Article

Assessment
1–13

DOI: 10.1177/1073191118765365
journals.sagepub.com/home/asm
SAGE

Generalizability and Decision Studies of a Treatment Adherence Instrument

Michael A. Southam-Gerow[1], Wes Bonifay[2], Bryce D. McLeod[1], Julia R. Cox[1], Stephanie Violante[1], Philip C. Kendall[3], and John R. Weisz[4]

Abstract

Observational measurement of treatment adherence has long been considered the gold standard. However, little is known about either the generalizability of the scores from extant observational instruments or the sampling needed. We conducted generalizability (G) and decision (D) studies on two samples of recordings from two randomized controlled trials testing cognitive–behavioral therapy for youth anxiety in two different contexts: research versus community. Two doctoral students independently coded 543 session recordings from 52 patients treated by 13 therapists. The initial G-study demonstrated that context accounted for a disproportionately large share of variance, so we conducted G- and D-studies for the two contexts separately. Results suggested that reliable cognitive–behavioral therapy adherence studies require at least 10 sessions per patient, assuming 12 patients per therapists and two coders—a challenging threshold even in well-funded research. Implications, including the importance of evaluating alternatives to observational measurement, are discussed.

Keywords

treatment integrity, youth anxiety, generalizability study, psychometrics, cognitive–behavioral therapy

Treatment integrity, a construct that includes adherence (how closely treatment delivered matches the intended plan), differentiation (the extent to which nonprescribed treatment content is present), and competence (the quality of treatment delivery) has become a focus of clinical research (e.g., Sanetti & Kratochwill, 2014; Schoenwald et al., 2011; Southam-Gerow & McLeod, 2013). At a basic level, treatment integrity is a manipulation check; it represents a measurement of the extent to which the independent variable (i.e., psychosocial treatment, hereafter called *treatment*) was delivered as designed (Perepletchikova & Kazdin, 2005). Treatment integrity is therefore an important construct for the advancement of clinical science.

Despite the importance of treatment integrity, scientists have not afforded much attention to its measurement and study. Reviews of the treatment integrity literature have concluded that measurement of the construct is lacking (e.g., Perepletchikova, Treat, & Kazdin, 2007; Webb, DeRubeis, & Barber, 2010), particularly in the area of child and adolescent (hereafter, youth) treatment (Goense, Boendermaker, van Yperen, Stams, & van Laar, 2014; McLeod, Southam-Gerow, & Weisz, 2009). The integrity measurement that has been reported usually focuses on verifying the presence of the independent variable, assessed in a binary fashion (e.g., Perepletchikova et al., 2007; Schoenwald et al., 2011). With a focus on adherence, several research groups in the youth treatment field have advanced treatment integrity measurement by publishing promising findings related to caregiver-report (Huey, Henggeler, Brondino, & Pickrel, 2000; Schoenwald, Sheidow, & Chapman, 2009), therapist-report (Hogue, Dauber, Lichvar, Bobek, & Henderson, 2015; Ward et al., 2013), and observational (Hogue et al., 2008; Southam-Gerow et al., 2016) integrity instruments. However, there is general consensus that the evidence base for treatment integrity measurement needs further development (e.g., Bruhn, Hirsch, & Lloyd, 2015; Webb et al., 2010).

Observation is commonly used to assess treatment integrity (Perepletchikova et al., 2007; Webb et al., 2010). Though some consider this approach to be the gold standard (e.g., McLeod et al., 2009), observational measurement is both time and labor intensive. Perhaps as a result, studies that employ observational treatment integrity instruments

[1]Virginia Commonwealth University, Richmond, VA, USA
[2]University of Missouri, Columbia, MO, USA
[3]Temple University, Philadelphia, PA, USA
[4]Harvard University, Cambridge, MA, USA

Corresponding Author:
Michael A. Southam-Gerow, Department of Psychology, Virginia Commonwealth University, 806 West Franklin Street, P.O. Box 842018, Richmond, VA 23284-2018, USA.
Email: masouthamger@vcu.edu

only rate an average of 2.50 sessions per patient (Dennhag, Gibbons, Barber, Gallop, & Crits-Christoph, 2012). Research has suggested that this number of sessions may not be sufficient to provide a reliable estimate of treatment integrity (Crits-Christoph et al., 2011; Dennhag et al., 2012). Data-driven guidance is needed to provide estimates of how many (and also how many minutes of) sessions of youth treatment need to be coded to produce a reliable estimate of treatment adherence (Perepletchikova et al., 2007).

One way to produce such an estimate is to conduct a generalizability and decision study (hereafter, G-study and D-study; e.g., Brennan, 2010; Dennhag et al., 2012; Gresham, Dart, & Collins, 2017). These methodologies help identify appropriate sampling strategies, thereby improving efficiency (e.g., Crits-Christoph et al., 2011; Dennhag et al., 2012). Generalizability theory (Cronbach, Rajaratnam, & Gleser, 1963; Shavelson & Webb, 1991; Wasserman, Levy, & Loken, 2009) is an empirical approach developed to determine how applicable the scores from one sample are to the scores of the broader population of interest from which the sample was drawn. A G-study generates indices of the dependability of the sample. Specifically, the generalizability coefficients provide an estimate of the accuracy of one observed score as a representation of the population mean. The analysis also identifies important sources of variability in the item scores—in the case of treatment integrity, this would mean the variability in adherence across relevant design facets (e.g., therapists, patients, and sessions).

A G-study can be followed by a D-study, an analysis particularly relevant to the issue of efficiency. The D-study provides guidance on the optimal research design needed for a dependable measurement of the variable of interest. Specifically, a D-study helps inform future studies by identifying combinations of design facets (e.g., number of sessions × number of patients) that will produce reliable estimates. Of particular relevance to observational treatment integrity measurement is determining the number of sessions that need to be coded. A D-study provides a sampling recommendation driven by data rather than convention.

G- and D-studies have been used for a variety of research areas. Education research has used these studies to determine the optimal number of days and raters needed to use limited resources while increasing reliability of observational measures of, for example, the quality of classroom interactions (Mashburn, Downer, Rivers, Brackett, & Martinez, 2014), teacher performance (Hill, Charalambous, & Kraft, 2012), and implementation of a classroom-based intervention (e.g., Good Behavior Game; Gresham et al., 2017). Similarly, medical researchers use G- and D-studies to determine the optimal balance between number of test procedures performed and number of raters needed to get a reliable measure of trainee (e.g., Konge, Vilmann, Clementsen, Annema, & Ringsted, 2012; Todsen et al., 2015) or diagnostic test performance (Nielsen, Jensen, & O'Neill, 2015) or to assess the generalizability of a medical school admissions test (Sebok, Luu, & Klinger, 2014). Several recent papers have promoted G- and D-studies as important and underused for informing research design across many applied settings (e.g., Briesch, Swaminathan, Welsh, & Chafouleas, 2014; Preuss, 2013).

