampled from N = 616 available). Session content was coded using the Therapist Integrity in Evidence Based Interventions (TIEBI) coding system (see Bearman, Herren, & Weisz, 2012; Bearman, Schneiderman, & Zoloth, 2017; Weisz, Bearman, Santucci, & Jensen-Doss, 2017). Within the TIEBI, session recordings are coded, in 5-min segments, for the presence/absence of 22 items reflecting therapist adherence and competence in the use of MATCH. Coding of adherence is based on the percent of 5-min segments in which prescribed content from MATCH was present. Coding of therapist competence is based on coders' global ratings of skillfulness of delivery of each item of MATCH content, ranging from 0 = *not at all* to 4 = *expert*. TIEBI coders (N = 10) were bachelor's- and masters'-level research assistants in the primary coding system developer's laboratory. Following coder training, practice sessions, and reliability screening, the coders coded sessions that were randomly selected using the following procedures: (a) First sessions were omitted (these often included clinic administrative content), (b) all remaining sessions were divided into thirds (early, middle, late phase of treatment), and (c) one session was randomly selected for coding from each of these three phases, omitting sessions shorter than 15 min or longer than 75 min (these were typically unrepresentative—e.g., clinic paperwork). Of the 1,788 total audible recordings, 493 were randomly assigned by session phase to the 10 coders, with each recording masked as to participant characteristics and study condition.

To generate a mean percent adherence score for each coded treatment session for each MATCH content item, the number of 5-min segments in which that item was coded present was summed, multiplied by five (number of minutes in each segment), and divided by the total time of the session in minutes. The resulting mean percentages were averaged across all the coded sessions for each youth's treatment episode. To generate a mean competence score for each youth's treatment episode for each MATCH content item, each of the global competency codes for each present item were averaged across all global competence codes in a given session, and these means were, in turn, averaged across all coded sessions in the treatment episode. A composite competence score was also calculated by identifying the highest competency score across all MATCH items present in a session and then calculating the mean of those highest scores averaged across all sessions. The composite score reflected the fact that each session could require concentrated emphasis on a particular MATCH skill (e.g., practicing or exposure), and thus, the highest level of competence achieved for any one among the various MATCH items item is especially important to assess. To summarize, the coding system generated three measures of fidelity: therapist adherence, mean therapist competence, and composite therapist competence.

Over the course of the coding, 83 sessions were randomly selected for double coding to assess agreement between independent coders. Reliability, across pairs of coders, ranged from intraclass correlation coefficient (ICC; 1,1) = 0.71 to ICC (1,1) = 0.99, with a mean of ICC (1,1) = 0.92 for MATCH adherence, and from ICC (1,1) = 0.70 to ICC (1,1) = 0.97, with a mean of ICC (1,1) = 0.88 for MATCH competence.

Experimental Design and Group Comparisons: Study Questions 1 and 2

The study design involved a series of two-group comparisons on MATCH fidelity and clinical outcome measures. By comparing Sustain 1 with Sustain 2, we addressed study question 1: Are clinicians who are trained and experienced in STEPs (i.e., the Sustain group) able to maintain their previous levels of fidelity and youth outcomes after external STEPs supervision is withdrawn and replaced by internal clinic supervision? By comparing Sustain 2 with ES 2, we addressed study question 2: When trained and experienced STEPs clinicians (i.e., the Sustain group) shift to internal staff supervision, how do their fidelity and youth outcomes compare with the fidelity and youth outcomes achieved concurrently by clinicians who are using STEPs with external expert supervision (i.e., the ES group)?

