Available online at www.sciencedirect.com

ScienceDirect

Behavior Therapy 50 (2019) 225–240

Behavior Therapy

www.elsevier.com/locate/bt

Durability of Cognitive Behavioral Therapy Effects for Youth and Adolescents With Anxiety, Depression, or Traumatic Stress: A Meta-Analysis on Long-Term Follow-Ups

Leslie R. Rith-Najarian
Bita Mesri
Alayna L. Park
Michael Sun
Denise A. Chavira
Bruce F. Chorpita
University of California, Los Angeles

Cognitive behavioral therapies (CBT) for youth with anxiety, traumatic stress, and depression have demonstrated strong effects in individual studies and meta-analyses. Relatively more attention has been given to posttreatment effects, though, and assessment of follow-up effects has been limited at the meta-analytic level. The current meta-analysis aimed to (a) examine the effects of youth CBT at posttreatment, 1-month, 3-month, 6-month, 1-year, and long-term (2+ years) follow-up as well as (b) identify research-related variables (e.g., measure respondent type) that relate to effects. Using a random effects model across 110 child and adolescent CBT groups, within-group effect sizes were large at posttreatment (g = 1.24) and from 1-month through long-term follow-up (g = 1.23–1.82), and effect sizes did not significantly differ by treatment target (i.e., anxiety, traumatic stress, depression). However, availability of outcome data for effect sizes diminished across later follow-up assessments. Moreover, effect sizes were significantly associated with outcome respondent type across assessment timing, with outcome measures from caregiver and youth respondents associated with smaller effect sizes (B = -0.97, p < 0.001) relative to outcome measures that were evaluator-reported. Results provide initial support for the durability of treatment effects for youth CBTs and highlight the importance of some confounding variables. Implications for improving treatment research standards and prioritizing assessment of long-term follow-up assessment are discussed.

Keywords: cognitive behavioral therapy; youth; anxiety; depression; traumatic stress

Address correspondence to Leslie Rith-Najarian, Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Box 951563, Los Angeles, CA 90095; e-mail: leslierrn@ucla.edu

ANXIETY AND DEPRESSIVE disorders are the most prevalent psychological disorders in children and adolescents, across global surveys (Polanczyk, Salum, Sugaya, Caye, & Rohde, 2015). Having anxiety or depression in childhood or adolescence is predictive of greater impairment and worse functioning in adulthood (e.g., Copeland, Angold, Shanahan, & Costello, 2014; Dunn & Goodyer, 2006), underscoring the importance of early intervention for internalizing disorders. Anxiety, traumatic stress, and depression have common symptoms and mechanisms, such as rumination and worry (Olatunji, Naragon-Gainey, & Wolitzky-Taylor, 2013; Verstraeten, Bijttebier, Vasey, & Raes, 2011) and behavioral avoidance (Aldao, Nolen-Hoeksema, & Schweizer, 2010; Briggs & Price, 2009). Accordingly, their respective youth evidence-based treatments share many cognitive and behavioral skills (Chorpita & Daleiden,

2009). Compared with other modalities, cognitive behavioral therapy (CBT) is the treatment approach with the most well-established support for improving symptoms in youth with anxiety (e.g., Higa-McMillan, Francis, Rith-Najarian, & Chorpita, 2016), trauma (e.g., Dorsey et al., 2017), and depression (e.g., Weersing, Jeffreys, Do, Schwartz, & Bolano, 2017). According to multiple systematic reviews and meta-analyses examining youth CBTs specifically, the majority of CBTs targeting anxiety, traumatic stress, and depression in children and adolescents show moderate to large effects (Gutermann, Schwartzkopff, & Steil, 2017; James, James, Cowdrey, Soler, & Choke, 2015; Kendall & Peterman, 2015; Spielmans, Pasek, & McFall, 2007).

Current standards typically deem a treatment successful if the treatment has demonstrated evidence of significant improvement by posttreatment, relative to a control (e.g., Kazdin, 2008). One strength of using posttreatment assessment findings as a standard of evidence is that such data are readily available and interpretable and provide a reasonable indication that improvement from baseline status has been achieved. However, a limitation of this approach is that "posttreatment" timing can be confounded with "still in treatment" as such assessments often occur within a week of the last treatment session. Therefore, posttreatment assessment cannot provide information about the durability of a treatment's effects. Unfortunately, when it has come to establishing standards of evidence, relatively less attention has been devoted to examining the durability of treatment effects after a youth is no longer attending regular therapy.

Important insights into the durability of youth CBT effects could be gained by focusing more on assessment of follow-up effects. Longitudinal studies have provided evidence that youth internalizing symptoms persist into adulthood when left untreated (e.g., Copeland et al., 2014; Dunn & Goodyer, 2006), but there is still more to understand about how symptoms unfold in the long term for those youth who *are* treated. If youth maintain significant improvement in outcomes after treatment ends, then the lack of relapse can be captured in durable effect sizes—or effect sizes that maintain their magnitude—across follow-up assessments. Many researchers have discussed why durability of treatment effects is important, especially for youth CBTs. First, assessment of follow-up effects can convey whether the changes made by youth during treatment are internalized in a way that lasts over a significant period of time. Such lasting improvement in functioning can allow for critical developmental opportunities, which otherwise might have been missed due to cascading impairments and negative consequences (Burt & Paysnick, 2012; Stewart, Stavness, King, Antle, & Law, 2006). Second, although CBTs have been found to be relatively more time- and cost-effective than usual care, they can still be expensive and time-intensive; for example, one study found CBT to cost an average of $3,221.34 per child with treatment lasting 24 weeks on average (Weisz et al., 2009). Thus, the effects of treatment should be lasting so that additional resources are not required for future treatment. Third, despite the strong evidence for youth CBTs' posttreatment effectiveness, relapse rates can be as high as a third to a half of treated youth (e.g., Ginsburg et al., 2014; TADS Team et al., 2009). Thus, it is important to examine CBT effect sizes at an aggregate level at both posttreatment as well as follow-up, allowing for evaluation of overall changes across time. The findings of such a meta-analysis would also inform the gap in evidence about youth CBT's long-term effects, thus potentially providing new information that can be important for patients, providers, and other stakeholders.

This current review compared treatment effects over time from posttreatment through long-term follow-up in a sample of articles on youth CBT for anxiety, traumatic stress, and depression. Individual studies have demonstrated support for long-term CBT gains among youth with internalizing disorders, especially anxiety (e.g., Kendall & Peterman, 2015); beyond individual studies, a meta-analytic review allows for a more complete assessment of the current state of all available evidence. Examination of youth CBT follow-up effects has been conducted in previous meta-analyses, and some support has been reported for long-term durability of treatment effects. However, of those meta-analyses reporting "long-term" follow-up effects, many of the included follow-up assessments occurred only months posttreatment (e.g., Gutermann et al., 2017; James et al., 2015). Additionally, the timing of follow-up assessments across studies varies widely, and consequently meta-analyses that aim to assess youth psychotherapy follow-up effects must purposefully address this variability in their methodology. One approach has been presenting follow-up effects for each study individually rather than in aggregate (e.g., Dorsey et al., 2017; Higa-McMillan et al., 2016). Another approach has been selecting a single follow-up interval on which to focus (e.g., reviewing studies with 1-year follow-up results; Manassis et al., 2014). A third approach involves collapsing effect sizes across all follow-up intervals ranging from months to years (e.g., James et al., 2015; Weisz et al., 2013, 2017). Unfortunately, these approaches limit the interpretation of the magnitude of follow-

up effect sizes at different lengths of follow-up timing. In contrast, a meta-analysis on youth depression treatments that did consider timing of follow-up assessment found that follow-up treatment effect size was negatively correlated with length of time until follow-up (Weisz, McCarty, & Valeri, 2006). More attention to later follow-up (i.e., 1 year or longer) is needed to understand the durability of our existing youth CBTs. Another challenge for meta-analysis of youth CBTs is that each RCT can use different outcome measures from different respondents—youth self-report, parent-report, diagnostic interview, etc.—and thus the resulting effect sizes may be capturing slightly different constructs. Moreover, different respondent measures within studies are not always in agreement with each other (De Los Reyes et al., 2015). To more accurately represent youth CBT study findings, meta-analysis can include effect sizes for all primary symptom outcome measures, while examining if effects vary by respondent, as well as by other research-related variables (e.g., Weisz et al., 2017). Overall, this review intended to more completely assess the current state of evidence for youth CBT long-term follow-up effects for internalizing symptom outcomes.