Specific to therapy process measurement (e.g., integrity, alliance), two recent studies (Crits-Christoph et al., 2011; Dennhag et al., 2012) have attempted to determine the appropriate number of sessions per patient and patients per therapist to establish dependable ratings. Dennhag et al. (2012) calculated generalizability coefficients from session-level observational adherence and competence ratings generated during a randomized controlled trial (RCT) of three active treatments for cocaine dependence. At the patient level, the number of sessions needed to achieve a "good" generalizability coefficient (≥.80; Cardinet, Johnson, & Pini, 2010; Wass, van der Vleuten, Shatzer, & Jones, 2001) ranged from 5 to 10, depending on treatment and measurement of adherence or competence. At the therapist level, the number of patients necessary ranged from 4 to more than 14, depending on the number of observed sessions, treatment modality, and measurement of adherence or competence (Dennhag et al., 2012). Crits-Christoph et al. (2011) used patient-reported alliance ratings from a trial of alliance-fostering treatment for major depressive disorder to calculate patient- and therapist-level generalizability statistics. Patient-level generalizability was .93 at four assessments of alliance; therapist-level generalizability, however, was much more difficult to get a reliable estimate and required a large number of patients per therapist. In sum, findings in G- and D-studies of treatment integrity have suggested (a) generalizability may vary by treatment and (b) far more observed sessions per patient and patients per therapist are likely needed to establish dependable treatment integrity ratings than are commonly found in treatment integrity studies.

The present study reports results from a G- and subsequent D-study. We conducted two separate G/D studies. The first drew on data from an efficacy trial of individual cognitive–behavioral therapy (ICBT) for youth anxiety conducted in a university-based research clinic. The second smaller set of studies drew on data from an effectiveness trial conducted in community clinics. We used these two data sets to gauge the generalizability of the adherence scores and then to determine the sampling parameters needed to maximize generalizability in future. Our sampling dimensions included the number of recordings (sessions), number of cases/patients, and number of therapists (cf. Crits-Christoph et al., 2011; Dennhag et al., 2012). Also of interest was whether these parameters needed to differ to obtain reliable estimates across context (i.e., research vs. community). This is the first G- and D-study of treatment integrity in youth.

Method

Data Sources and Participants

The first study was conducted on data from an efficacy RCT conducted by Kendall, Hudson, Gosch, Flannery-Schroeder, and Suveg (2008), who compared the efficacy of ICBT, family cognitive–behavioral therapy (CBT), and an active control condition. Our second study used data drawn from the RCT conducted by Southam-Gerow et al. (2010), who compared the effectiveness of ICBT with usual care. We focused solely on the two ICBT conditions and because the main difference between the two was the setting or context in which they were delivered, we refer to the two groups as the *research* (Kendall et al., 2008) and *community* (Southam-Gerow et al., 2010) contexts. Our data were recorded treatment sessions collected in each RCT. To be included in this study, youth had to: (a) have at least two audible recorded sessions and (b) have received ICBT from a single therapist (see Kendall et al., 2008; Southam-Gerow et al., 2010, for more details). As will be detailed later, we also only included youth whose therapists had more than one patient in the study. Without multiple sessions per patient and patients per therapist, there would be no variability within the adherence ratings of a given therapist, and the G-study would not be able to quantify the effects of these key facets.

The initial sample pool included 51 youth participants from the research context (Kendall et al., 2008) and 17 from the community context (Southam-Gerow et al., 2010). Applying our inclusion and exclusion criteria, our final analyzed sample included 45 youth from the research context and 7 from the community context. In the research sample, youth aged 7 to 14 years (M = 10.20, SD = 1.90) were 42.2% female; these youth identified as 88.9% White, 6.7% African American, 2.2% Latino/a, and 2.2% other race/ethnicity. In the community sample, youth aged 8 to 14 years (M = 11.40, SD = 2.50) were 71.4% female; these youth identified as 75.0% White, 25.0% Latino/a, with three participants not reporting ethnicity.

Clinical psychology doctoral trainees and licensed psychologists delivered ICBT in the research context (n = 10; 90.0% female, 77.8% White). Therapists in the community context were clinic employees who volunteered to participate and were randomly assigned to either receive training (or not) in ICBT for youth anxiety. Community therapists (n = 3) were 66.6% female and 100% White. The professional makeup was 66.6% social workers and 33.3% marriage and family therapist.

Individual Cognitive–Behavioral Therapy

Therapists in both RCTs delivered *Coping Cat*, an ICBT program for youth diagnosed with anxiety disorders (Kendall & Hedtke, 2006a, 2006b), which consists of 16 sessions; 14 sessions are conducted individually with the youth and two sessions are conducted with the parents. The first half focuses on anxiety management skills training (e.g., relaxation, problem solving), whereas the second half emphasizes exposure. Homework is regularly assigned to the youth throughout the program. Gold standard quality control methods, including the use of a treatment manual, a training workshop, and ongoing, weekly supervision with an expert in CBT for youth anxiety (Sholomskas et al., 2005), were used in both RCTs. Furthermore, adherence to *Coping Cat* was measured with the *Coping Cat* Brief Adherence Scale (see Kendall, 1994; Kendall et al., 1997), an observational scale that uses a checklist format (presence/absence of interventions) to determine if core ICBT interventions were delivered. Based on the scale, therapists in both studies showed more than 90.0% adherence (see Kendall et al., 2008; Southam-Gerow et al., 2010, for details).

Treatment Adherence Instrument

CBT for Anxiety in Youth Adherence Scale (CBAY-A; Southam-Gerow et al., 2016). The CBAY-A is a 22-item instrument gauging three facets of treatment: (a) *Standard*, 4 items that represent general CBT interventions (e.g., Homework Assigned), (b) *Model*, 12 items that assess ICBT-specific content (e.g., Relaxation, Exposure), and (c) *Delivery*, 6 items that assess how model items are delivered (e.g., Modeling, Rehearsal). CBAY-A items are scored on a 1 to 7 extensiveness scale that reflects the frequency and the thoroughness with which the therapist delivered the intervention (cf. Hogue, Liddle, & Rowe, 1996): 1 = *not at all*, 3 = *somewhat*, 5 = *considerably*, 7 = *extensively*. For the current study, we focused on the Model item scale (11 items; e.g., Psychoeducation, Problem Solving, Exposure). We focused on Model items because the model items are the most face valid in the instrument and thus the most likely to be used in future studies to measure adherence. The other two sets of items measure either more general therapist behaviors (i.e., standard items) or else the manner in which a therapist delivered a model item (i.e., the delivery items).

The CBAY-A model items have demonstrated strong intercoder reliability, ICC(2, 2) (where ICC is intraclass correlation coefficient), with a mean of 0.84 (SD = 0.15; range: 0.49-0.93) reported in the initial psychometric paper (Southam-Gerow et al., 2016). The same study provided evidence of construct validity, including scale scores that discriminated between therapists delivering ICBT across research/community settings and therapists delivering usual care (Southam-Gerow et al., 2016). For the present sample, the average item intercoder reliability (n = 550), ICC(2, 2), was 0.83 (SD = 0.15).