Results

Data Analysis Plan

Study analyses were modeled after those used in the original STEPs trial (Weisz et al., 2012) and were conducted using SPSS Statistics version 24.0 (IBM Corp., 2016) or HLM 7.01 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011). Prior to analysis, patterns of missing data were examined. Given the varying number of outcome assessments across participants, the first and last values for each of these variables were used in the missing data analysis. All variables had less than 5% missing data other than the adherence variables (13.7%). Little's missing completely at random (MCAR) tests, conducted once using the parent-report data in the full sample and again including the youth-report data in the subsample old enough to provide youth report data, indicated data were missing completely at random [full sample: $\chi^2(78) = 55.37$, $p = .98$; youth report sample: $\chi^2(194) = 81.42$, $p = 1.0$]. Ten multiply imputed data sets were generated to replace the missing adherence data using the Blimp program, version 1.0 (Keller & Enders, 2017), which uses a chained equations approach for imputing multilevel data (Enders, Keller, & Levy, 2017).

Analyses of outcome measures were conducted via random-effects, three-level multilevel models (repeated measure nested within participants who were nested within supervisors[1]). At level 1, each participant's loglinear rate of change over time was estimated using all available measurements (see Hedeker & Gibbons, 2006), and the independent variables (Sustain 2 vs. Sustain 1 for question 1; Sustain 2 vs. ES 2 for question 2) were entered as level two predictors of intercept and slope. Given that there were only three clinics, clinic was dummy coded and entered as intercept predictors at level 3. To assess the need for a three-level modeling approach, the proportion of variance in model intercepts and slopes at the youth and supervisor levels were computed. Across outcome variables, between .03% and 12.5% of variance in intercept and between .04% and 45.2% of the variance in slope were at the supervisor level. Taking cluster size into account, this yielded design effect estimates ranging from 1.0 to 4.2. Given the small average cluster size of nine for the first research question and 11 for the second, design effects greater than 1.1 were considered nonignorable (Lai & Kwok, 2015), suggesting three-level models were appropriate.

Analyses of fidelity scores were similar, although these involved two-level models (clients within supervisors). The need for this nesting was supported by the fact that 1.1–5.6% of the variance in fidelity was at the supervisor level (design effects = 1.1–1.3). These analyses were conducted using the multiple imputation feature in HLM 7.01, which conducts separate analyses for each dataset and aggregates them using a combination of averaging and more complex equations that account for error variance due to error in imputations (Raudenbush et al., 2011).

For the outcome analyses, effect sizes were calculated using the same approach used by Weisz et al. (2012), dividing the Group X slope parameter by the square root of the variance for the slope parameter. For the fidelity measures, the group differences parameter was divided by the raw data standard deviation. Both types of effect sizes are estimates of Cohen's (1988) d, where .2 is considered small, .5 medium, and .8 large. The Optimal Design Plus Empirical Evidence version 3.01 (Spybrook et al., 2011) was used to assess power based on number of clusters (i.e., therapists), sample size within clusters (i.e., average number of clients per therapist), and intraclass correlation coefficients (i.e., proportion of variance in slope accounted for by clinician). For study question 1, the outcome analyses had power of .80 to detect effect sizes between .67 and .86, and the fidelity analyses were powered to detect effect sizes between .70 and .88. For study question 2, the outcome analyses had power of .80 to detect effect sizes between .46 and .55, and the fidelity analyses were powered to detect effect sizes between .47 and .54.

Question 1: Fidelity and Outcomes After Transition to Internal Supervision

By comparing the Sustain 1 and Sustain 2 samples, we addressed study question 1: Are clinicians who are trained and experienced in STEPs (i.e., the Sustain group) able to maintain their previous levels of fidelity and youth outcomes after external STEPs supervision is withdrawn and replaced by internal clinic supervision?

MATCH fidelity. The analyses for Question 1 are presented in Table 3. Taken together, the findings indicated that the Sustain clinicians did maintain their previous level of MATCH fidelity after they transitioned from external supervision to internal clinic staff supervision. This was reflected in the measure of therapist adherence and in the mean and composite measures of therapist competence in MATCH. None of the fidelity differences between Sustain 1 and Sustain 2 approached significance. The effect sizes for these comparisons ranged from .09 to .13, favoring Sustain 2.