For our first aim, we examined whether effect sizes were durable through 1-year follow-up and longer. To address this aim, we: (a) investigated how frequently effect sizes were assessed across follow-up timepoints, (b) compared aggregated effect sizes at: posttreatment, 1-month follow, 3-month follow-up, 6-month follow-up, 1-year follow-up, and long-term follow-up, and (c) examined effect sizes over these time periods by groups for each treatment target—anxiety, traumatic stress, depression. In the presence of durable effects, our second aim was to assess how research study variables might contribute to effects. To address this aim, we compared aggregated effect sizes by respondent types—youth, caregiver, and evaluator—across all the assessment intervals. We also conducted post hoc analyses that compared the posttreatment and follow-up effect sizes of youth CBTs that did and did not have 1 year or longer follow-up. Examination of such results can demonstrate whether effect sizes of youth CBTs are consistent across assessments regardless of research variables, or if effect sizes from certain respondents or certain types of studies show relatively more or less durability.

Method

IDENTIFICATION OF STUDIES

Eligible articles were identified from the PracticeWise Evidence-Based Services (PWEBS) literature database, a database of youth mental health treatment research articles (PracticeWise, 2017). All interventions included in the database were tested in studies using youth samples with a mean age of 19 years old or younger. The PWEBS database is updated twice annually by incorporating new articles identified from online databases (e.g., PsycINFO) and personal nominations submitted by trained consultants and other individuals in the professional community (Chorpita et al., 2011). The database was accessed in February 2017 under a research agreement with PracticeWise, LLC, at which time PWEBS included 987 youth treatment articles published from 1965 through 2017.[1]

Inclusion eligibility of youth CBT articles was determined based on information coded already in PWEBS. To be included in this review, a study must have tested at least one CBT treatment group. The CBT intervention could be: provided to young children (ages 6 and under), middle youth (ages 7–12), adolescents (ages 13 and older), or a combination of these age groups; in individual sessions or group sessions; and could be combined with medication or parent/teacher training. The remaining inclusion criteria were: (a) use of random assignment; (b) treatment of anxiety (including obsessive-compulsive symptoms), traumatic stress, or depression; and (c) inclusion of at least one outcome measure to assess a target symptom area (e.g., assessing functional outcome only, such as school attendance, was insufficient). We did not include open trials as their purpose is typically one of initial feasibility rather than effectiveness, and thus follow-up assessments (especially longer-term ones) are often not conducted.

Data published in the original RCTs was supplemented in two ways. First, to identify additional follow-up data that may have been published separately for any of the included CBTs, a separate literature search was conducted for any papers published prior to March 10, 2017. Online searches on PsycINFO, Google Scholar, PubMed and Web of Science were refined by (a) publications that cited the original study article, (b) publications authored by one of the original study authors, and (c) the terms "long-term," "follow-up," "longitudinal," or "outcomes." Abstracts were examined to confirm if an article was indeed a follow-up for the original study. Then, the same inclusion criteria were applied. During this literature search, we also manually reviewed these search results to confirm that no RCT articles not yet in our sample met our inclusion criteria. Second, we contacted authors of any articles that reported

[1] Not all youth treatment articles published in or recently prior to February 2017 were necessary included in our eligibility review, due to lag time between article identification by PracticeWise staff and the completion of full coding of the article into the PWEBS database.

incomplete data, and heard back from seven of eight authors.

Articles were then excluded during the coding process based on reviewing the full content of the original study article and any separately published follow-up articles. Exclusion criteria were: (a) no follow-up assessment conducted; (b) no report of necessary statistics (i.e., *M*, *SD*, *N*) for calculating effect sizes at both posttreatment assessment and follow-up assessment (unless these statistics were provided when we contacted the authors); (c) a solely subclinical youth sample; and (d) unconventional delivery design (i.e., single-day interventions, electronically delivered interventions, parent-administered interventions, bibliotherapy). We included only clinician-delivered, multisession youth CBTs in order to reduce heterogeneity due to delivery-related variables, which were not related to our primary aims and research questions. There were no restrictions based on publication year, geography, or culture. A flowchart of the inclusion and exclusion progress is provided in Figure 1.

Given our aim to assess effect durability, it is important to consider if CBT groups in our sample

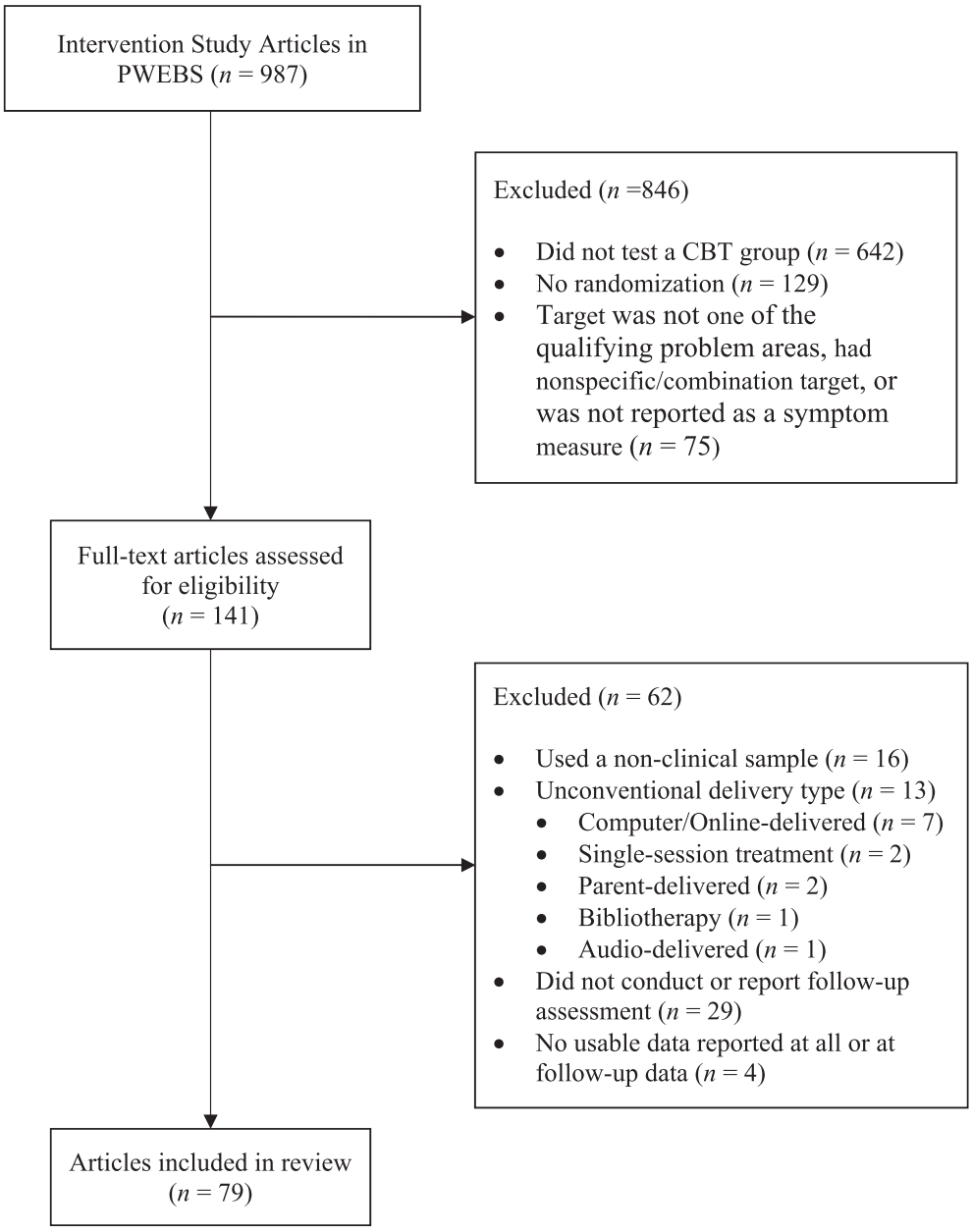

FIGURE 1 Article exclusion flowchart, outlining procedure to select final sample of articles meeting inclusion criteria.

had "winning" effects already at posttreatment. However, our inclusion criteria did not require posttreatment superiority against a comparison group, as not all study designs are the same, making "winning" at posttreatment difficult to define. For example, some RCTs were designed as noninferiority trials to demonstrate that one treatment group did not perform worse than another treatment with previously established efficacy. Additionally, some studies compared different types of youth CBTs (e.g., group sessions versus individual sessions) without including a third inactive comparison group; in such cases, a nonsignificant posttreatment difference between groups is not clearly interpreted as a "win" or a "loss" for either group. Therefore, posttreatment winning status was not used as one of the inclusion criteria so that we did not inadvertently exclude potentially strong youth CBT groups from our sample. We did, however, code information for each group to define "winning," when possible (outlined below).