Coding and Session Sampling Procedures

Two doctoral students in clinical psychology (M_{age} = 26.80, SD = 1.70; 50.0% female, one Latina and one Caucasian)

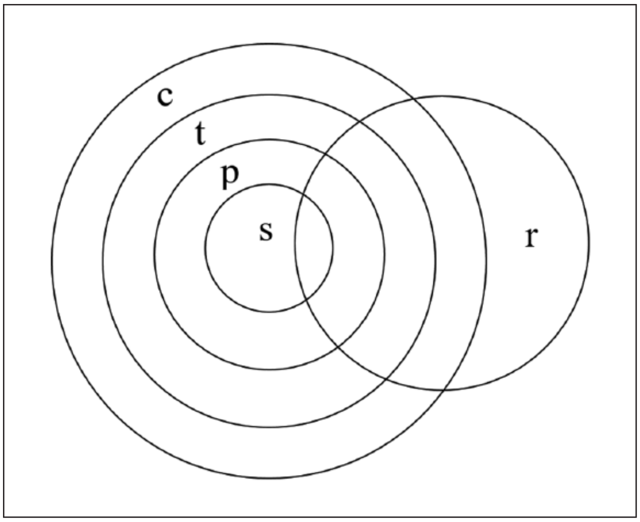

Figure 1. Venn diagram for a sessions within patients within therapists by coders [(s:p:t:c) × r] design.
Note. s = session; p = patient; t = therapist; c = context; r = coder.

comprised the coding team. Coders were trained over a 3-month period to reach adequate prestudy reliability at the item level, ICC(2, 2) > .59 (Cicchetti, 1994). Training progressed through three steps: *Step 1*. Coders received didactic instruction and discussion of CBT for youth anxiety and the scoring manual, reviewed sessions with the trainers, and engaged in coding exercises designed to test and expand understanding of each item. *Step 2*. Next, coders engaged in independent practice coding of recordings. In weekly meetings, results of the practice coding were discussed and specific illustrative segments reviewed. *Step 3*. Last, coders independently coded 32 recordings. Reliability for each coder was assessed against master codes. Once coders met "good" reliability on each item, ICC(2, 2) > .59 (Cicchetti, 1994), independent coding commenced. Once coding commenced, coders met regularly for the duration of coding to prevent coder drift (Margolin et al., 1998). Coding order for each coder was determined by random assignment. Each session was double-coded for reliability purposes. Coders were naïve to study hypotheses and data sources.

Analytic Approach

Data Structure. The data structure were as follows: treatment sessions (s) were nested within patients (p), who were nested within therapists (t), who were nested within contexts (c), and all of these facets were crossed with coders, that is, raters (r). Thus, the model for the G-study was specified as (s:p:t:c) × r, as depicted in Venn diagram form in Figure 1. To perform a logical G-study, some cases had to be removed prior to analysis. The patient facet was unbalanced; some therapists had a single patient, while others had as many as nine. All therapists who had just a single patient were eliminated—given that patients were nested within therapists, including such cases would allow the variance component associated with single patients to become confounded with the variance component associated with the therapist. Consequently, all therapists in the following analysis had a minimum of two patients.

The sessions facet was particularly difficult to handle. Not every session was recorded and some therapists conducted more sessions than others. That is, two therapists may have each conducted 15 sessions, but one may have had 3 recorded sessions, whereas the other had 14 recorded sessions. Previous studies with similar data structures addressed this issue by using just 2 sessions from each patient—one drawn randomly from the first 11 sessions and the other drawn randomly from Session 12 or higher (Barber, Foltz, Crits-Christoph, & Chittams, 2004; Dennhag et al., 2012). Rather than adopting this approach, the present study treated the sessions as unbalanced and included every recorded session in the analysis. It is therefore important to note that throughout this article, the variable labeled sessions (s) refers to the number of *coded sessions* and not the number of conducted (but uncoded) sessions.

As noted, the data were collected in two contexts. Within the research context, there were 10 therapists who had 6, 3, 2, 2, 8, 2, 3, 5, 9, and 5 patients, respectively, and the number of recorded sessions ranged from 3 to 14 (M = 10.36, SD = 2.67). Within the community context, there were 3 therapists who had 2, 2, and 3 patients, respectively, and the number of recorded sessions varied from 2 to 19 (M = 12.00, SD = 5.51).

Generalizability Study. G-studies are designed to identify and estimate variation due to the target of measurement (therapists) and all observable facets of measurement error in the data (number of contexts, patients, sessions, and raters). In our initial analysis, we examined the variance components across both contexts (research and community). The therapist-level generalizability coefficient (ρ^2_t) across contexts was computed using the following formula (extended from a less complex model presented in Brennan [2001a]):

$$\rho^2_t = \frac{\sigma^2_c + \sigma^2_{t:c}}{\sigma^2_c + \sigma^2_{t:c} + \frac{\sigma^2_{p:t:c}}{n_p} + \frac{\sigma^2_{s:p:t:c}}{n_s n_p} + \frac{\sigma^2_{cr}}{n_r} + \frac{\sigma^2_{(t:c)r}}{n_r} + \frac{\sigma^2_{(p:t:c)r}}{n_p n_r} + \frac{\sigma^2_{(s:p:t:c)r,e}}{n_s n_p n_r}} \tag{1}$$

where σ^2_c is the variance attributed to contexts, $\sigma^2_{t:c}$ is the variance attributed to therapists nested within contexts, $\sigma^2_{p:t:c}$ is the variance attributed to patients nested within

Table 1. Proportion of Variance Explained by the Model Terms Across Both the Research and Community Contexts.

Effect	*df*	*SS*	*MS*	Variance component	Variance percentage
c	1	51.90	51.90	.179	49.5
t:c	11	12.78	1.16	.007	2.1
p:t:c	39	16.60	0.43	.008	2.4
s:p:t:c	498	129.81	0.26	.103	28.6
r	1	2.90	2.90	.005	1.5
cr	1	0.03	0.03	.001	0.2
(t:c)r	1	1.60	0.15	.002	0.7
(p:t:c)r	11	1.86	0.05	.001	0.2
(s:p:t:c)r,e	39	27.20	0.05	.055	15.2

Note. df = degrees of freedom; *SS* = sum of squares; *MS* = mean square; c = context; t = therapist; p = patient; s = session; r = coder; e = error; n_t = 13; n_p = 3.02 (harmonic mean); n_s = 9.01 (harmonic mean); n_r = 2.

therapists nested within contexts, n_p is the number of patients, $\sigma^2_{s:p:t:c}$ is the variance attributed to sessions nested within patients nested within therapists nested within contexts, n_s is the number of sessions, σ^2_{cr} is the variance attributed to the coder × context interaction, n_r is the number of raters, $\sigma^2_{(t:c)r}$ is the variance attributed to the rater × therapist (nested within contexts) interaction, $\sigma^2_{(p:t:c)r}$ is the variance attributed to the rater × patient (nested within therapists nested within contexts) interaction, and $\sigma^2_{(s:p:t:c)r,e}$ is the variance that is not attributable to other factors in the design (i.e., error variance).

Next, we examined the variance components associated with each context in isolation. The separate generalizability coefficients for the research and community contexts were computed using a simpler version of Equation 1, in which σ^2_c (variance due to context) was removed from the numerator and denominator, along with all "c" subscripts from the remaining components.

Finally, it is important to note that both contexts entailed an unbalanced design in which the *n* values at certain levels differed. Following the advice given by Brennan (2001a), n_p and n_s in Equation 1 were computed using the harmonic mean of patients and sessions, respectively, within each context. The generalizability coefficient ρ^2_t in Equation 1 (and that of the research and community contexts in isolation) is analogous to the classical test theory index of reliability (coefficient α; see O'Brien, 1995) and is interpreted using the same guidelines (e.g., .70 to .80 = "acceptable"; .80 to .90 = "good"; .90+ = "excellent"; Cronbach, 1951; Kline, 2000).

After conducting the G-study, we carried out a D-study by using the observed numbers of sessions, patients, and raters to determine hypothetical combinations of facets needed to provide stable adherence ratings. Generalizability coefficients were computed at the therapist level for varying numbers of patients (8-20), sessions (10-19), and raters (2-4), and the number of therapists per context were fixed at the actual values observed in the G-study above. For our purposes, acceptable generalizability was defined as a coefficient between .70 and .80, while good generalizability was indicated by coefficients greater than or equal to .80 (Brennan, 2001a).