Clinical outcomes. There were no statistically significant differences in youth outcome between the Sustain 1 and Sustain 2

[1] Participants were also nested within therapists, but these therapists were not fully nested within supervisors, and some supervisors supervised only a small number of clinicians. Given the complexity involved in modeling this type of nesting structure relative to our sample size, we decided to account for only one type of nesting in the study analyses. Given that therapists accounted for similar amounts of variance in outcomes to supervisors (design effects 1.0–3.2) and the focus of this paper was on type of supervision, the decision was made to run the analyses with supervisor-level nesting. The results for models with therapist-level nesting were nearly identical to those reported here and are available from the authors by request.

Table 3
Fidelity and Clinical Outcomes of Sustain Clinicians Under External Expert Supervision (Sustain 1) and Internal Supervision (Sustain 2)

Measures	Sustain 2 vs Sustain 1[a]	*p*	*d*	Sustain 1 Mean	Sustain 2 Mean		
Fidelity							
Mean adherence percent	3.06%	.70	.09	64.78%	67.84%		
Mean competence	.11	.67	.12	2.17	2.28		
Composite competence	.15	.63	.13	2.48	2.63		
				Estimated 1-year change Sustain 1	Estimated 1-year change Sustain 2	Estimated 2-year change Sustain 1	Estimated 2-year change Sustain 2
Clinical outcomes							
BPC parent total	.35	.35	.24	−7.08	−5.01	−7.91	−5.60
BPC youth total	.13	.75	.09	−4.86	−4.11	−5.43	−4.59
BPC parent internalizing	.22	.41	.24	−3.16	−1.86	−3.53	−2.08
BPC youth internalizing	.03	.91	.03	−1.95	−1.79	−2.18	−2.00
BPC parent externalizing	.15	.49	.18	−3.97	−3.08	−4.43	−3.44
BPC youth externalizing	.08	.68	.13	−2.89	−2.41	−3.23	−2.69
TPA parent mean	−.16	.42	−.20	−4.79	−5.75	−5.36	−6.42
TPA youth mean	−.21	.44	−.25	−4.59	−5.84	−5.13	−6.53
CBCL total	−.07	.78	−.09	−6.92	−7.35	−7.73	−8.22
YSR total	−.68	.11	−.77	−6.98	−11.02	−7.80	−12.31
CBCL internalizing	−.02	.95	−.02	−7.84	−7.97	−8.77	−8.91
YSR internalizing	−.58	.21	−.58	−7.18	−10.61	−8.02	−11.85
CBCL externalizing	−.10	.74	−.10	−6.26	−6.83	−7.00	−7.64
YSR externalizing	−.59	.12	−.99	−4.22	−7.67	−4.71	−8.57

Note. BPC = Brief Problem Checklist; TPA = Top Problems Assessment; CBCL = Child Behavior Checklist; YSR = Youth Self-Report. Group membership was dummy coded as 1 = Sustain 2, 0 = Sustain 1, so negative regression coefficients indicated lower fidelity and faster symptom improvement in the Sustain 2 group; d = estimated Cohen's d effect size.
[a] Regression coefficient is from multilevel models.

clients. ESs ranged from −.99 favoring Sustain 2 to .24 favoring Sustain 1. Three of the nonsignificant tests were associated with medium to large effect sizes, but all of those (YSR Total Problems, YSR Internalizing, and YSR Externalizing) showed trends indicating better clinical outcomes for Sustain 2 than the Sustain 1 group—that is, better outcomes after transition to internal clinic supervision.

Question 2: Fidelity and Outcomes With Concurrent Internal Versus External Supervision

By comparing Sustain 2 with ES 2, we addressed study question 2: When trained and experienced STEPs clinicians (i.e., the Sustain group) shift to internal staff supervision, how do their fidelity and youth outcomes compare with the fidelity and youth outcomes achieved concurrently by clinicians who are using STEPs with external expert supervision (i.e., the ES group)?

MATCH fidelity. Findings for question 2 are presented in Table 4. Taken together, the findings indicate that the Sustain clinicians, after transitioning to internal staff supervision, showed MATCH fidelity that was highly similar to the fidelity shown by ES clinicians being supervised by external experts. On none of the three fidelity measures—that is, therapist adherence, mean therapist competence, and composite therapist competence—did group differences approach significance; the three effect sizes ranged from −.16 favoring Sustain 1 to +.08 favoring Sustain 2.