DATA EXTRACTION AND CODING

Youth CBT characteristics (e.g., delivery type, primary symptom target) and study information (e.g., sample demographics) were extracted directly from PWEBS. These variables were already collected using the PracticeWise coding system (PracticeWise, 2012), a coding procedure that has produced excellent interrater agreement, as reported elsewhere (Cohen's κ ranging from .84 to 1.0; e.g., Chorpita & Daleiden, 2009; Chorpita et al., 2011). Then, study outcome data were aggregated for each CBT group from its original article and any corresponding follow-up study articles. If an article had multiple eligible youth CBT treatment groups tested, then we coded information for each group separately.

Six article coders were trained by the principal investigator to use an outcome coding manual and all achieved sufficient interrater reliability after coding practice articles. Coders independently coded subsets of articles and the principal investigator coded all study articles. Discrepancies were addressed in weekly meetings between the principal investigator and the respective second coder. Interrater reliability was calculated with Cohen's kappa and intraclass correlations using two-way random models assessing for absolute agreement.

Eligible Measures

Coders identified all "target symptom" measures prior to extracting data for each article. Target symptoms of each CBT study were determined by identifying the symptom area that was common to: (a) sample inclusion (e.g., inclusion of youth with anxiety); (b) treatment target (e.g., treatment protocol designed to address anxiety symptoms); and (c) reported outcome measures (e.g., anxiety measures). For example, if a study included youth with comorbid depression and trauma, but the treatment was explicitly designed to reduce depression symptoms, and depression outcome measures were primarily reported, then the target symptom outcome domain would be depression. If a measure collapsed different domains into one score (e.g., anxious-depressive symptoms, or symptoms combined with functional impairment), then we considered its scores as capturing more than the target symptoms, and the measure was not coded.[2] Additionally, unstandardized experimenter-created or experimenter-defined measures were excluded because they have been found to significantly inflate effect sizes (e.g., Cheung & Slavin, 2016). Continuous target symptom measures could derive from a diagnostic assessment (e.g., severity, total number of symptoms). We also coded diagnostic case data (e.g., frequency of participants meeting criteria for a disorder or with scores above a cut-off). Finally, to be included in effect size calculation for this review, the measure had to be assessed and reported at pretreatment, posttreatment, and follow-up. If outcome data were missing for a measure at any one of those time points (e.g., a measure assessed only at pretreatment and posttreatment), then the extracted data from that measure were not included in effect sizes for that youth CBT.

Measure Outcome Data

For each eligible measure, the mean, standard deviation, and sample size was coded for each assessment period; if such data were missing, then we reverse calculated them from available alternative statistics (e.g, *SE*, effect size with reported formula), if reported. If the data reported in an article were presented in different ways (e.g., intent-to-treat, raw), we coded the data type that was most consistently presented across assessment periods. In cases when both intent-to-treat and completer data were reported at all assessment periods, we defaulted to collecting the completer data, as it was more frequently the type of data reported across our article sample. Of the articles in this review sample, 49 (62%) had completer or raw data, and the remaining 30 had intent-to-treat or imputed data. Interrater reliability for collected measure outcome data was excellent, using intraclass correlation with average measures analysis (means *ICC* = 0.97; standard deviations *ICC* = 0.97; sample size *ICC* = 0.99).

[2] Note: None of these cases resulted in exclusion of an article, as all articles had at least one other eligible measure.

Assessment Period

Options to code included baseline, posttreatment (immediately or briefly after active treatment has ended), or follow-up (after posttreatment assessment). To be defined as a "follow-up" assessment, there was no minimum amount of time required, but the treatment did need to be finished in terms of content and therapeutic contact (with the exception of one or two booster sessions, n = 22 groups). Assessment period type was coded with excellent interrater reliability (Cohen's κ = .97). The duration of time between posttreatment and each follow-up was recorded in months, with excellent interrater reliability (*ICC* = .98). Assessment interval categories were created post hoc: posttreatment, 1-month follow-up, 3-month follow-up, 6-month follow-up, 12-month follow-up, or long-term follow-up (i.e., 2 or more years). If a follow-up occurred at an intermediate time, the effect size was assigned to the nearest follow-up interval; for example, 4-month follow-up data would be used for a 3-month follow-up effect size.

Outcome Respondent Type

For every measure, we also coded who completed the measure. There were three respondent types to code for continuous measures: youth (i.e., self-report measure), caregiver (parent or other adult responsible for youth's care), or evaluator (individual who completed measure or structured interview as part of research study, not otherwise part of youth's life). Measures from other reporters (e.g., teacher, provider) were not coded due to relatively lower frequency across articles in the sample.[3] Some CBTs were assessed by measures from multiple reporters and some CBTs were assessed by measures only from a single reporter type. Measure respondent type was coded with excellent inter-rater reliability (Cohen's κ = 0.96). For the evaluator respondent type, there was also clinical improvement data reported by studies as frequencies rather than continuous data. These data may have been based on a single evaluator measure or a combination, with each study providing its own definition of an "improved case." If clinically improved case data were represented in multiple ways for assessing a CBT group, we selected the data that were reported for the most assessment intervals, and then prioritized data type in the order of: (a) diagnostic status, (b) falling below measure severity or cut-off scores, and (c) achieving clinically significant change (as defined by measures of improvement or degree of decrease in symptom measures).

[3] Note: The decision not to code measures from other reporters did not result in exclusion of any articles, as the few articles that were affected still had measures from other respondent types.

"Winning" Status

Reported statistical results were coded for each RCT to determine which youth CBT groups did win or did not win against a comparison group within their respective study. A CBT group was credited with a "win" if there was a significant between-group result demonstrated that favored more improvement for youth in that CBT group. Assessment of the between-group effect required report of both the statistical test value and its significance (i.e., p-value) specifically for the CBT group. Significant between-group results could be demonstrated by a statistically significant posttreatment group effect or a statistically significant group-by-time interaction. If between-group differences were found at baseline, an appropriate statistical test was required to address this (e.g., ANCOVA). If two CBT groups within the same study separately demonstrated superiority over another comparison group, both CBT groups were credited with a "win." Winning status was coded with excellent interrater reliability (Cohen's κ = 0.90). We marked CBT groups as "undetermined" if their study tested only active comparison groups (e.g., two CBT groups), and this group failed to produce a "win."

EFFECT SIZE CALCULATION

We calculated within-subject pre- to posttreatment and pre- to follow-up effect sizes, rather than between-group effect sizes. We did so for three reasons. First, many comparison groups are eventually provided treatment (i.e., post-waitlist) and so the majority of RCTs do not have a true comparison group to assess at follow-up, especially at longer term follow-ups. Second, our article sample included various types of comparison groups (e.g., waitlist, active controls), which would likely affect the resulting effect sizes, as has been found in other research (e.g., Weisz et al., 2017). We also did not want to shrink our coverage of youth CBTs or sample size by excluding articles due to comparison group type. Third, because we were interested in how effect sizes for CBT groups change across longer follow-up intervals, within-subject effect sizes are more representative of durable effects without picking up on changes by comparison group.

Effect sizes for continuous measures were calculated using Cohen's $d_{average}$ with Hedges' correction (Hedges, 1981), with the pretreatment statistics in M_2 and SD_2, and the respective post or follow-up statistics in M_1 and SD_1.

$$Hedges' \; g_{average} = \frac{M_1 - M_2}{\frac{SD_1 + SD_2}{2}} \times \left(1 - \frac{3}{4(n_2 - 1) - 1}\right)$$

We used the $d_{average}$ as opposed to the $d_{repeatedmeasure}$ here. The $d_{repeatedmeasure}$ is a more conservative metric, but it requires the correlation coefficient of pretreatment and posttreatment means, which most studies in our sample did not report. Although we could have substituted an arbitrary standard correlation coefficient (e.g., $r = .70$) for all studies missing this statistic, we decided that using $d_{repeatedmeasure}$ would not provide additive benefit, as it may have biased our results for or against those few studies that did report a correlation.

Effect sizes for frequency data (e.g., number of cases free of diagnosis) were calculated by transforming odds ratios using the formula below (Hasselblad & Hedges, 1995). Respective posttreatment/follow-up data (number of cases with improvement or no diagnosis, over the sample size) were provided in the numerator, with the pretreatment data (number of cases without diagnosis or elevated symptoms, over the sample size) provided in the denominator.

$$\text{converted } OR_d = \ln\left(\frac{A/B}{C/D}\right) \times \frac{\sqrt{3}}{\pi}$$

Because odds ratios cannot have a value of 0 in the B, C, or D part of the formula, any time an effect size calculation would encounter such data, 0.5 was added to A, B, C, and D, as recommended by Agresti (1996). Once converted into a standardized mean difference, the same Hedges' correction was also applied to these effect sizes.

To adjust for heterogeneity of variance, we weighted each effect size by its inverse variance (Hedges & Olkin, 1985), giving more weight to those studies that produce a more precise effect size, i.e., an effect size with smaller standard error. Each average effect size reported in the results is the average of these weighted effect sizes.