Given that our study possessed two samples from different contexts that differed across multiple possibly relevant dimensions (e.g., patient characteristics, therapist characteristics), we had a contingency plan if the initial G-study suggested that the majority of the variability was due to context. If that was the case, it would not be beneficial to use results from the G- (or D-) study to make recommendations for future research. As a result, if the initial G-study found that a majority of variance was attributable to the context, we planned to conduct G- and D-studies for each context separately: research and community.

Results

The following analysis was performed using urGENOVA version 2.1 (Brennan, 2001b), a software package for use with unbalanced designs. A G-study determined the specific variance components that explained the observed differences in therapists' adherence ratings across both contexts (research and community). As it happened, we did need to conduct the G- and subsequent D-studies within each context separately.

Study 1: Research and Community Contexts

G-Study. The G-study of the combined contexts produced the variance components displayed in Table 1. The largest proportion of variance was attributed to context, which accounted for 49.5% of the total variance in adherence scores. That is, the primary contributor to varying adherence scores was whether the therapist was part of the research (M = 1.83, SD = 0.40) or the community (M = 1.39, SD = 0.31) context. The number of sessions (nested within patients nested within therapists nested within contexts; s:p:t) accounted for 28.6% of the total variance in adherence scores, meaning the degree of therapist adherence varied significantly across coded sessions. Residual error accounted for 15.2% of the variance in adherence scores in the studies.

All of the remaining terms in the model accounted for minimal proportions of variance. Therapists (nested within contexts; t:c) accounted for 2.0% of the variance, patients (nested within therapists within contexts; p:t:c) accounted for 2.4% of the variance, and raters (r) accounted for 1.5% of the variance in adherence ratings. The three interaction terms—context × rater (cr), therapists (within context) × rater [(t:c)r], and patient (within therapists within context) × rater [(p:t:c)r]—each accounted for less than 1.0% of the variance and are therefore ignorable. Overall, more than

Table 2. Proportion of Variance Explained by the Model Terms in the Research Context.

Effect	*df*	*SS*	*MS*	Variance component	Variance percentage
t	9	11.87	1.32	.0080	3.9
p:t	35	15.46	0.44	.0078	3.8
s:p:t	421	124.94	0.30	.1179	57.7
r	1	2.67	2.67	.0053	2.6
tr	9	1.58	0.18	.0028	1.4
(p:t)r	35	1.67	0.05	.0013	0.6
(s:p:t)r,e	421	27.68	0.06	.0610	29.9

Note. df = degrees of freedom; *SS* = sum of squares; *MS* = mean square; t = therapist; p = patient; s = session; r = coder; e = error; n_t = 10; n_p = 3.37 (harmonic mean); n_s = 9.38 (harmonic mean); n_r = 2.

85.0% of the total variance in adherence scores was accounted for by the five facets in our G-study: treatment context and number of sessions accounted for the largest proportions of variance in adherence scores and all other terms explained trivial amounts of variance.

In research and community contexts, n_p ranged from 2 to 9 patients (harmonic mean = 3.02) and n_s ranged from 2 to 19 sessions (harmonic mean = 9.01). Plugging in the values from Table 1 into Equation 1 reveals the generalizability coefficient associated with the two contexts: ρ_t^2 = .95. This high-generalizability coefficient is due to the large impact of context; by including the contexts (c) facet in the generalizability equation, the numerator (c + the minor effect of therapists) is almost as large as the denominator (c + the minor effects of every facet except number of sessions and error), resulting in a coefficient close to 1.0.

Study 2: Research Context

G-Study. The G-study of the research context produced the variance components displayed in Table 2. The large proportion of variance was attributed to sessions (nested within patients nested within therapists; s:p:t), which accounted for 57.7% of the total variance in adherence scores. That is, the degree of therapist adherence varied significantly across coded sessions. The second largest proportion of variance was associated with residual error, as 29.9% of the variance in adherence scores in the research context could not be attributed to any of the four facets in the design. All of the remaining terms in the model accounted for minimal proportions of variance. Therapists accounted for 3.8% of the variance, patients (nested within therapists; p:t) accounted for 3.8% of the variance, and raters (r) accounted for 2.6% of the variance in adherence ratings. The two interaction terms—therapist × rater (tr), and patient (within therapists) × rater [(p:t)r]—each accounted for approximately 1.0% of the variance and are therefore ignorable. Overall, sessions and residual error accounted for the largest proportions of variance in adherence scores; all other terms explained trivial amounts of variance.

In the research context, n_p ranged from 2 to 9 patients (harmonic mean = 3.37) and n_s ranged from 3 to 14 sessions (harmonic mean = 9.38). Plugging in the values from Table 2 into the modified (context-free) version of Equation 1 reveals the generalizability coefficient associated with the research context: ρ_t^2 = .48.

D-Study. The results of the D-study based on the research variance components are shown in Table 3. In the research context, the number of patients per therapist ranged from 2 to 9, with a harmonic mean of 3.37, the number of sessions per patient ranged from 3 to 14 with a harmonic mean of 9.38, and there were two raters. The rows of Table 3 reveal that these numbers were too low to achieve an acceptable generalizability coefficient (≥.70). The most impactful facet was the number of raters. With only two raters, ρ_t^2 would not reach a good level of generalizability (≥.80), even with 18 patients who each have 19 coded sessions. With four raters, however, good generalizability could be achieved with 13 patients and 14 sessions.

Study 3: Community Context

G-Study. The G-study of the community context produced the variance components displayed in Table 4. Again, the largest proportion of variance was attributed to sessions (nested within patients nested within therapists; s:p:t), though this facet accounted for just 35.9% of the total variance in adherence scores (compared with 57.7% in research). The second largest proportion of variance was associated with residual error, which was slightly higher in the community (32.7%) than it was in the research (29.9%). Unlike research, however, the community G-study revealed that the number of patients (nested within therapists) accounted for a considerable percentage (15.3%) of the variance in adherence scores. Therapists (t) accounted for just 4.5% of the variance and raters (r) accounted for 5.0%. The therapist × rater (tr) interaction accounted for 2.5% of the variance and the patient (within therapists) × rater [(p:t)r] interaction accounted for 4.1%. Overall, sessions, residual error, and patients accounted for the largest proportions of variance in adherence scores; all other terms explained lesser amounts of variance.

Inputting the values from Table 4 into the modified (context-free) version of Equation 1 reveals the generalizability coefficient associated with the community context: ρ_t^2 = .27. This suboptimal generalizability coefficient is likely explained by the high-error variance as well as the low number of patients per therapist in the community context. This study included only three therapists, each of whom had very few patients; specifically, the therapists had 2, 2, and 3 patients (harmonic mean = 2.25). As we will see in the

Table 3. Generalizability Coefficients at the Therapist Level for Adherence Ratings Across Various Numbers of Patients, Sessions, and Coders in the Research Trial.