Clinical outcomes. As detailed in Table 4, there were no statistically significant differences between the Sustain 2 and ES 2 clients. ESs ranged from −.16 favoring Sustain 2 to .35 favoring Sustain 1; all effect sizes were either small or below the cutoff for a small effect size.

Discussion

We tested the sustainability of the transdiagnostic STEPs model when expert external supervision was withdrawn and replaced by internal clinic staff who had been trained to supervise STEPs. We focused on fidelity to the MATCH program used in STEPs and on clinical outcomes achieved by the treated youths, and we used two kinds of comparisons. First, we investigated whether clinicians who were trained and experienced in STEPs (i.e., the Sustain group) were able to maintain their previous levels of fidelity and youth outcomes after external supervision was withdrawn and replaced by internal clinic supervision. Second, we asked, when staff clinicians who had previous training and experience in STEPs (i.e., the Sustain group) were shifted to internal staff supervision, how their fidelity and youth outcomes would compare with the fidelity and youth outcomes achieved concurrently by clinicians who were using STEPs with external expert supervision (i.e., the ES group).

The answers to these questions showed a consistent pattern: All the group comparisons—across three fidelity measures and 14 clinical outcome measures—showed nonsignificant group differences. This was true in the comparison of Sustain clinicians before versus after their transition to internal clinic supervision and also true in the concurrent comparison of internally supervised Sustain clinicians with externally supervised ES clinicians. When we calculated effect sizes for the 34 nonsignificant group differences, we found that 85% fell at or below the mean found in a meta-analysis of EST versus usual care comparisons (.29, Weisz et al., 2013), only three fell within the medium-to-large effect range, and these three effects showed the Sustain clinicians performing better after their transition to internal staff supervision than they had under

Table 4
Fidelity and Clinical Outcomes of Sustain Clinicians and External Supervision Clinicians With Randomly Assigned Youths in Phase 2

Measures	Sustain 2 vs. ES 2[a]	*p*	*d*	ES 2 Mean	Sustain 2 Mean		
Fidelity							
Mean adherence percent	−5.30%	.25	−.16	69.86	64.57		
Mean competence	.074	.67	.08	2.11	2.18		
Composite competence	−.056	.83	−.05	2.59	2.53		
				Estimated 1-year change ES 2	Estimated 1-year change Sustain 2	Estimated 2-year change ES 2	Estimated 2-year change Sustain 2
Clinical outcomes							
BPC parent total	.33	.30	.22	−7.43	−5.49	−8.31	−6.13
BPC youth total	.36	.22	.32	−6.80	−4.68	−7.60	−5.23
BPC parent internalizing	.16	.39	.22	−3.08	−2.12	−3.44	−2.37
BPC youth internalizing	.28	.17	.35	−3.95	−2.31	−4.42	−2.58
BPC parent externalizing	.17	.32	.23	−4.42	−3.42	−4.93	−3.83
BPC youth externalizing	.07	.65	.13	−2.82	−2.42	−3.15	−2.71
TPA parent mean	−.13	.45	−.16	−5.23	−6.01	−5.85	−6.71
TPA youth mean	.18	.44	.19	−6.73	−5.66	−7.52	−6.33
CBCL total	.23	.34	.29	−9.09	−7.72	−10.15	−8.63
YSR total	−.11	.78	−.14	−11.89	−12.51	−13.28	−13.98
CBCL internalizing	.25	.42	.26	−9.95	−8.49	−11.12	−9.48
YSR internalizing	.20	.62	.27	−13.31	−12.12	−14.87	−13.55
CBCL externalizing	.15	.57	.16	−7.98	−7.09	−-8.92	−7.92
YSR externalizing	−.11	.78	−.14	−7.35	−8.56	−8.22	−9.57

Note. BPC = Brief Problem Checklist; TPA = Top Problems Assessment; CBCL = Child Behavior Checklist; YSR = Youth Self-Report. Group membership was dummy coded as 1 = Sustain 2, 0 = ES 2, so negative regression coefficients indicated lower fidelity and faster symptom improvement in the Sustain 2 group; d = estimated Cohen's d effect size.
[a] Regression coefficient is from multilevel models.

external supervision. Taken as a whole, the results did not clearly favor one group over another, and if anything they slightly favored the more experienced internal supervision group, suggesting that STEPs fidelity and clinical outcomes can be sustained.