For each CBT group, one effect size was first calculated for each measure at each of its respective assessment intervals. Then, all target symptom measures' effect sizes were averaged for each assessment interval. If a measure's mean results were reported by subscales (rather than a single total score mean), these subscale results were first collapsed into a single measure effect size, before being averaged with other measure effect sizes.

DATA ANALYSES

Heterogeneity and Random Effects Model

Given that the youth CBTs in our sample are characteristically heterogeneous (e.g., treatment target, age, delivery format), we selected a random effects model. A random effects model is more appropriate for interpreting effects that theoretically represent a relatively diverse "population" of CBTs, which in our case is a generalized context of youth CBTs for internalizing symptoms.

Effect Size and Meta-Regression Analyses

Once all the effect sizes had been calculated using the formulas above, we conducted analyses using Comprehensive Meta-Analysis software version 3 (CMA; Borenstein, Hedges, Higgins, & Rothstein, 2013) and SPSS Statistics 24.0 (IBM Corp, 2016). We used CMA to calculate the average weighted effect sizes with a random effects model. For meta-regressions examining between-group predictor variables only, we used CMA to conduct random effects, method of moments meta-regressions with a Knapp-Hartung adjustment. Given that our effect sizes were not normally distributed, the method of moments approach was the more robust option (over unrestricted or restricted maximum likelihood approaches) for estimating the true between-studies variance, as this approach relies on no assumptions about the random effects' distribution. Because the true population variance of the effect sizes for these youth CBTs was not known, the Knapp-Hartung adjustment was used rather than a z-distribution so that the estimation of between-group was based on the sample variance (Knapp & Hartung, 2003). For the meta-regressions that examined outcome respondent type as a predictor, we used an SPSS macro based on Hedges, Tipton, and Johnson's (2010) for meta-regression with robust variance estimation (RVE) approach. The RVE approach is more appropriate for a correlated effects model, which was necessary to use given that effect sizes from multiple measurement outcomes on the same individuals within studies were being predicted. Robust variance estimation adjusts standard errors of the effect sizes to account for dependence between them.

Power Analysis

Although meta-regression power analysis formulas (e.g., Hedges & Pigott, 2001) and corresponding software macros (e.g., Cafri, Kromrey, & Brannick, 2009) have been developed, proper procedures for the various types of meta-regressions, especially with complicated multi-level data, are still being debated (e.g., López-López, Marín-Martínez, Sánchez-Meca, Van den Noortgate, & Viechtbauer, 2014). Therefore, we used formulas and procedures for standard random effects power analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009) to calculate the minimum number of studies (k) necessary for each meta-regression to have adequate power ($p = .80$). This estimated power analysis approach has been used by other meta-analyses (e.g., Weisz et al., 2013). Power was

computed with the aim of detecting small effect sizes (d = .20) at an alpha level of .05. The selected formula for the variance of the summary effect (V_d) assumes a large amount of heterogeneity between the studies, and thus multiplies the within-study variance (0.076) by 2, and then divides the product by the average sample size of the included studies (26.6).

$$Power = .80<1-\Phi(1.96-\lambda) + \Phi(-1.96-\lambda)$$

$$\lambda = \frac{d}{\sqrt{V_d}} \qquad V_d = \frac{2^* 0.076}{k}$$

Results showed that detecting an effect size this small with adequate power would require at least 30 studies to be included in any given meta-analysis. Therefore, we did not report or interpret any meta-regression results that relied on $n_{\text{studies}} < 30$.

Assessment of Biases

We assessed two sources of publication bias —(a) availability of published studies versus nonpublished and (b) selective reporting by the published studies—as well as a third source of potential bias due to incomplete outcome data due to dropout. We addressed the first problem, also known as the file drawer problem (Rosenthal, 1979), by using a funnel plot (Torgerson, 2006), with standard error on the y-axis and posttreatment effect size on the x-axis. The funnel plot was followed by Egger's weighted regression test (Egger, Smith, Schneider, & Minder, 1997), and then by computing a fail-safe N (Rosenthal, 1979). We addressed the second problem, selective reporting bias, by running a meta-regression on posttreatment effect sizes with a binary "complete reporting" independent variable associated with effect size. A study was assigned a 1 on the "complete reporting" variable if all target symptoms measures listed in the methods were reported in the results; otherwise, the study was assigned a 0. Finally, we addressed the third problem, bias related to incomplete outcome data, by assigning a "data type" variable for each study, defined as the type of reported data on which the effect size was based—raw, completer, intent-to-treat, or imputed—and ran a meta-regression with this variable as a predictor of posttreatment and follow-up effect size.

Results

RESULTING SAMPLE

The sample that passed all inclusion and exclusion criteria (see Figure 1) included 79 articles and 18 follow-up articles, published between 1986 and 2017, covering 110 youth CBT groups. Treatment target by CBT group were: anxiety (n = 71), traumatic stress (n = 24), and depression (n = 15). Primary delivery formats were: individual sessions (with child-only or with family members; n = 59) and group sessions (n = 51). There were 74 youth CBT groups that met criteria for "winning" status at posttreatment, out of 79 groups that had determinable status based on posttreatment results. Treatments lasted an average of 80 days (SD = 30). Studies were designed to compare youth CBTs against: other CBTs only (n = 25); other CBTs plus inactive control groups (i.e., attentional control, waitlist, no treatment control; n = 18); other CBTs plus other active treatments (n = 6); other active treatments only (n = 11); other active treatments plus inactive control groups (i.e., waitlist, no treatment control; n = 19); treatment as usual (n = 3); waitlist groups only (n = 23); no treatment control groups only (n = 3); and attentional control groups with/without separate waitlist groups (n = 2). Of the 79 studies, 57% stated using masked assessments from baseline through postintervention (though more may have been masked without explicitly reporting it, while others still relied solely on self-report measures). Refer to Appendix A for more details on group characteristics, effect size forest plots, and review sample article references.

OVERALL EFFECT SIZES ACROSS ASSESSMENTS

The mean effect sizes were 1.24 at posttreatment (n = 110, SE = 0.06, 95% CI = [1.13, 1.35]), 1.23 at 1-month follow-up (n = 10, SE = 0.18, 95% CI = [0.88, 1.58]), 1.43 at 3-month follow-up (n = 52, SE = 0.09, 95% CI = [1.26, 1.60]), 1.57 at 6-month follow-up (n = 43, SE = 0.10, 95% CI = [1.37, 1.76]), and 1.65 at 1-year follow-up (n = 51, SE = 0.09, 95% CI = [1.47, 1.82]), and 1.82 at long-term follow-up (n = 12; SE = 0.18, 95% CI = [1.47, 2.16]) assessment periods.[4] It is important to note that the studies represented at each of the follow-up assessment times vary. The inconsistency of follow-up assessment timing across studies prevented us from analyzing the significant change in effect size over time using developed statistical procedures for meta-analysis of effect sizes at multiple time points (e.g., Musekiwa, Manda, Mwambi, & Chen, 2016). The issue with our dataset was the degree of missingness across intervals (i.e., only nine of the studies in our sample conducted assessments at three or more of the possible follow-up assessment times), which is not addressable with currently existing meta-analytic procedures. Nevertheless, to demonstrate

[4] Effect sizes were also calculated with studies – rather than groups – as the unit of analyses. The average effect sizes across assessment times changed minimally (-0.03 to 0.04). Therefore, unless otherwise noted, all analyses use groups as the unit of analyses, which provided larger overall sample size and more power.

the average change in effect sizes over time, effect sizes were also calculated from posttreatment to each respective follow-up assessment for each CBT group. These effect sizes are more demonstrative of *change* in effect size over time because they are not affected by the variable inclusion of different groups at different assessment time points. In this way, groups that had anomalous effect sizes already at posttreatment of groups could not over inflate or deflate the average follow-up effects at their respective follow-up intervals. These posttreatment to follow-up effect sizes are plotted for all youth CBT groups as well as for the subset of CBT groups that met criteria for posttreatment "winning" in Figure 2.