	Sessions																													
	10			11			12			13			14			15			16			17			18			19		
	Coders			Coders			Coders			Coders			Coders			Coders			Coders			Coders			Coders			Coders		
Patients	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4
5	.57	.60	.62	.58	.61	.63	.59	.62	.64	.60	.63	.65	.61	.64	.65	.61	.64	.66	.62	.65	.67	.62	.66	.67	.63	.66	.68	.63	.66	.68
6	.60	.63	.65	.61	.65	.66	.62	.64	.67	.63	.66	.68	.64	.67	.69	.64	.68	.69	.65	.68	.70	.65	.69	.70	.66	.69	.71	.66	.69	.71
7	.62	.66	.68	.64	.67	.69	.65	.68	.70	.65	.69	.71	.66	.69	.71	.66	.70	.72	.67	.70	.72	.67	.71	.73	.68	.71	.73	.68	.72	.74
8	.65	.68	.70	.66	.69	.71	.67	.70	.72	.67	.71	.73	.68	.71	.73	.68	.72	.74	.69	.72	.74	.69	.73	.75	.70	.73	.75	.70	.75	.76
9	.67	.70	.72	.68	.71	.73	.68	.72	.74	.69	.72	.74	.69	.73	.75	.70	.74	.76	.70	.74	.76	.71	.74	.76	.71	.75	.77	.71	.75	.77
10	.68	.72	.74	.69	.73	.75	.70	.73	.75	.70	.74	.76	.71	.74	.76	.71	.75	.77	.72	.75	.77	.72	.76	.78	.72	.76	.78	.72	.76	.78
11	.69	.73	.75	.70	.74	.76	.71	.75	.77	.71	.75	.77	.72	.76	.78	.72	.76	.78	.73	.76	.78	.73	.77	.79	.73	.77	.79	.73	.77	.79
12	.70	.74	.76	.71	.75	.77	.72	.76	.78	.72	.76	.78	.73	.77	.79	.73	.77	.79	.73	.77	.79	.74	.78	.80	.74	.78	.80	.74	.78	.80
13	.71	.75	.77	.72	.76	.78	.73	.76	.79	.73	.77	.79	.74	.77	.80	.74	.78	.80	.74	.78	.80	.75	.78	.81	.75	.79	.81	.75	.79	.81
14	.72	.76	.78	.73	.77	.79	.73	.77	.79	.74	.78	.80	.74	78	.80	.75	.79	.81	.75	.79	.81	.75	.79	.81	.75	.79	.82	.76	.80	.82
15	.73	.77	.79	.74	.77	.80	.74	.78	.80	.75	.78	.81	.75	.79	.81	.75	.79	.81	.76	.79	.82	.76	.80	.82	.76	.80	.82	.76	.80	.82
16	.74	.78	.80	.74	.78	.80	.75	.79	.81	.75	.79	.81	.75	.79	.82	.76	.80	.82	.76	.80	.82	.76	.80	.82	.77	.81	.83	.77	.81	.83
17	.74	.78	.80	.75	.79	.81	.75	.79	.81	.76	.80	.82	.76	.80	.82	.76	.80	.82	.77	.81	.83	.77	.81	.83	.77	.81	.83	.77	.81	.83
18	.75	.79	.81	.75	.79	.81	.76	.80	.82	.76	.80	.82	.76	.80	.83	.77	.81	.83	.77	.81	.83	.77	.81	.83	.77	.81	.84	.78	.82	.84

Note. $N_{therapists}$ = 10. Light gray indicates generalizability coefficients between .70 and .80; dark gray indicates generalizability coefficients higher than .80.

Table 4. Proportion of Variance Explained by the Model Terms in the Community Context.

Effect	*df*	*SS*	*MS*	Variance component	Variance percentage
t	2	0.92	0.46	.0027	4.5
p:t	4	1.13	0.28	.0091	15.3
s:p:t	77	4.87	0.06	.0217	35.9
r	1	0.26	0.26	.0031	5.0
tr	2	0.03	0.01	.0015	2.5
(p:t)r	4	0.18	0.05	.0025	4.1
(s:p:t)r,e	77	1.53	0.02	.0198	32.7

Note. df = degrees of freedom; *SS* = sum of squares; *MS* = mean square; t = therapist; p = patient; s = session; r = coder; e = error; n_t = 3; n_p = 2.25 (harmonic mean); n_s = 7.21 (harmonic mean); n_r = 2.

D-study below, adding more patients per therapist would have greatly benefited the community context.

D-Study. The D-study regarding the community context trial revealed even more extreme recommendations for good generalizability, as displayed in Table 5. Here, adding more raters or sessions would not make a considerable impact on generalizability: across all number of sessions, four raters only slightly outperformed two or three raters. In the community context, the most important factor in achieving adequate generalizability was the number of patients per therapist. In this context, two raters would produce an acceptable ρ_t^2 coefficient (≥.70), but only when there is a minimum of 12 patients per therapist and 11 sessions per patient.

Discussion

We tested the generalizability of the scores of an observational adherence instrument designed for ICBT for youth anxiety. We also conducted a D-study to determine optimal design for future studies using the instrument. We conducted these analyses on a data set drawn from two RCTs testing the same treatment program conducted in two different contexts: research and community. Overall, the initial G-study (Study 1) that included both treatment contexts together found that context accounted for a large proportion of variance. Differences between research and community contexts across patient, therapist, and agency characteristics are possible contributors to this finding and have been documented elsewhere (e.g., Ehrenreich-May et al., 2011; Southam-Gerow, Weisz, & Kendall, 2003). Accordingly, we conducted separate studies within the university-based research sample (Study 2) and the community-based sample (Study 3). We found that across both contexts, number of coded sessions and residual error were the largest contributors to explaining the variance in adherence scores; for the community context alone, the number of youth also accounted for nonignorable variance (nearly 16.0%). The D-study for each context suggested that more youth per therapist and a higher number of sessions per therapist were needed, with these requisites being higher in the community context. The D-study also suggested that adherence scores would become more stable if the number of coders was increased.

Although the observed generalizability coefficients were below the optimal level, they were in line with previous research. Dennhag et al. (2012) conducted a G-study regarding the number of sessions and patients required to create stable adherence and competence scores regarding cocaine dependence treatment. Across three types of treatment (self-expressive, cognitive, and individual drug counseling), they found weak adherence ρ_t^2 values of .44, .64, and .48, respectively. In the present study, the main issue that appears to be affecting ρ_t^2 was the residual error term: even after accounting for number of therapists, patients, sessions, coders, and their interactions, almost 30.0% of the variance in adherence scores was still unexplained. Though this is consistent with previous work (e.g., Dennhag et al., 2012), future work will help determine if this is a consistent finding.

In both contexts, considerably more sessions per youth and youth per therapist were needed to produce generalizable adherence scores than past research has suggested. In both the research and community contexts, up to 10 sessions were needed, given 12 youth per therapist and two coders to produce a generalizability coefficient of .70 or higher. Increasing the number of coders reduced the number of patients needed, especially for the research context. For example, with three coders, the research context could produce generalizable scores with nine youth; for the community context, 11 youth are needed. The efficiency of the research context becomes clearer when considering the effect of adding sessions. Even with 19 sessions per youth, the community context would always need at least 10 youth per therapist, even with four coders. However, in the research context, the number of youth per therapist could go as low as six with as few as 16 sessions per patient.

These findings present a somewhat pessimistic perspective on the efficiency of observational adherence measurement. Ratios of sessions per youth and youth per therapist at the level needed as suggested by these data are quite difficult to accomplish even in well-funded research endeavors. By including more coders, the ratios become somewhat better, but adding coders does not enhance efficiency. These estimates apply solely to the adherence instrument used in the present study, the CBAY-A, and it is possible that other adherence instruments might produce generalizable estimates with fewer sessions, youth, therapists, and coders. However, the suggested need for more recorded sessions and more therapists per youth is consistent with the handful of past studies (e.g., Dennhag et al., 2012). Furthermore,

Table 5. Generalizability Coefficients at the Therapist Level for Adherence Ratings Across Various Numbers of Patients, Sessions, and Coders in the Community Trial.

