Generalizability and Decision Studies of a Treatment Adherence Instrument

Michael A. Southam-Gerow, Wes Bonifay, Bryce D. McLeod, Julia R. Cox, Stephanie Violante, Philip C. Kendall, and John R. Weisz

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 therapist 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
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, Clemens, 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).
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 ($\hat{\rho}_t^2$) across contexts was computed using the following formula (extended from a less complex model presented in Brennan [2001a]):

$$\hat{\rho}_t^2 = \frac{\sigma_t^2 + \sigma_{tx}^2}{\sigma_t^2 + \sigma_{tx}^2 + \frac{1}{n_p}\sigma_{txt}\frac{1}{n_p}\sigma_{txt} + \frac{1}{n}\sigma_{tx} + \frac{1}{n_t}\sigma_{tx} + \frac{1}{n_{rt}}\sigma_{rt} + \frac{1}{n_{rt}}\sigma_{rt} + \frac{1}{n_{rtc}}\sigma_{rtc}}$$

where $\sigma_t^2$ is the variance attributed to contexts, $\sigma_{tx}^2$ is the variance attributed to therapists nested within contexts, $\sigma_{txt}^2$ is the variance attributed to patients nested within
therapists nested within contexts, \( n_p \) is the number of patients, \( \sigma^2_{p t c} \) is the variance attributed to sessions nested within patients nested within therapists nested within contexts, \( n_t \) is the number of therapists, \( \sigma^2_{t c r} \) 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 \( \sigma^2_{r} \) 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 \( \sigma^2_c \) (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 \( n_t \) in Equation 1 were computed using the harmonic mean of patients and therapists, respectively, within each context. The generalizability coefficient \( \rho^2_c \) in Equation 1 (and that of the research and community contexts in isolation) is analogous to the classical test theory index of reliability (coefficient \( \alpha \); 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 (c:r), 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 1. Proportion of Variance Explained by the Model Terms Across Both the Research and Community Contexts.**

<table>
<thead>
<tr>
<th>Effect</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>Variance component</th>
<th>Variance percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>1</td>
<td>51.90</td>
<td>51.90</td>
<td>.179</td>
<td>49.5</td>
</tr>
<tr>
<td>tc</td>
<td>11</td>
<td>12.78</td>
<td>1.16</td>
<td>.007</td>
<td>2.1</td>
</tr>
<tr>
<td>p:tc</td>
<td>39</td>
<td>16.60</td>
<td>0.43</td>
<td>.008</td>
<td>2.4</td>
</tr>
<tr>
<td>s:p:tc</td>
<td>498</td>
<td>129.81</td>
<td>0.26</td>
<td>.103</td>
<td>28.6</td>
</tr>
<tr>
<td>r</td>
<td>1</td>
<td>2.90</td>
<td>2.90</td>
<td>.005</td>
<td>1.5</td>
</tr>
<tr>
<td>c:r</td>
<td>1</td>
<td>0.03</td>
<td>0.03</td>
<td>.001</td>
<td>0.2</td>
</tr>
<tr>
<td>(tc):r</td>
<td>11</td>
<td>1.60</td>
<td>0.15</td>
<td>.002</td>
<td>0.7</td>
</tr>
<tr>
<td>(p:tc):r</td>
<td>11</td>
<td>1.86</td>
<td>0.05</td>
<td>.001</td>
<td>0.2</td>
</tr>
<tr>
<td>(s:p:tc):r:e</td>
<td>39</td>
<td>27.20</td>
<td>0.05</td>
<td>.055</td>
<td>15.2</td>
</tr>
</tbody>
</table>

Note. \( df \) = degrees of freedom; \( SS = \) sum of squares; \( MS = \) mean square; \( c = \) context; \( t = \) therapist; \( p = \) patient; \( s = \) session; \( r = \) coder; \( e = \) error; \( n_c = 13; n_p = 3.02 \) (harmonic mean); \( n_t = 9.01 \) (harmonic mean); \( n_r = 2 \).
Inputting the values from Table 4 into the modified (context-free) version of Equation 1 reveals the generalizability coefficient associated with the research context: \( \rho_t^2 = .48 \).

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 (\( \geq .70 \)). The most impactful facet was the number of raters. With only two raters, \( \rho_t^2 \) would not reach a good level of generalizability (\( \geq .80 \)), 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 \( \times \) rater (tr) interaction accounted for 2.5% of the variance and the patient (within therapists) \( \times \) 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 trivial amounts of variance.

In the research context, \( n_t \) 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: \( \rho_t^2 = .48 \).

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 was attributed to residual error, as 29.9% of the variance in adherence scores. The two interaction terms—therapist \( \times \) rater (tr), and patient (within therapists) \( \times \) 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 research and community contexts, \( n_t \) ranged from 2 to 9 patients (harmonic mean = 3.02) and \( n_s \) ranged from 3 to 14 sessions (harmonic mean = 9.01). Plugging in the values from Table 2 into the modified (context-free) version of Equation 1 reveals the generalizability coefficient associated with the community context: \( \rho_t^2 = .27 \). 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 2. Proportion of Variance Explained by the Model Terms in the Research Context.