More broadly, the findings suggest one process through which a rather complex intervention program might be sustained over time after a period of initial training and external supervision have ended—and sustained with protocol fidelity and clinical outcomes maintained. The process involves preparing internal clinic staff supervisors to take over the supervision role of the external experts. This seems certain to reduce cost and practical barriers to program sustainability (e.g., limited availability of external expert supervisors, even if funds were available to pay them), but it may offer additional advantages: Internal supervisors are likely to draw not only on their knowledge of the treatment protocol but also on their own clinical expertise, on their knowledge of and relationships with the clinicians they are supervising, and on their understanding of the clinic clientele, social and economic conditions in the clinic's geographic area, and opportunities for support through clinic resources and networks. That said, implementation of this sustainment model would require funding for the initial training and supervision needed, and some compensation for lost reimbursement income to the clinicians and clinics—funding that might well exceed internal resources of most clinics, necessitating some form of outside support. Implementing the model might also require adjustments in the ways supervisory roles are defined and assigned, potentially altering traditional clinic practice in some respects. Certainly, considerable tracking over time will be required to gauge the capacity of existing clinic structures to support extended, long-term application of the sustainment model examined here, both for STEPs and other empirically tested treatments.

Our study had certain limitations that suggest directions for future research. First, whereas our results are encouraging, they reflect a cautious, stepwise approach toward independent implementation of STEPs by clinics. The approach included preservation of STEPs training by external experts, followed by a gradual transition to internal supervision, via procedures designed to support skill-building and supervision effectiveness by the internal clinic staff supervisors; and the Sustain 2 clinicians had acquired a substantial dose of STEPs experience during Phase 1, which might help explain their trend toward better outcomes compared with Sustain 1, and their similarity to the ES2 group in fidelity and clinical outcomes. The dose of supervised experience in Phase 1 reduces the generalizability of our findings to internal supervision models involving less (or no) experience prior to the sustainment phase. Given the apparent success of our cautious scaffolding approach, it may now be appropriate for future tests to involve faster reduction in the scaffolding provided to determine whether there is a tipping point beyond which fidelity and clinical outcomes are undermined. Such a future direction could be facilitated by measures of fidelity to the supervision model and measures of supervisor readiness. Given the array of clinical skills involved in the supervision model, we chose to rely on the judgment of MATCH experts, but if it were actually possible to create valid measures of those supervision skills, such measures could be quite useful in future research. Whether through the use of measures or expert feedback, the development of optimum strategies for evaluating supervisor readiness and supporting supervisors in developing that readiness should be a focus of future research. Another useful future direction will involve testing what happens when researchers leave the clinics and clinic supervisors are left on their own to both train and supervise new clinicians who join their

clinic. A project of that kind would be strengthened by the development of training fidelity measures to complement assessment of supervision fidelity.

A limitation of our study design was that it did not include a usual care condition; comparison with usual care could provide information on the benefits of STEPs under alterative supervisory models relative to current clinic services and would thus be very helpful to include in future research. Finally, we tested sustainability over a period of about 3 years following transition to internal clinic staff supervision. It would be useful, if possible in the future, to monitor fidelity and outcomes over even more extended periods to provide increasingly robust tests of long-term sustainability. With that said, we should note that conducting research across multiple years in service-oriented community clinics imposes certain reality constraints, some of which can produce study limitations. One example relates to staff turnover in such settings, which has been documented in previous studies at levels as high as 25–50% per year (e.g., Woltmann et al., 2008). Significant turnover in the clinics participating in the present project meant that the clinicians in our Sustain group during Phase 2 did not include all the Sustain clinicians who had been included in Phase 1, and this would have been increasingly true had the study duration been extended.