	Sessions														
	10			11			12			13			14		
	Coders			Coders			Coders			Coders			Coders		
Patients	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4
9	.64	.66	.66	.65	.66	.67	.65	.66	.67	.66	.67	.67	.66	.67	.68
10	.67	.68	.69	.67	.68	.69	.68	.69	.69	.68	.69	.70	.68	.69	.70
11	.69	.70	.71	.69	.70	.71	.70	.71	.71	.70	.71	.72	.70	.71	.72
12	.71	.72	.72	.71	.72	.73	.71	.73	.73	.72	.73	.73	.72	.73	.74
13	.72	.73	.74	.73	.74	.74	.73	.74	.75	.73	.74	.75	.74	.75	.75
14	.74	.75	.75	.74	.75	.76	.74	.75	.76	.75	.76	.76	.75	.76	.76
15	.75	.76	.77	.75	.76	.77	.76	.77	.77	.76	.77	.77	.76	.77	.78
16	.76	.77	.78	.77	.78	.78	.77	.78	.78	.77	.78	.79	.77	.78	.79
17	.77	.78	.79	.78	.79	.79	.78	.79	.79	.78	.79	.80	.78	.79	.80
18	.78	.79	.80	.79	.80	.80	.79	.80	.80	.79	.80	.80	.79	.80	.81
19	.79	.80	.81	.80	.80	.81	.80	.81	.81	.80	.81	.81	.80	.81	.82
20	.80	.81	.81	.80	.81	.82	.81	.81	.82	.81	.82	.82	.81	.82	.82

	Sessions														
	15			16			17			18			19		
	Coders			Coders			Coders			Coders			Coders		
Patients	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4
9	.66	.67	.68	.66	.67	.68	.66	.67	.68	.67	.68	.68	.67	.68	.69
10	.68	.70	.70	.69	.70	.70	.69	.70	.70	.69	.70	.71	.69	.70	.71
11	.70	.72	.72	.71	.72	.72	.71	.72	.72	.71	.72	.73	.71	.72	.73
12	.72	.73	.74	.72	.73	.74	.73	.74	.74	.73	.74	.74	.73	.74	.74
13	.74	.75	.75	.74	.75	.75	.74	.75	.76	.74	.75	.76	.75	.75	.76
14	.75	.76	.77	.75	.76	.77	.76	.77	.77	.76	.77	.77	.76	.77	.77
15	.76	.77	.78	.77	.78	.78	.77	.78	.78	.77	.78	.78	.77	.78	.78
16	.78	.79	.79	.78	.79	.79	.78	.79	.79	.78	.79	.79	.78	.79	.80
17	.79	.80	.80	.79	.80	.80	.79	.80	.80	.79	.80	.80	.79	.80	.80
18	.80	.80	.81	.80	.81	.81	.80	.81	.81	.80	.81	.81	.80	.81	.81
19	.80	.81	.82	.81	.81	.82	.81	.82	.82	.81	.82	.82	.81	.82	.82
20	.81	.82	.82	.81	.82	.83	.82	.82	.83	.82	.82	.83	.82	.83	.83

Note. $N_{therapists} = 3$. Light gray indicates generalizability coefficients between .70 and .80; dark gray indicates generalizability coefficients higher than the .80 threshold prior to rounding.

studies from other research areas have suggested that sample size needs for observational research are greater than might be anticipated and/or feasible (e.g., Gage, Prykanowski, & Hirn, 2014; Mashburn et al., 2014). Accordingly, our results contribute to the emerging conclusion that traditional sampling approaches (e.g., 20.0% of a sample) for observational measurement may lead to unreliable adherence estimates.

Given these findings, it may be useful to delineate some practical applications of G-study/D-study. The accumulation of scientific findings that leads to theory building relies on measurements that are generalizable across studies. Greater generalizability in assessment tools will lead to greater confidence that separate findings can be successfully integrated into a more coherent body of research. Furthermore, by establishing that the findings of one particular study can be generalized to diverse settings, samples, or research designs, a careful G-study/D-study can increase the likelihood of successful replication (Brennan, 2010).

When testing psychosocial treatments, gauging the extent to which the independent variable was manipulated as intended is an important task (Perepletchikova & Kazdin, 2005). From a practical perspective, the generalizability of such measurements across trials may be less important when different treatment programs are being conducted. However, from a broader public health perspective, there is a need to consider how to measure treatment integrity beyond discrete investigator-led trials. As stakeholders attempt to take treatment programs to scale across diverse service contexts the need for a integrity measurements that generalize across contexts becomes an important issue (e.g., McLeod, Southam-Gerow, Bair, Rodriguez, & Smith, 2013).

Our findings raise questions about the practicality of the gold standard observational integrity instruments for research that stresses the importance of generalizability. The CBAY-A was modeled off of exemplar observational treatment integrity instruments (e.g., Hogue, Rowe, Liddle, & Turner, 1994; Sifry et al., 1994). Though these observational instruments are very thorough, they are complex and require extensive training. Our findings suggest that it may be important to investigate if shorter observational integrity instruments that are easier to use produce more dependable estimates. The development of "pragmatic" instruments that are brief and easy to use represents an important goal of implementation research (Glasgow & Riley, 2013). Reducing the number of items and simplifying the coding process may help reduce the number of sessions needed to produce a dependable estimate, which would be important for researchers interested in producing generalizable treatment integrity estimates.

Considering these findings, it may also be important to consider different ways of measuring treatment integrity. Though therapist-report instruments have generally not converged with observational instruments in past studies, (e.g., Chapman, McCart, Letourneau, & Sheidow, 2013; Hogue, Dauber, & Henderson, 2014), some positive findings have been reported (e.g., Ward et al., 2013). There may be ways to improve the score validity of therapist-report integrity instruments, including involving them in instrument development, training them in the use the instrument, and decoupling reporting integrity from therapist evaluation and/or payment. In addition, some studies have demonstrated that scores on patient-report of integrity converge with therapist-report ratings as well as predict treatment outcomes (e.g., Ellis, Naar-King, Templin, Frey, & Cunningham, 2007; Schoenwald et al., 2009). That said, patient-report poses a challenge for children (age < 12 years), but may be worth exploring.

One factor related to interpreting findings pertaining to the number of coded sessions bears mentioning. Growing evidence suggests that adherence scores may systematically change over the course of treatment (see, e.g., Boswell et al., 2013; Smith et al., 2016). If this is accurate, it poses a potential problem for interpreting findings related to the D-study. Generalizability theory has its roots in classical test theory, so it rests on the assumption that there exists a true score for each facet (e.g., Brennan, 2010). D-studies are designed to estimate how many observations are needed to achieve a reliable estimate of the true score for particular facets. However, if adherence scores systematically change over treatment, then variation in scores from session to session are not due to error. For example, it could mean that scores vary across phases of treatment (e.g., skill building and exposure phases) as opposed to a course of treatment. Thus, the findings may not apply to the number of coded sessions needed to achieve a reliable estimate. In other research areas (e.g., school observations, medical settings), it will be worth considering whether such variation over time is expected and adjust sampling accordingly.