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

Note. df = degrees of freedom; SS = sum of squares; MS = mean square; \( t = \) therapist; \( p = \) patient; \( s = \) session; \( r = \) coder; \( e = \) error; \( n_t = 10; n_s = 3.37 \) (harmonic mean); \( n_p = 3.37 \) (harmonic mean); \( n_r = 2 \).
<table>
<thead>
<tr>
<th>Sessions</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.57 .60 .62</td>
<td>.58 .61 .63</td>
<td>.59 .62 .64</td>
<td>.60 .63 .65</td>
<td>.61 .64 .66</td>
<td>.62 .65 .67</td>
<td>.62 .66 .67</td>
<td>.63 .66 .68</td>
<td>.63 .66 .68</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>.60 .63 .65</td>
<td>.61 .65 .66</td>
<td>.62 .64 .67</td>
<td>.63 .66 .68</td>
<td>.64 .67 .69</td>
<td>.64 .68 .69</td>
<td>.65 .68 .70</td>
<td>.65 .69 .70</td>
<td>.66 .69 .71</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>.62 .66 .68</td>
<td>.64 .67 .69</td>
<td>.65 .68 .70</td>
<td>.65 .69 .71</td>
<td>.66 .69 .71</td>
<td>.66 .70 .72</td>
<td>.67 .70 .72</td>
<td>.67 .71 .73</td>
<td>.68 .71 .73</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>.65 .68 .70</td>
<td>.66 .69 .71</td>
<td>.67 .70 .72</td>
<td>.67 .71 .73</td>
<td>.68 .71 .73</td>
<td>.68 .72 .74</td>
<td>.69 .72 .74</td>
<td>.69 .73 .75</td>
<td>.70 .73 .75</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>.67 .70 .72</td>
<td>.68 .71 .73</td>
<td>.68 .72 .74</td>
<td>.69 .72 .74</td>
<td>.69 .73 .75</td>
<td>.70 .74 .76</td>
<td>.70 .74 .76</td>
<td>.71 .74 .76</td>
<td>.71 .75 .77</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>.68 .72 .74</td>
<td>.69 .73 .75</td>
<td>.70 .73 .75</td>
<td>.70 .74 .76</td>
<td>.71 .74 .76</td>
<td>.71 .75 .77</td>
<td>.72 .75 .77</td>
<td>.72 .76 .78</td>
<td>.72 .76 .78</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>.69 .73 .75</td>
<td>.70 .74 .76</td>
<td>.71 .75 .77</td>
<td>.71 .75 .77</td>
<td>.72 .76 .78</td>
<td>.72 .76 .78</td>
<td>.73 .76 .78</td>
<td>.73 .77 .79</td>
<td>.73 .77 .79</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>.70 .74 .76</td>
<td>.71 .75 .77</td>
<td>.72 .76 .78</td>
<td>.72 .76 .78</td>
<td>.73 .77 .79</td>
<td>.73 .77 .79</td>
<td>.73 .77 .79</td>
<td>.74 .78 .80</td>
<td>.74 .78 .80</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>.71 .75 .77</td>
<td>.72 .76 .78</td>
<td>.73 .76 .79</td>
<td>.73 .77 .79</td>
<td>.74 .77 .80</td>
<td>.74 .78 .80</td>
<td>.74 .78 .80</td>
<td>.75 .78 .81</td>
<td>.75 .78 .81</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>.72 .76 .78</td>
<td>.73 .77 .79</td>
<td>.73 .77 .79</td>
<td>.74 .78 .80</td>
<td>.74 .78 .80</td>
<td>.75 .79 .81</td>
<td>.75 .79 .81</td>
<td>.75 .79 .81</td>
<td>.75 .79 .81</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>.73 .77 .79</td>
<td>.74 .77 .80</td>
<td>.74 .78 .80</td>
<td>.74 .78 .80</td>
<td>.75 .79 .81</td>
<td>.75 .79 .81</td>
<td>.76 .79 .82</td>
<td>.76 .80 .82</td>
<td>.76 .80 .82</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>.74 .78 .80</td>
<td>.74 .78 .80</td>
<td>.75 .79 .81</td>
<td>.75 .79 .81</td>
<td>.75 .79 .81</td>
<td>.76 .79 .82</td>
<td>.76 .80 .82</td>
<td>.76 .80 .82</td>
<td>.77 .81 .83</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>.74 .78 .80</td>
<td>.75 .79 .81</td>
<td>.75 .79 .81</td>
<td>.76 .80 .82</td>
<td>.76 .80 .82</td>
<td>.76 .80 .82</td>
<td>.77 .81 .83</td>
<td>.77 .81 .83</td>
<td>.77 .81 .83</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>.75 .79 .81</td>
<td>.75 .79 .81</td>
<td>.76 .80 .82</td>
<td>.76 .80 .82</td>
<td>.76 .80 .82</td>
<td>.77 .81 .83</td>
<td>.77 .81 .83</td>
<td>.77 .81 .83</td>
<td>.77 .81 .83</td>
</tr>
</tbody>
</table>