The community clinic effectiveness context of our study, whereas it did pose certain challenges, was a strength in relation to the clinical representativeness of our findings. Arguably a test of the sustainability of empirically supported treatment in community clinics could hardly be based elsewhere. Thus, the study was located entirely within practicing mental health service clinics, in which youths referred through normal community pathways (without any advertising or recruiting) were treated by practicing clinicians in the context of their everyday work under usual employment conditions—including, for example, clinic productivity requirements, large nonstudy caseloads, substantial administrative paperwork, and other significant time-management challenges. The workplace context of the study speaks to the ecological validity of findings, suggesting that they may be relevant to youth psychotherapy because it is genuinely practiced in everyday clinical care.

As noted in the introduction, this study is one part of a broader effort by many investigators to understand what is needed to make implementation of empirically supported interventions sustainable in service settings (see, e.g., Aarons et al., 2011; Glasgow et al., 2006; Simpson & Flynn, 2007). Some of this work has focused on external factors such as financing (e.g., Stewart et al., 2016) and some on internal organizational factors such as the culture and climate of service agencies (e.g., Glisson & Williams, 2015). As a complement to these important lines of research, it is useful to examine what those who develop, test, and implement treatments can do in concert with clinicians and clinical administrators to facilitate increasingly independent, sustained use of empirically supported interventions within clinical service organizations. The present study is one step toward that objective.

References

Aarons, G. A., Glisson, C., Green, P. D., Hoagwood, K., Kelleher, K. J., Landsverk, J. A., . . . the Research Network on Youth Mental Health. (2012). The organizational social context of mental health services and clinician attitudes toward evidence-based practice: A United States national study. *Implementation Science, 7,* 56. http://dx.doi.org/10.1186/1748-5908-7-56

Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health, 38,* 4–23. http://dx.doi.org/10.1007/s10488-010-0327-7

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont, Research Center for Youths, Youth and Families.

Bearman, S. K., Herren, J., & Weisz, J. R. (2012). *Therapy integrity in evidence based interventions: Observational coding system, coding manual*. Austin, TX: 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, 44,* 293–307. http://dx.doi.org/10.1007/s10488-016-0723-8

Chamberlain, P., Roberts, R., Jones, H., Marsenich, L., Sosna, T., & Price, J. M. (2012). Three collaborative models for scaling up evidence-based practices. *Administration and Policy in Mental Health, 39,* 278–290. http://dx.doi.org/10.1007/s10488-011-0349-9

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,* 13–25. http://dx.doi.org/10.1037/ccp0000133

Chorpita, B. F., Daleiden, E. L., & Weisz, J. R. (2005). Identifying and selecting the common elements of evidence based interventions: A distillation and matching model. *Mental Health Services Research, 7,* 5–20. http://dx.doi.org/10.1007/s11020-005-1962-6

Chorpita, B. F., Reise, S., Weisz, J. R., Grubbs, K., Becker, K. D., Krull, J. L., & the Research Network on Youth Mental Health. (2010). Evaluation of the Brief Problem Checklist: Child and caregiver interviews to measure clinical progress. *Journal of Consulting and Clinical Psychology, 78,* 526–536. http://dx.doi.org/10.1037/a0019602

Chorpita, B. F., & Weisz, J. R. (2009). *Modular Approach to Therapy for Children with Anxiety, Depression, Trauma, or Conduct Problems (MATCH-ADTC)*. Satellite Beach, FL: PracticeWise, LLC.