Study limitations warrant attention. First, our sample size was still smaller than optimal: we were not able to code every session held because not all sessions were recorded. Numerous reasons account for these omissions, including technical errors (e.g., recording not audible) and user error (e.g., therapist neglected to press record button). Amplifying this problem was the fact that we limited the sample to therapists who saw more than one patient to examine variance regarding the number of patients per therapist. This had a particularly deleterious impact on the community sample, focusing on seven patients and three therapists. The sample size was also compounded by the overall unbalanced design (i.e., the presence of different *n* values among the facets). This type of data structure introduced interpretation difficulties and disallowed the use of certain statistical methods that are commonly used when analyzing variance components from balanced designs. As a result, our findings related to the community sample should be viewed skeptically, and future studies of CBT adherence should strive for balanced designs.

Second, the conclusions are limited to consideration of adherence measurement within the context of ICBT for youth anxiety. It will be important to replicate these findings with other populations and treatment approaches, as it is conceivable that generalizability of adherence scores vary by treatment type. Finally, it is important to note that residual error accounted for considerable variance in both contexts, suggesting that there are other variables we did not assess that account for variance in adherence scores. Among the many candidates for this variance that we did not assess include therapist expertise with CBT, variability in difficulty of program content (e.g., some elements of CBT may be more challenging to deliver with adherence than others), patient or therapist diversity, and patient clinical complexity.

Despite the limitations, the current study is the first G- and D-study we have seen reported in the youth anxiety literature and includes the largest sample of coded recordings that we have seen. Our findings suggest that generalizable measurement of adherence requires a large number of sessions, youth, and therapists, underscoring the need to develop more efficient methods of measuring this important construct.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preparation of this article was supported in part by a grant from the National Institute of Mental Health Grant (RO1 MH086529; McLeod & Southam-Gerow).

ORCID iDs

Michael A. Southam-Gerow https://orcid.org/0000-0002-4545-6752
John R. Weisz https://orcid.org/0000-0002-5560-6814

References

Barber, J. P., Foltz, C., Crits-Christoph, P., & Chittams, J. (2004). Therapists' adherence and competence and treatment discrimination in the NIDA Collaborative Cocaine Treatment Study. *Journal of Clinical Psychology*, *60*, 29-41. doi:10.1002/jclp.10186

Boswell, J. F., Gallagher, M. W., Sauer-Zavala, S. E., Bullis, J., Gorman, J. M., Shear, M. K., . . . Barlow, D. H. (2013). Patient characteristics and variability in adherence and competence in cognitive-behavioral therapy for panic disorder. *Journal of Consulting and Clinical Psychology*, *81*, 443-454. doi:10.1037/a0031437

Brennan, R. L. (2001a). *Generalizability theory*. New York, NY: Springer-Verlag.

Brennan, R. L. (2001b). *Manual for urGENOVA*. Iowa City: Iowa Testing Programs, University of Iowa.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*, 1-21. doi:10.1080/08957347.2011.532417

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*, 13-35. doi:10.1016/j.jsp.2013.11.008

Bruhn, A. L., Hirsch, S. E., & Lloyd, J. W. (2015). Treatment integrity in school-wide programs: A review of the literature (1993-2012). *Journal of Primary Prevention*, *36*, 335-349. doi:10.1007/s10935-015-0400-9

Cardinet, J., Johnson, J., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Routledge.

Chapman, J. E., McCart, M. R., Letourneau, E. J., & Sheidow, A. J. (2013). Comparison of youth, caregiver, therapist, trained, and treatment expert raters of therapist adherence to a substance abuse treatment protocol. *Journal of Consulting and Clinical Psychology*, *81*, 674-680. doi:10.1037/a0033021

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-290. doi:10.1037/1040-3590.6.4.284

Crits-Christoph, P., Johnson, J., Gallop, R., Gibbons, M. B. C., Ring-Kurtz, S., Hamilton, J. L., & Tu, X. (2011). A generalizability theory analysis of group process ratings in the treatment of cocaine dependence. *Psychotherapy Research*, *21*, 252-266. doi:10.1080/10503307.2010.551429

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334. doi:10.1007/BF02310555

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*, 137-163. doi:10.1111/j.2044-8317.1963.tb00206.x

Dennhag, I., Gibbons, M. B. C., Barber, J. P., Gallop, R., & Crits-Christoph, P. (2012). How many treatment sessions and patients are needed to create a stable score of adherence and competence in the treatment of cocaine dependence? *Psychotherapy Research*, *22*, 475-488. doi:10.1080/10503307.2012.674790

Ehrenreich-May, J., Southam-Gerow, M. A., Hourigan, S. E., Wright, L. R., Pincus, D. B., & Weisz, J. R. (2011). Characteristics of anxious and depressed youth seen in two different clinical contexts. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*, 398-411. doi:10.1007/s10488-010-0328-6

Ellis, D. A., Naar-King, S., Templin, T., Frey, M. A., & Cunningham, P. B. (2007). Improving health outcomes among youth with poorly controlled Type I diabetes: The role of treatment fidelity in a randomized clinical trial of multisystemic therapy. *Journal of Family Psychology*, *21*, 363-371.

Gage, N. A., Prykanowski, D., & Hirn, R. (2014). Increasing reliability of direct observation measurement approaches in emotional and/or behavioral disorders research using generalizability theory. *Behavioral Disorders*, *39*, 228-244.

Glasgow, R. E., & Riley, W. T. (2013). Pragmatic measures: What are they and why we need them. *American Journal of Prevention Medicine*, *45*, 237-243. doi:10.1016/j.amepre.2013.03.010

Goense, P., Boendermaker, L., van Yperen, T., Stams, G. J., & van Laar, J. (2014). Implementation of treatment integrity procedures: An analysis of outcome studies of youth interventions targeting externalizing behavioral problems. *Zeitschrift für Psychologie*, *222*, 12-21. doi:10.1027/2151-2604/a000161

Gresham, F. M., Dart, E. H., & Collins, T. A. (2017). Generalizability of multiple measures of treatment integrity: Comparisons among direct observation, permanent products, and self-report. *School Psychology Review*, *46*, 108-121. doi:10.17105/SPR46-1.108-121

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*, 56-64. doi:10.3102/0013189X12437203

Hogue, A., Dauber, S., Chinchilla, P., Fried, A., Henderson, C., Inclan, J., . . . Liddle, H. A. (2008). Assessing fidelity in individual and family therapy for adolescent substance abuse. *Journal of Substance Abuse Treatment*, *35*, 137-147. doi:10.1016/j.jsat.2007.09.002

Hogue, A., Dauber, S., & Henderson, C. E. (2014). Therapist self-report of evidence-based practices in usual care for adolescent behavior problems: Factor and construct validity. *Administration and Policy in Mental Health and Mental Health Services Research*, *41*, 126-139. doi:10.1007/s10488-012-0442-8

Hogue, A., Dauber, S., Lichvar, E., Bobek, M., & Henderson, C. E. (2015). Validity of therapist self-report ratings of fidelity to evidence-based practices for adolescent behavior problems: Correspondence between therapists and observers. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*, 229-243. doi:10.1007/s10488-014-0548-2

Hogue, A., Liddle, H. A., & Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy: Theory, Research, Practice, Training*, *33*, 332-345. doi:10.1037/0033-3204.33.2.332

Hogue, A., Rowe, C., Liddle, H. A., & Turner, R. (1994). *Scoring manual for the Therapist Behavior Rating Scale (TBRS)* (Unpublished manuscript). Center for Research on Adolescent Drug Abuse, Temple University, Philadelphia, PA.