EFFECT SIZES BY TREATMENT TARGET

At posttreatment, the mean effect size was 1.75 (*SE* = 0.11, *95% CI* = [1.53, 1.97]) for anxiety-targeted groups, 1.31 (*SE* = 0.13, *95% CI* = [1.06, 1.56]) for trauma-targeted groups, and 1.67 (*SE* = 0.27, *95% CI* = [1.14, 2.20]) for the depression-targeted groups. See Table 1 for effect sizes by target across follow-up time. By using meta-regression with a robust variance estimation approach, all effect sizes across studies and follow-up assessments were considered at the same time. Treatment target was not a significant predictor of effect size (Tau^2 = 0.41, B = 0.004, p = .95). Using a method of moments meta-regression, effect size differences by treatment target were analyzed at each follow-up assessment, with anxiety as the reference group. Treatment target was not a significant predictor at posttreatment (R^2 analog = 0.00, F[2, 107] = 0.98, p = .38), 3-month follow-up: (R^2 analog = 0.00, F[2, 49] = 1.09, p = .34), or 6-month follow-up: (R^2 analog = 0.00, F[2, 40] = 1.24, p = .30). Treatment target was a significant predictor at 1-year follow-up (R^2 analog = 0.36, F[2, 48] = 3.56, p = .04), specifically

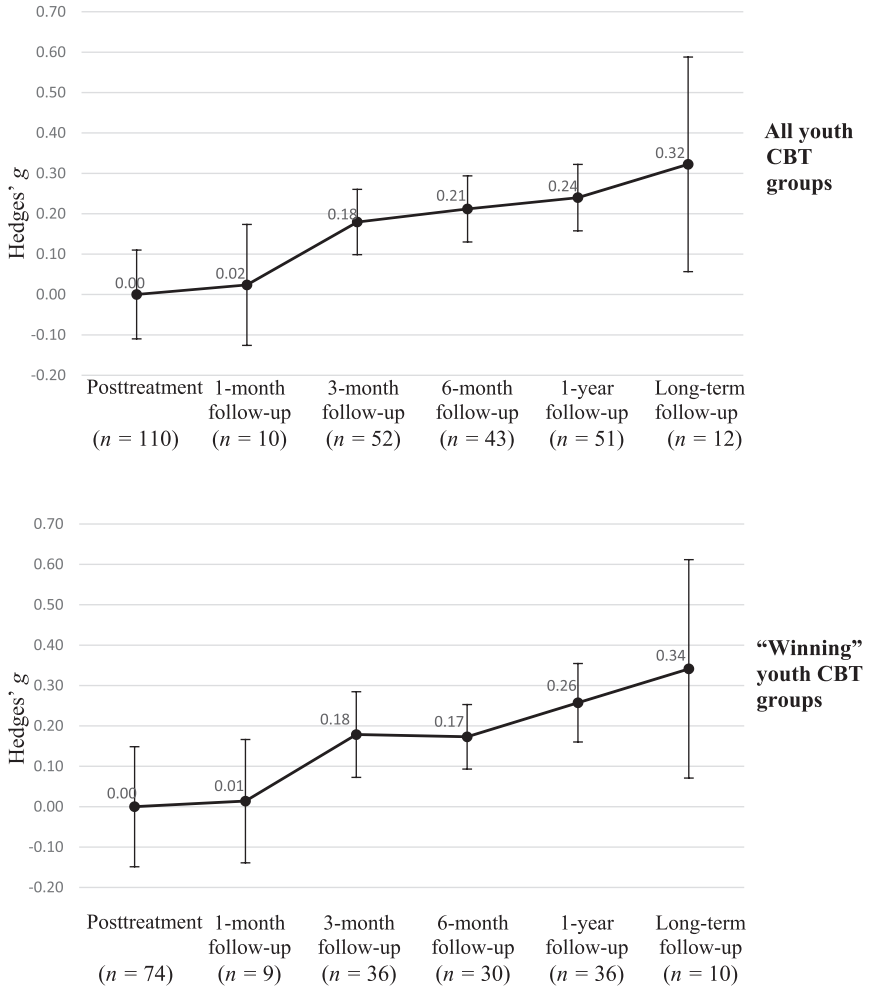

FIGURE 2 Effect sizes from posttreatment to follow-up assessments. Error bars represent 95% confidence intervals.

Table 1
Effect Sizes Across Assessment Timepoints, Treatment Targets, and Respondent Type

Target	Measure respondent	Post-Tx: *g* (*se*)	1-mo FU: *g* (*se*)	3-mo FU: *g* (*se*)	6-mo FU: *g* (*se*)	1-yr FU: *g* (*se*)	LT FU: *g* (*se*)
All targets	Measures from any respondent	1.24, (0.06) *n* = 110	1.23, (0.18) *n* = 10	1.43, (0.09) *n* = 52	1.57, (0.1) *n* = 43	1.65, (0.09) *n* = 51	1.82, (0.18) *n* = 12
	Child/ Parent	0.89, (0.05) *n* = 95	0.91, (0.18) *n* = 7	1.18, (0.07) *n* = 43	1.21, (0.08) *n* = 33	1.27, (0.09) *n* = 44	1.07, (0.11) *n* = 7
	Evaluator	1.97, (0.1) *n* = 84	2.15, (0.35) *n* = 6	1.99, (0.15) *n* = 32	2.22, (0.14) *n* = 35	2.29, (0.14) *n* = 45	2.58, (0.33) *n* = 10
Anxiety	All measures	1.75, (0.11) *n* = 71	2.00, (0.41) *n* = 4	1.61, (0.15) *n* = 32	1.97, (0.14) *n* = 26	2.02, (0.14) *n* = 38	2.79, (0.51) *n* = 7
Depression	All measures	1.31, (0.13) *n* = 24	0.58, (0.24) *n* = 1	1.37, (0.16) *n* = 15	1.93, (0.22) *n* = 9	1.96, (0.19) *n* = 5	2.16, (0.38) *n* = 2
Traumatic Stress	All measures	1.67, (0.27) *n* = 15	1.21, (0.22) *n* = 5	2.07, (0.39) *n* = 5	1.84, (0.4) *n* = 8	2.29, (0.34) *n* = 8	1.55, (0.4) *n* = 3

Note. Post-Tx = Posttreatment; 1-mo FU = one month follow-up; 3-mo FU = three month follow-up; 6-mo FU = six month follow-up; 1-yr FU = one year follow-up; LT FU = long-term (2+ years) follow-up.

with depression-targeted CBTs associated with larger effect sizes ($B = 0.55$, $t = 2.31$, $p = .03$) than anxiety-target CBTs. However, this result was no longer significant when studies were used as the unit of analyses ($p = .10$), as there were only 6 depression studies included in the 1-year follow-up time, one of which had a notably larger effect size for its two CBT groups at that follow-up time (respectively 0.94 and 1.09 standard deviations larger than the average effect size of the other six groups). For the two other follow-up assessment times, we deemed it inappropriate to conduct meta-regression, given the small number of studies per covariate at 1-month follow-up ($df = 2,7$) and at long-term follow-up ($df = 2,9$).

EFFECT SIZES BY OUTCOME RESPONDENT TYPE

For the following analyses, we collapsed the outcome respondent types into two categories that grouped together well conceptually: (a) youth- and caregiver-reported outcomes and (b) evaluator-reported outcomes (for continuous measures and frequency data). Given the variation of measure type included in each study, these broader categories were simultaneously represented in a larger number of studies, thus increasing power of subsequent analyses. Moreover, the overall effect sizes across all assessment times for the caregiver- and youth-reported outcomes were more similar (caregiver $d = 0.97$; youth $d = 1.17$) and nonsignificantly different from each other ($Tau^2 = 0.21$, $p = 0.10$) as compared to the evaluator-reported continuous and frequency data outcomes, which were more similar to each other (evaluator continuous measure $d = 2.08$; evaluator frequency outcome $d = 2.38$) and nonsignificantly different ($Tau^2 = 0.41$, $p = 0.07$). In contrast, youth-reported effect sizes were significantly different from both evaluator-reported continuous measures ($Tau^2 = 0.21$, $p < 0.001$) and evaluator-reported frequency outcomes ($Tau^2 = 0.23$, $p < 0.001$), as were caregiver-reported effect sizes from evaluator-reported continuous measures ($Tau^2 = 0.36$, $p < 0.001$) and evaluator-reported frequency outcomes ($Tau^2 = 0.32$, $p < 0.001$).

At posttreatment, the mean effect size was 0.89 ($SE = 0.05$, *95% CI* = [0.80, 0.98]) for client-reported outcomes (youth/caregiver-reported) and 1.97 ($SE = 0.10$, *95% CI* = [1.79, 2.16]) for evaluator-reported outcomes. See Table 1 for effect sizes by respondent type across follow-up time. By using meta-regression with a robust variance estimation approach, all effect sizes across studies and follow-up assessments were considered at the same time, and respondent type was a significant predictor of effect size ($n = 789$ effect sizes, $n = 79$ studies, $Tau^2 = 0.26$, $p < 0.001$, with conservative $rho = 1.0$), with outcome measures from caregiver and youth respondents associated with smaller effect sizes ($B = -0.97$, $p < 0.001$). The same Tau^2 value (0.26) was found via sensitivity analysis with all other values of *rho*. Next, we ran meta-regressions for all effect sizes across studies for each assessment time individually (all with conservative $rho = 1.0$). Respondent type was significantly associated with effect size at posttreatment ($n = 323$ effect sizes, $n = 79$ studies, $Tau^2 = 0.23$, $p < 0.001$), at 3-month follow-up ($n = 128$ effect sizes, $n = 36$ studies, $Tau^2 = 0.28$, $p = 0.003$), at 6-month follow-up ($n = 121$ effect sizes, $n = 29$ studies, $Tau^2 = 0.23$, $p < 0.001$), and at 1-year follow-up ($n = 169$ effect sizes, $n = 34$ studies, $Tau^2 = 0.39$, $p = 0.001$). For each meta-regression, the respective Tau^2 value remained the same across sensitivity analyses with

all other values of *rho*. In each respective meta-regression, outcome measures from caregiver or youth respondents (as opposed to evaluator) were associated with smaller effect sizes at posttreatment (B = -1.01, $p < 0.001$), 3-month follow-up (B = -0.92, $p = 0.003$), 6-month follow-up (B = -0.84, $p < 0.001$), and 1-year follow-up (B = -0.87, $p = 0.001$). A meta-regression was not run for either 1-month or long-term follow-up effect sizes, as they would have been underpowered (1-month follow-up: n = 21 effect sizes, n = 9 studies; long-term follow-up: n = 27 effect sizes, n = 8 studies).