Chorpita, B. F., Weisz, J. R., Daleiden, E. L., Schoenwald, S. K., Palinkas, L. A., Miranda, J., . . . the 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,* 999–1009. http://dx.doi.org/10.1037/a0034200

Dorsey, S., Lyon, A. R., Pullmann, M. D., Jungbluth, N., Berliner, L., & Beidas, R. (2017). Behavioral rehearsal for analogue fidelity: Feasibility in a state-funded children's mental health initiative. *Administration and Policy in Mental Health, 44,* 395–404. http://dx.doi.org/10.1007/s10488-016-0727-4

Ebesutani, C., Bernstein, A., Martinez, J. I., Chorpita, B. F., & Weisz, J. R. (2011). The youth self report: Applicability and validity across younger and older youths. *Journal of Clinical Child and Adolescent Psychology, 40,* 338–346. http://dx.doi.org/10.1080/15374416.2011.546041

Edmunds, J. M., Kendall, P. C., Ringle, V. A., Read, K. L., Brodman, D. M., Pimental, S. S., & Beidas, R. S. (2013). An examination of behavioral rehearsal during consultation as a predictor of training outcomes. *Administration and Policy in Mental Health, 40,* 456–466. http://dx.doi.org/10.1007/s10488-013-0490-

Enders, C. K., Keller, B. T., & Levy, R. (2017). A chained equations imputation approach for multi-level data with categorical and continuous variables. *Psychological Methods,* Advance online publication. http://dx.doi.org/10.1037/met0000148

Forgatch, M. S., & DeGarmo, D. S. (2011). Sustaining fidelity following the nationwide PMTO™ implementation in Norway. *Prevention Science, 12,* 235–246. http://dx.doi.org/10.1007/s11121-011-0225-6

Glasgow, R. E., Klesges, L. M., Dzewaltowski, D. A., Estabrooks, P. A., & Vogt, T. M. (2006). Evaluating the impact of health promotion programs: Using the RE-AIM framework to form summary measures for decision making involving complex issues. *Health Education Research, 21,* 688–694. http://dx.doi.org/10.1093/her/cyl081

Glisson, C., & Williams, N. J. (2015). Assessing and changing organizational social contexts for effective mental health services. *Annual Review of Public Health, 36,* 507–523.

Greif, R., Becker, C. B., & Hildebrandt, T. (2015). Reducing eating disorder risk factors: A pilot effectiveness trial of a train-the-trainer approach to dissemination and implementation. *International Journal of Eating Disorders, 48,* 1122–1131. http://dx.doi.org/10.1002/eat.22442

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New York, NY: Wiley.

IBM Corp. (2016). *IBM SPSS statistics for windows* (Version 24.0). Armonk, NY: Author.

Institute of Medicine, National Academy of Sciences. (2015). *Psychosocial interventions for mental and substance use disorders: A framework for establishing evidence-based standards*. Washington, DC: National Academies Press.

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science, 6,* 21–37. http://dx.doi.org/10.1177/1745691610393527

Keller, B. T., & Enders, C. K. (2017). *Blimp user's manual* (Version 1.0). Retrieved from http://www.appliedmissingdata.com/blimpuserguide-5.pdf

Kendall, P. C., & Beidas, R. (2007). Smoothing the trail for dissemination of evidence-based practices for youth: Flexibility within fidelity. *Professional Psychology, Research and Practice, 38,* 13–20. http://dx.doi.org/10.1037/0735-7028.38.1.13

Lai, M. H. C., & Kwok, O. (2015). Examining the rule of thumb of not using multilevel modeling: The "design effect smaller than two" rule. *Journal of Experimental Education, 83,* 423–438. http://dx.doi.org/10.1080/00220973.2014.907229

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. A. (2012). A "SMART" design for building individualized treatment sequences. *Annual Review of Clinical Psychology, 8,* 21–48. http://dx.doi.org/10.1146/annurev-clinpsy-032511-143152

Nock, M. K., Goldman, J. L., Wang, Y., Albano, A. M., & Jellinek, M. S. (2004). From science to practice: The flexible use of evidence-based treatments in clinical settings. *Journal of the American Academy of Child & Adolescent Psychiatry, 43,* 777–780. http://dx.doi.org/10.1097/01.chi.0000120023.14101.58

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. Chicago, IL: Scientific Software International.