Huey, S. J., Henggeler, S. W., Brondino, M. J., & Pickrel, S. G. (2000). Mechanisms of change in multisystemic therapy: Reducing delinquent behavior through therapist adherence and improved family and peer functioning. *Journal of Consulting and Clinical Psychology*, *68*, 451-467. doi:10.1037/0022-006X.68.3.451

Kendall, P. C. (1994). Treating anxiety disorders in children: Results of a randomized clinical trial. *Journal of Consulting and Clinical Psychology*, *62*, 100-110. doi:10.1037/0022-006X.62.1.100

Kendall, P. C., Flannery-Schroeder, E., Panichelli-Mindel, S. M., Southam-Gerow, M., Henin, A., & Warman, M. (1997). Therapy for youths with anxiety disorders: A second randomized clinical trial. *Journal of Consulting and Clinical Psychology*, *65*, 366-380. doi:10.1037/0022-006X.65.3.366

Kendall, P. C., & Hedtke, K. A. (2006a). *Cognitive-behavioral therapy for anxious children: Therapist manual*. Ardmore, PA: Workbook.

Kendall, P. C., & Hedtke, K. A. (2006b). *The coping cat workbook* (2nd ed.). Ardmore, PA: Workbook.

Kendall, P. C., Hudson, J. L., Gosch, E., Flannery-Schroeder, E., & Suveg, C. (2008). Cognitive-behavioral therapy for anxiety disordered youth: A randomized clinical trial evaluating child and family modalities. *Journal of Consulting and Clinical Psychology*, *76*, 282-297. doi:10.1037/0022-006X.76.2.282

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London, England: Routledge.

Konge, L., Vilmann, P., Clementsen, P., Annema, J. T., & Ringsted, C. (2012). Reliable and valid assessment of competence in endoscopic ultrasonography and fine-needle aspiration for mediastinal staging of non-small cell lung cancer. *Endoscopy*, *44*, 928-933. doi:10.1055/s-0032-1309892

Margolin, G., Oliver, P., Gordis, E., O'Hearn, H., Medina, A., Ghosh, C., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child and Family Psychology Review*, *1*, 195-213. doi:10.1023/A:1022608117322

Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2014). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science*, *15*, 146-155. doi:10.1007/s11121-012-0357-3

McLeod, B. D., Southam-Gerow, M. A., Bair, C. E., Rodríguez, A., & Smith, M. M. (2013). Making a case for treatment integrity as a quality indicator. *Clinical Psychology: Science & Practice*, *20*, 14-32.

McLeod, B. D., Southam-Gerow, M. A., & Weisz, J. R. (2009). Conceptual and methodological issues in treatment integrity measurement. *School Psychology Review*, *38*, 541-546.

Nielsen, D. G., Jensen, S. L., & O'Neill, L. (2015). Clinical assessment of transthoracic echocardiography skills: A generalizability study. *BMC Medical Education*, *15*, 9. doi:10.1186/s12909-015-0294-5

O'Brien, R. M. (1995). Generalizability coefficients are reliability coefficients. *Quality & Quantity*, *29*, 421-428.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*, 365-383. doi:10.1093/clipsy.bpi045

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology*, *75*, 829-841. doi:10.1037/0022-006X.75.6.829

Preuss, R. A. (2013). Using generalizability theory to develop clinical assessment protocols. *Physical Therapy*, *93*, 562-569. doi:10.2522/ptj.20120368

Sanetti, L. M. H., & Kratochwill, T. R. (2014). *Treatment integrity: A foundation for evidence-based practice in applied psychology*. Washington, DC: American Psychological Association.

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L, Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*, 32-43. doi:10.1007/s10488-010-0321-0

Schoenwald, S. K., Sheidow, A. J., & Chapman, J. E. (2009). Clinical supervision in treatment transport: Effects on adherence and outcomes. *Journal of Consulting and Clinical Psychology*, *77*, 410-421. doi:10.1037/a0013788

Sebok, S. S., Luu, K., & Klinger, D. A. (2014). Psychometric properties of the multiple mini-interview used for medical admissions: Findings from generalizability and Rasch analyses. *Advances in Health Sciences Education*, *19*, 71-84. doi:10.1007/s10459-013-9463-7

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Sholomskas, D. E., Syracuse-Siewert, G., Rounsaville, B. J., Ball, S. A., Nuro, K. F., & Carroll, K. M. (2005). We don't train in vain: A dissemination trial of three strategies of training clinicians in cognitive-behavioral therapy. *Journal of Consulting and Clinical Psychology*, *73*, 106-115. doi:10.1037/0022-006X.73.1.106

Sifry, R. L., Nuro, K. F., Ball, S., Corvino, J., Bisighini, R. M., & Carroll, K. M. (1994). *Rater's manual for Yale psychotherapy development center treatment rating scale* (Unpublished manuscript). Yale University, New Haven, CT.

Smith, M. M., McLeod, B. D., Southam-Gerow, M. A., Jensen-Doss, A., Kendall, P. C., & Weisz, J. R. (2016). Does the delivery of CBT for youth anxiety differ across research and practice settings? *Behavior Therapy*, *48*, 501-516. doi:10.1016/j.beth.2016.07.004

Southam-Gerow, M. A., & McLeod, B. D. (2013). Advances in applying treatment integrity research for dissemination and implementation science: Introduction to special issue. *Clinical Psychology: Science and Practice*, *20*, 1-13. doi:10.1111/cpsp.12019

Southam-Gerow, M. A., McLeod, B. D., Arnold, C. C., Rodríguez, A., Cox, J. R., Reise, S. P., . . . Kendall, P. C. (2016). Initial development of a treatment adherence measure for cognitive–behavioral therapy for child anxiety. *Psychological Assessment*, *28*, 70-80. doi:10.1037/pas0000141

Southam-Gerow, M. A., Weisz, J. R., Chu, B. C., McLeod, B. D., Gordis, E. B., & Connor-Smith, J. K. (2010). Does CBT for youth anxiety outperform usual care in community clinics? An initial effectiveness test. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*, 1043-1052. doi:10.1016/j.jaac.2010.06.009

Southam-Gerow, M. A., Weisz, J. R., & Kendall, P. C. (2003). Youth with anxiety disorders in research and service clinics: Examining client differences and similarities. *Journal of Clinical Child & Adolescent Psychology*, *32*, 375-385.

Todsen, T., Tolsgaard, M. G., Olsen, B. H., Henriksen, B. M., Hillingsø, J. G., Konge, L., . . . Ringsted, C. (2015). Reliable and valid assessment of point-of-care ultrasonography. *Annals of Surgery*, *261*, 309-315. doi:10.1097/SLA.0000000000000552

Ward, A., Regan, J., Chorpita, B., Starace, N., Rodriguez, A., Okamura, K., . . . Weisz, J. R. (2013). Tracking evidence-based practice with youth: Validity of the MATCH and Standard Manual Consultation Records. *Journal of Clinical Child & Adolescent Psychology*, *42*, 44-55. doi:10.1080/15374416.2012.700505

Wass, V., van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *Lancet*, *357*, 945-949.

Wasserman, R. H., Levy, K. N., & Loken, E. (2009). Generalizability theory in psychotherapy research: The impact of multiple sources of variance on the dependability of psychotherapy process ratings. *Psychotherapy Research*, *19*, 397-408. doi:10.1080/10503300802579156

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *78*, 200-211. doi:10.1037/a0018912