POST HOC ANALYSES

Given that outcome respondent type was significantly associated with effect size, we conducted further analyses to determine what confounding variables might have affected the magnitude of effect sizes across follow-up assessments. First, 83% of the groups (10 of 12) assessed at long-term follow-up and 90% of the groups (45 of 49) assessed at 1-year follow-up featured evaluator outcomes, which was disproportionately larger than the 76% of groups that featured evaluator outcomes at posttreatment (84 of the 110). Second, on average, studies that assessed outcomes through 1-year follow-up had effect sizes based on 92% of the target symptom measures in the study, whereas studies that assessed outcomes through long-term follow-up had effect sizes based on 78% of the target symptom measures in the study. Thus, the effect sizes for groups with later follow-up assessments were based on a smaller proportion of the total number of target symptom measures. Third, method of moments meta-regression revealed significant between-group variance of posttreatment effect sizes explained (R^2 analog = 0.10, $F[1, 108]$ = 4.07, p = .047), with larger effect sizes for those groups in studies using follow-up assessment of 1 year or more (B = 0.25, t = 2.01, p = .047). The goodness-of-fit test revealed that significant unexplained within-group variance remained (I^2 = 65.93%, Tau^2 = 0.18, Q = 317.03, df = 108, $p <$ .001). The results of this analysis suggest that the studies that had later follow-up assessments already had significantly larger effect sizes at posttreatment.

ASSESSMENT OF DATA BIAS

First, we assessed our data for potential publication bias. The funnel plot plotting effect size and standard error was asymmetric and Egger's weighted regression test was significant (t = 5.26, $p <$ .001). These results indicate that larger effect sizes were disproportionately included in our article sample, suggesting publication bias (Jüni, Holenstein, Sterne, Bartlett, & Egger, 2002). However, the fail-safe N indicated that 6,780 studies with an effect size of 0 at posttreatment would need to be added to our article sample before the average posttreatment effect size would become nonsignificantly different from zero. Next, we assessed potential bias due to selective reporting. Of the 110 groups, 85% reported usable statistics for all target symptom measures listed in their methods. However, the method of moments meta-regression found no significant effect of complete versus incomplete reporting of measures on CBT group effect size (posttreatment, p = .41; follow-up: p = .37). Finally, we assessed for bias due to incomplete data reporting (e.g., due to dropout over time). Using the raw data type as the index group, no significant effect was found for intent-to-treat, completer, or imputed data types (posttreatment, p = .25 – .68; follow-up: p = .12 – .42). Taken together, these findings suggest that our obtained effect sizes were likely not impacted by incomplete data or selective reporting biases.

Discussion

This meta-analysis examined CBT effects for youth anxiety, depression, and traumatic stress symptoms across follow-up assessment time points. Previous meta-analyses and reviews have often focused on posttreatment outcomes only or on follow-up outcomes in a limited way. In contrast, our primary aim was to assess if youth CBTs for internalizing disorders are durable through follow-up one year or longer. Overall, our results are promising for the durability of youth CBTs' effects through one-year or more of follow-up assessment. Given that our sample included youth CBTs regardless of their posttreatment performance, "durability" of effects in our meta-analytic results should be interpreted as retained magnitude of effect size from posttreatment to follow-up assessment, relative for each group. Our second aim was to examine how effect sizes related to research-related variables. A closer look at our results demonstrated how incomplete the picture still is for youth CBT effect durability, and how difficult it is to disentangle if longer-term follow-up effects are being driven by the CBTs themselves or research-related variables.

Effect sizes for youth CBTs in our sample were large at posttreatment, all well above Cohen's (Cohen, 1988) suggested guideline of 0.80 as an indicator of large effects. Somewhat larger effect sizes were found for youth CBTs targeting anxiety (g = 1.75, SE = 0.11) and traumatic stress (g = 1.67, SE = 0.27) as compared to those targeting depression (g = 1.31, SE = 0.13). These findings are consistent with meta-analyses that have reviewed psychotherapies for such internalizing disorders separately; overall, youth CBTs for

internalizing symptoms have established strong empirical support meta-analytically, with those targeting depression producing relatively more moderate effect sizes (e.g., Bennett et al., 2016; Morina, Koerssen, & Pollet, 2016; Weisz et al., 2006). The effect sizes presented in the current review are even larger still than what you would see in many other meta-analyses on youth CBTs, likely due to our calculation of within-group effect sizes, as opposed to the more commonly used formulas that account for change in a comparison group.

Effect sizes were also large at all follow-up assessments for youth CBTs targeting anxiety, traumatic stress, and depression. The average post-to-follow-up effect size at 1-month, 3-month, and long-term follow-up demonstrated no significant change from posttreatment effect size, as their confidence intervals overlapped. The 6- and 12-month follow-ups demonstrated significant additional improvement, as their average effect sizes were larger and their confidence intervals did not include values within the 95% confidence interval of the posttreatment intercept. The additional improvement was true only for the 12-month follow-up effect sizes when examining just "winning" CBT groups. These findings of effect durability are consistent with the findings of a recent meta-analysis of long-term outcomes for youth CBTs targeting anxiety (Gibby, Casline, & Ginsburg, 2017); our findings extend support of long-term effect durability for youth CBTs targeting traumatic stress and depression as well. The effect sizes did not vary by treatment target at posttreatment or any follow-up assessment. The variability in timing and low rates of longer-term follow-up assessments is worth noting: 100% of groups had follow-up of 1 month or more, 95% had follow-up of 3 months or more, 59% of groups had follow-up of 6 months or more, 46% of groups had follow-up of 1 year or more, and 11% of groups had follow-up of 2 years or more. Therefore, although effect sizes were still large at long-term follow-up, they were based on a small subsample of youth CBTs.

Moreover, the follow-up effect size results need to be considered within the context of confounding research-related variables. Outcome respondent type was a significant predictor of effect size, with larger effect sizes found for evaluator-reported measures. This result was found across all assessment times, except 1-month and long-term follow-up, for which there were too few studies to run the test. It is unclear whether these findings can be interpreted as indication of underreporting of symptom improvement by youth and their caregivers, or of overreporting of symptom improvement by study evaluators. Regardless, these findings highlight the importance of considering how the type of measures that are included in a study can drive the magnitude of effect size at posttreatment and through follow-up assessment. Problematically for our aim of assessing durability of longer-term follow-up effects, the few studies that included a 1-year or long-term follow-up also more frequently included evaluator-reported outcomes. These same studies also had effect sizes based, on average, on 78% of their original symptom measures. Taken together, assuming that such selective reporting would favor significant effects over nonsignificant effects, it seems likely that the effect sizes calculated in this meta-analysis for the 1-year and long-term follow-up times are overestimates of what the true effect sizes would have been if studies had reported all results, including more youth and caregiver effects. Finally, studies that had 1-year or longer follow-up assessments already had statistically larger effect sizes at posttreatment. This finding might indicate that there is something qualitatively different about such studies that ultimately results in larger effect sizes for the tested youth CBT. We can reason that studies with a planned longer-term follow-up assessment may have more funding, larger baseline enrollment, and/or more collaborators, which could improve their ability to deliver a CBT or at least improve their ability to collect data from a larger sample size. It is also possible that some of the other studies had in fact planned longer-term follow-up analyses, but given a relatively smaller (but still large) effect size at posttreatment, follow-up assessments were cancelled or follow-up results were not significant and hence not included in publications.