Schoenwald, S. K., Chapman, J. E., Kelleher, K., Hoagwood, K. E., Landsverk, J., Stevens, J., . . . the Research Network on Youth Mental Health. (2008). A survey of the infrastructure for children's mental health services: Implications for the implementation of empirically supported treatments (ESTs). *Administration and Policy in Mental Health, 35,* 84–97. http://dx.doi.org/10.1007/s10488-007-0147-6

Schoenwald, S. K., Kelleher, K., Weisz, J. R., & the 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,* 66–72. http://dx.doi.org/10.1007/s10488-007-0160-9

Simpson, D. D., & Flynn, P. M. (2007). Moving innovations into treatment: A stage-based approach to program change. *Journal of Substance Abuse Treatment, 33,* 111–120. http://dx.doi.org/10.1016/j.jsat.2006.12.023

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal design plus empirical evidence: Documentation for the "Optimal Design" software Version 3.0*. Retrieved from http://hlmsoft.net/od/od-manual-20111016-v300.pdf

Stewart, R. E., Adams, D. R., Mandell, D. S., Hadley, T. R., Evans, A. C., Rubin, R., . . . Beidas, R. S. (2016). The perfect storm: Collision of the business of mental health and the implementation of evidence-based practices. *Psychiatric Services, 67,* 159–161. http://dx.doi.org/10.1176/appi.ps.201500392

Tommeraas, T., & Ogden, T. (2017). Is there a scale-up penalty? Testing behavioral change in the scaling up of Parent Management Training in Norway. *Administration and Policy in Mental Health, 44,* 203–216. http://dx.doi.org/10.1007/s10488-015-0712-3

Weisz, J., Bearman, S. K., Santucci, L. C., & Jensen-Doss, A. (2017). Initial test of a principle-guided approach to transdiagnostic psychotherapy with children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 46,* 44–58. http://dx.doi.org/10.1080/15374416.2016.1163708

Weisz, J. R., & Chorpita, B. F. (2011). Mod squad for youth psychotherapy: Restructuring evidence-based treatment for clinical practice. In P. C. Kendall (Ed.), *Child and adolescent therapy: Cognitive-behavioral procedures* (4th ed., pp. 379–397). New York, NY: Guilford Press.

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,* 369–380. http://dx.doi.org/10.1037/a0023307

Weisz, J. R., Chorpita, B. F., Palinkas, L. A., Schoenwald, S. K., Miranda, J., Bearman, S. K., . . . the 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,* 274–282. http://dx.doi.org/10.1001/archgenpsychiatry.2011.147

Weisz, J. R., Kuppens, S., Eckshtain, D., Ugueto, A. M., Hawley, K. M., & Jensen-Doss, A. (2013). Performance of evidence-based youth psychotherapies compared with usual clinical care: A multilevel meta-analysis. *JAMA Psychiatry, 70,* 750–761. http://dx.doi.org/10.1001/jamapsychiatry.2013.1176

Weisz, J. R., Ng, M. Y., & Bearman, S. K. (2014). Odd couple? Re-envisioning the relation between science and practice in the dissemination-implementation era. *Clinical Psychological Science, 2,* 58–74. http://dx.doi.org/10.1177/2167702613501307

Woltmann, E. M., Whitley, R., McHugh, G. J., Brunette, M., Torrey, W. C., Coots, L., . . . Drake, R. E. (2008). The role of staff turnover in the implementation of evidence-based practices in mental health care. *Psychiatric Services, 59,* 732–737. http://dx.doi.org/10.1176/appi.ps.59.7.732

Yeh, M., & Weisz, J. R. (2001). Why are we here at the clinic? Parent-child (dis)agreement on referral problems at outpatient treatment entry. *Journal of Consulting and Clinical Psychology, 69,* 1018–1025. http://dx.doi.org/10.1037/0022-006X.69.6.1018

Received August 16, 2017
Revision received May 28, 2018
Accepted June 8, 2018 ■