Our findings should be interpreted in the context of a few methodological limitations. First, our ability to represent the durability of youth CBT effects for internalizing symptoms is constrained by if or how studies actually assessed follow-up effects. We have no way of knowing what the follow-up effects of the excluded youth CBT articles (n = 27) would have been, had their studies conducted follow-up assessments. Similarly, we also excluded articles that did conduct a follow-up assessment (according to their methods) but either did not report follow-up results (n = 2) or reported results in an unusable way (n = 4; e.g., by responder only; re-randomization to differing maintenance conditions). Second, even though we tried to assess for bias due to incomplete data, we cannot truly know what the effect sizes would have been if all studies were able to collect posttreatment and follow-up data from their full sample. This issue is even more pronounced for calculation of follow-up effects, as

attrition often increases over time, and so effects are even more biased due to completer or intent-to-treat data. Third, the youth CBTs included in this sample had only one treatment target, and were not designed to target comorbid conditions. Comorbidity is more common in real-world clinical settings (Bearman & Weisz, 2015), so the current review's findings may not generalize to CBT treatment of youth with more complicated presenting problems. Fourth, follow-up effect sizes may be representing not only effects from the original treatment, but also from any other services in which youth and families received later. This issue is especially pertinent to longer-term follow-ups; for example, one youth anxiety treatment review found that one in three studies included youth who had received additional services during the time between study posttreatment and their long-term follow-up assessments (Gibby et al., 2017). Fifth, the principal investigator served as a coder on all articles and also conducted the research analyses, which is a limitation to controlling expectancy effects. However, all articles were double coded with high interrater reliability; therefore, final data was not dependent on the principal investigator's extracted data alone. Finally, given the variability of comparison group type in the included studies, we deemed it inappropriate to calculate and present between-group effect sizes across time, especially given the variability of their inclusion across follow-up timing (see Appendix A, Supplemental Table 2). Unfortunately, without between-group effect size results, comparison of the current meta-analysis to other relevant published meta-analysis studies is limited, as the majority of such studies report between-group effect sizes.

Limitations notwithstanding, our findings pose some important implications. Given the inconsistency in the selection of follow-up length across studies, psychotherapy research trials would benefit from developing more standardized procedures for follow-up assessment timing. In our review, greatest consistency was found for 9 studies which each had a (1) 6-month, (2) 1-year, and (3) either 3-month or long-term follow-up. Such variable timing has been observed in other meta-analyses examining long-term effects of psychotherapies for youth (e.g., Gibby et al., 2017; Gutermann et al., 2017) and adults (Flückiger, Del Re, Munder, Heer, & Wampold, 2014), and is often cited as the reason why follow-up assessment intervals get collapsed at the meta-analytic level. In contrast, a meta-analysis of 17 postoperative cancer treatments was able to identify a sample that all used 6-, 12-, 18-, and 24-month follow-up assessments (Musekiwa et al., 2016). Establishing such consistency in follow-up assessment timing for psychotherapy research would improve precision of long-term effect size estimates.

Moreover, considering that so few trials conducted long-term follow-up, there are implications for how we design research studies and how we form policies related to research. Researchers could shift towards using longer-term follow-up assessments in more research trials. If a youth CBT demonstrates significant posttreatment effects, follow-up assessments are imperative to confirm that benefits have been maintained. However, conducting follow-up assessments is often expensive, time consuming, and burdensome for youth and caregivers. To this end, we need to prioritize developing follow-up procedures that are more feasible and accessible, to ensure that long-term follow-up is not just for the studies that can afford them. In turn, policies related to standards of evidence may need to incorporate ongoing evaluation for evidence-based treatments such that a treatment's status as "evidence-based" is not solidified once they have demonstrated large effects at posttreatment. If long-term follow-up effects for a certain youth CBT demonstrate significantly decreased magnitude (or are not assessed at all), it may be necessary to reconsider its "efficacious" status.

Finally, the findings related to outcome respondent type warrant further exploration. We are not the first study to find that outcome measure effect sizes vary by outcome respondent type (e.g., Becker-Haimes, Jensen-Doss, Birmaher, Kendall, & Ginsburg, 2018; Weisz et al., 2017). However, our findings extend previous research findings by demonstrating that such differences persist across follow-ups. It is possible that the evaluator effects are exaggerated by biased overreporting of improvement, as masking of evaluators is not always possible at follow-ups, given that all participants (including post-waitlist) likely should have received treatment by this point. Alternatively, self-report measures may be less sensitive to detecting maintenance of treatment gains, as youth and caregivers are likely not reporting current symptoms with a fully accurate recollection of baseline symptom severity as a point of comparison. Future research should aim to investigate sources of respondent-driven effect size differences, in order to better understand how to interpret follow-up effects from different reporters. Regardless, our findings suggest that relying exclusively on evaluator-reported variables might not fully capture the degree of change that a youth has experienced, and policies regarding evidence standards would do better to incorporate criteria for multi-informant evidence.

Although we achieved our aims to investigate the current state of effect durability in youth CBTs for internalizing symptoms, we also discovered that the current state of evidence cannot fully address many of our research questions surrounding durability. That said, large effect sizes through 1-year and longer follow-up are promising, as our results could have alternatively suggested overall deterioration of symptom improvement. However, we believe it is equally important to underscore that there is much room for improvement in long-term follow-up assessment. This implication is not unique to youth treatments, to CBT treatments, or to treatments for internalizing disorders, as any psychological treatments could benefit from more attention given to long-term follow-up assessment. Surrounded by hundreds of "evidence-based" psychotherapies—for youth CBT or otherwise—follow-up data may be useful in choosing among the multitude of treatment options. As a field, it is imperative to reconsider the significance of post-treatment effects on their own and the added value of follow-up data to the consumer, clinician, and organizational setting.

Role of Funding Sources

This research was supported in part by the University of California – Los Angeles Graduate Summer Research Mentorship Program and by the Dolores Zohrab Liebmann Fund. The funding sponsors had no direct involvement in any research activities.

Conflict of Interest Statement

Leslie Rith-Najarian and Alayna Park have received consulting fees from PracticeWise, LLC, a company that offers services discussed in this paper. Dr. Bruce Chorpita is a partner in PracticeWise, LLC. The content is solely the responsibility of the authors and does not necessarily represent the official views of either funding organization or of PracticeWise, LLC.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.beth.2018.05.006.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. Hoboken, NJ: Wiley, John & Sons.

Aldao, A., Nolen-Hoeksema, S., & Schweizer, S. (2010). Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical Psychology Review*, *30*, 217–237. https://doi.org/10.1016/j.cpr.2009.11.004

Bearman, S. K., & Weisz, J. R. (2015). Review: Comprehensive treatments for youth comorbidity - evidence-guided approaches to a complicated problem. *Child and Adolescent Mental Health*, *20*, 131–141. https://doi.org/10.1111/camh.12092

Becker-Haimes, E. M., Jensen-Doss, A., Birmaher, B., Kendall, P. C., & Ginsburg, G. S. (2018). Parent–youth informant disagreement: Implications for youth anxiety treatment. *Clinical Child Psychology and Psychiatry*, *23*, 42–56. https://doi.org/10.1177/1359104516689586

Bennett, K., Manassis, K., Duda, S., Bagnell, A., Bernstein, G. A., Garland, E. J., & Wilansky, P. (2016). Treating child and adolescent anxiety effectively: Overview of systematic reviews. *Clinical Psychology Review*, *50*, 80–94. https://doi.org/10.1016/j.cpr.2016.09.006

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Power analysis for meta-analysis. *Introduction to Meta-analysis* (pp. 258–276). Chichester, West Sussex, UK: John Wiley & Sons.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2013). *Comprehensive Meta-Analysis Version 3*. Englewood, NJ: Biostat.

Briggs, E. S., & Price, I. R. (2009). The relationship between adverse childhood experience and obsessive-compulsive symptoms and beliefs: The role of anxiety, depression, and experiential avoidance. *Journal of Anxiety Disorders*, *23*, 1037–1046. https://doi.org/10.1016/j.janxdis.2009.07.004

Burt, K. B., & Paysnick, A. a. (2012). Resilience in the transition to adulthood. *Development and Psychopathology*, *24*, 493–505. https://doi.org/10.1017/S0954579412000119

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2009). A SAS macro for statistical power calculations in meta-analysis. *Behavior Research Methods*, *41*, 35–46. https://doi.org/10.3758/BRM.41.1.35

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*, 283–292. https://doi.org/10.3102/0013189X16656615

Chorpita, B. F., & Daleiden, E. L. (2009). Mapping evidence-based treatments for children and adolescents: application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology*, *77*, 566–579. https://doi.org/10.1037/a0014565

Chorpita, B. F., Daleiden, E. L., Ebesutani, C., Young, J., Becker, K. D., Nakamura, B. J., & Starace, N. (2011). Evidence-based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. *Clinical Psychology: Science and Practice*, *18*, 154–172. https://doi.org/10.1111/j.1468-2850.2011.01247.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Copeland, W. E., Angold, A., Shanahan, L., & Costello, E. J. (2014). Longitudinal patterns of anxiety from childhood to adulthood: The Great Smoky Mountains study. *Journal of the American Academy of Child and Adolescent Psychiatry*, *53*, 21–33. https://doi.org/10.1016/j.jaac.2013.09.017

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-Informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *141*, 858–900. https://doi.org/10.1037/a0038498

Dorsey, S., McLaughlin, K. A., Kerns, S. E. U., Harrison, J. P., Lambert, H. K., Briggs, E. C., & Amaya-Jackson, L. (2017). Evidence base update for psychosocial treatments for children and adolescents exposed to traumatic events. *Journal of Clinical Child & Adolescent Psychology*, *46*, 303–330. https://doi.org/10.1080/15374416.2016.1220309

Dunn, V., & Goodyer, I. M. (2006). Longitudinal investigation into childhood- and adolescence-onset depression: Psychiatric outcome in early adulthood. *British Journal of Psychiatry*, *188*, 216–222. https://doi.org/10.1192/bjp.188.3.216

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple , graphical test measures of funnel plot asymmetry. *BMJ*, *315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Flückiger, C., Del Re, A., Munder, T., Heer, S., & Wampold, B. (2014). Enduring effects of evidence-based psychotherapies in acute depression and anxiety disorders versus treatment as usual at follow-up: A longitudinal meta-analysis. *Clinical Psychology Review*, *34*, 367–375. https://doi.org/10.1016/j.cpr.2014.05.001

Gibby, B. A., Casline, E. P., & Ginsburg, G. S. (2017). Long-term outcomes of youth treated for an anxiety disorder: A critical review. *Clinical Child and Family Psychology Review*, *20*, 201–225. https://doi.org/10.1007/s10567-017-0222-9

Ginsburg, G. S., Becker, E. M., Keeton, C. P., Sakolsky, D., Piacentini, J., Albano, A. M., & Kendall, P. C. (2014). Naturalistic follow-up of youths treated for pediatric anxiety disorders. *JAMA Psychiatry*, *71*, 310–318. https://doi.org/10.1001/jamapsychiatry.2013.4186

Gutermann, J., Schwartzkopff, L., & Steil, R. (2017). Meta-analysis of the long-term treatment effects of psychological interventions in youth with PTSD symptoms. *Clinical Child and Family Psychology Review*, *20*, 422–434. https://doi.org/10.1007/s10567-017-0242-5

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*, 167–178. https://doi.org/10.1037/0033-2909.117.1.167

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*, 107–128. https://doi.org/10.2307/1164588

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. San Diego, CA: Academic Press.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*, 203–217. https://doi.org/10.1037/1082-989X.6.3.203

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65. https://doi.org/10.1002/jrsm.5

Higa-McMillan, C. K., Francis, S. E., Rith-Najarian, L., & Chorpita, B. F. (2016). Evidence base update: 50 years of research on treatment for child and adolescent anxiety. *Journal of Clinical Child & Adolescent Psychology*, *45*, 91–113. https://doi.org/10.1080/15374416.2015.1046177

IBM Corp (2016). *IBM SPSS Statistics for Windows*. Armonk, NY: IBM Corp.

James, A. C., James, G., Cowdrey, F. A., Soler, A., & Choke, A. (2015). Cognitive behavioural therapy for anxiety disorders in children and adolescents. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD004690.pub4

Jüni, P., Holenstein, F., Sterne, J., Bartlett, C., & Egger, M. (2002). Direction and impact of language bias in meta-analyses of controlled trials: empirical study. *International Journal of Epidemiology*, *31*, 115–123. https://doi.org/10.1093/ije/31.1.115

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *The American Psychologist*, *63*, 146–159. https://doi.org/10.1037/0003-066X.63.3.146

Kendall, P. C., & Peterman, J. S. (2015). CBT for adolescents with anxiety: Mature yet still developing. *American Journal of Psychiatry*, *172*, 519–530. https://doi.org/10.1176/appi.ajp.2015.14081061

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*, 2693–2710. https://doi.org/10.1002/sim.1482

López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology*, *67*, 30–48. https://doi.org/10.1111/bmsp.12002

Manassis, K., Lee, T. C., Bennett, K., Zhao, X. Y., Mendlowitz, S., Duda, S., & Wood, J. J. (2014). Types of parental involvement in CBT with anxious youth: A preliminary meta-analysis. *Journal of Consulting and Clinical Psychology*, *82*, 1163–1172. https://doi.org/10.1037/a0036969

Morina, N., Koerssen, R., & Pollet, T. V. (2016). Interventions for children and adolescents with posttraumatic stress disorder: A meta-analysis of comparative outcome studies. *Clinical Psychology Review*, *47*, 41–54. https://doi.org/10.1016/j.cpr.2016.05.006

Musekiwa, A., Manda, S. O. M., Mwambi, H. G., & Chen, D. -G. (2016). Meta-analysis of effect sizes reported at multiple time points using general linear mixed model. *PLoS One*, *11*e0164898. https://doi.org/10.1371/journal.pone.0164898

Olatunji, B. O., Naragon-Gainey, K., & Wolitzky-Taylor, K. B. (2013). Specificity of rumination in anxiety and depression: A multimodal meta-analysis. *Clinical Psychology: Science and Practice*, *20*, 225–257. https://doi.org/10.1111/cpsp.12037

Polanczyk, G., Salum, G., Sugaya, L., Caye, A., & Rohde, L. (2015). Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *56*, 345–365. https://doi.org/10.1111/jcpp.12381

PracticeWise (2012). *Psychosocial and Combined Treatments Coding Manual*. Satellite Beach, FL: PracticeWise, LLC.

PracticeWise (2017). PracticeWise Evidence-Based Youth Mental Health Services Literature Database. Retrieved February 23, 2017, from. http://www.practicewise.com/pwebs_1/YouthSearch.aspx.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Spielmans, G. I., Pasek, L. F., & McFall, J. P. (2007). What are the active ingredients in cognitive and behavioral psychotherapy for anxious and depressed children? A meta-analytic review. *Clinical Psychology Review*, *27*, 642–654. https://doi.org/10.1016/j.cpr.2006.06.001

Stewart, D., Stavness, C., King, G., Antle, B., & Law, M. (2006). A critical appraisal of literature reviews about the transition to adulthood for youth with disabilities. *Physical and Occupational Therapy in Pediatrics*, *26*, 5–24. https://doi.org/10.1080/J006v26n04_02

TADS Team, March, J., Silva, S., Curry, J., Wells, K., Fairbank, J., & Bartoi, M. (2009). The Treatment for Adolescents With Depression Study (TADS): Outcomes over 1 year of naturalistic follow-up. *The American Journal of Psychiatry*, *166*, 1141–1149. https://doi.org/10.1176/appi.ajp.2009.08111620

Torgerson, C. J. (2006). Publication bias: The Achilles' heel of systematic reviews? *British Journal of Educational Studies*, *54*, 89–102. https://doi.org/10.1111/j.1467-8527.2006.00332.x

Verstraeten, K., Bijttebier, P., Vasey, M. W., & Raes, F. (2011). Specificity of worry and rumination in the development of anxiety and depressive symptoms in children. *The British Journal of Clinical Psychology / the British Psychological Society*, *50*, 364–378. https://doi.org/10.1348/014466510X532715

Weersing, V. R., Jeffreys, M., Do, M. -C. T., Schwartz, K. T. G., & Bolano, C. (2017). Evidence base update of psychosocial treatments for child and adolescent depression. *Journal of Clinical Child & Adolescent Psychology*, *46*, 11–43. https://doi.org/10.1080/15374416.2016.1220310

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. https://doi.org/10.1001/jamapsychiatry.2013.1176

Weisz, J. R., Kuppens, S., Ng, M. Y., Eckshtain, D., Ugueto, A. M., Vaughn-Coaxum, R., & Fordwood, S. R. (2017). What five decades of research tells us about the effects of youth psychological therapy: A multilevel meta-analysis and implications for science and practice. *American Psychologist*, *72*, 79–117. https://doi.org/10.1037/a0040360

Weisz, J. R., McCarty, C. A., & Valeri, S. M. (2006). Effects of psychotherapy for depression in children and adolescents: A meta-analysis. *Psychological Bulletin*, *132*, 132–149. https://doi.org/10.1037/0033-2909.132.1.132

Weisz, J. R., Southam-Gerow, M. A., Gordis, E. B., Connor-Smith, J. K., Chu, B. C., Langer, D. A., & Weiss, B. (2009). Cognitive-behavioral therapy versus usual clinical care for youth depression: an initial test of transportability to community clinics and clinicians. *Journal of Consulting and Clinical Psychology*, *77*, 383–396. https://doi.org/10.1037/a0013877

Received: December 21, 2017
Accepted: May 25, 2018
Available online: 1 June 2018