Note. $N_{\text{therapists}} = 10$. Light gray indicates generalizability coefficients between .70 and .80; dark gray indicates generalizability coefficients higher than .80.
in the Community Context.
the community context alone, the number of youth also
tributors to explaining the variance in adherence scores; for
of coded sessions and residual error were the largest con-
ple (Study 3). We found that across both contexts, number
research sample (Study 2) and the community-based sam-
Southam-Gerow, Weisz, & Kendall, 2003). Accordingly,
documented elsewhere (e.g., Ehrenreich-May et al., 2011;
are possible contributors to this finding and have been doc-
together found that context accounted for a large proportion
G-study (Study 1) that included both treatment contexts
ent contexts: research and community. Overall, the initial
anxiety. We also conducted a D-study to determine optimal
testing the same treatment program conducted in two differ-
ment for future studies using the instrument. We con-
duced these analyses on a data set drawn from two RCTs
testing the same treatment program conducted in two differ-
context, the number of patients per therapist could go
therapist and a higher number of sessions per therapist were
requisite being higher in the community context. The D-study also suggested that adherence scores
would become more stable if the number of coders was

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 $\rho^2$ values of .44, .64, and .48,
respectively. In the present study, the main issue that appears
to be affecting $\rho^2$ was the residual error term: even after
accounting for number of therapists, patients, sessions, cod-
ers, and their interactions, almost 30.0% of the variance in
adherence scores was still unexplained. Though this is consis-
tent 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 generaliz-
able adherence scores than past research has suggested. In
both the research and community contexts, up to 10 ses-
sions 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 pro-
duce generalizable scores with nine youth; for the com-

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

<table>
<thead>
<tr>
<th>Effect</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>Variance component</th>
<th>Variance percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>2</td>
<td>0.92</td>
<td>0.46</td>
<td>0.0027</td>
<td>4.5</td>
</tr>
<tr>
<td>$p:t$</td>
<td>4</td>
<td>1.13</td>
<td>0.28</td>
<td>0.0091</td>
<td>15.3</td>
</tr>
<tr>
<td>$s:p:t$</td>
<td>77</td>
<td>4.87</td>
<td>0.06</td>
<td>0.0217</td>
<td>35.9</td>
</tr>
<tr>
<td>$r$</td>
<td>1</td>
<td>0.26</td>
<td>0.26</td>
<td>0.0031</td>
<td>5.0</td>
</tr>
<tr>
<td>$t$</td>
<td>2</td>
<td>0.03</td>
<td>0.01</td>
<td>0.0015</td>
<td>2.5</td>
</tr>
<tr>
<td>$(p:t)r$</td>
<td>4</td>
<td>0.18</td>
<td>0.05</td>
<td>0.0025</td>
<td>4.1</td>
</tr>
<tr>
<td>$(s:p:t)r,e$</td>
<td>77</td>
<td>1.53</td>
<td>0.02</td>
<td>0.0198</td>
<td>32.7</td>
</tr>
</tbody>
</table>

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); $n_r = 2$.

Discussion

We tested the generalizability of the scores of an observa-
tional 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 con-
duced these analyses on a data set drawn from two RCTs
testing the same treatment program conducted in two differ-
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 $\rho^2$ coefficient ($\geq$70), but only when there is a
minimum of 12 patients per therapist and 11 sessions per

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 achiev-
ing adequate generalizability was the number of patients
per therapist. In this context, two raters would produce an
acceptable $\rho^2$ coefficient ($\geq$70), but only when there is a
minimum of 12 patients per therapist and 11 sessions per

These findings present a somewhat pessimistic perspec-
tive on the efficiency of observational adherence measure-
ment. Ratios of sessions per youth and youth per therapist at
the level needed as suggested by these data are quite diffi-
cult to accomplish even in well-funded research endeavors.
By including more coders, the ratios become somewhat bet-
ter, 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 esti-
mates 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.

<table>
<thead>
<tr>
<th>Sessions</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>.64</td>
<td>.66</td>
<td>.66</td>
<td>.65</td>
<td>.66</td>
<td>.67</td>
<td>.66</td>
<td>.67</td>
<td>.67</td>
<td>.66</td>
</tr>
<tr>
<td>10</td>
<td>.67</td>
<td>.68</td>
<td>.69</td>
<td>.68</td>
<td>.69</td>
<td>.69</td>
<td>.68</td>
<td>.69</td>
<td>.70</td>
<td>.68</td>
</tr>
<tr>
<td>11</td>
<td>.69</td>
<td>.70</td>
<td>.71</td>
<td>.69</td>
<td>.70</td>
<td>.71</td>
<td>.70</td>
<td>.71</td>
<td>.72</td>
<td>.70</td>
</tr>
<tr>
<td>12</td>
<td>.71</td>
<td>.72</td>
<td>.72</td>
<td>.71</td>
<td>.72</td>
<td>.73</td>
<td>.72</td>
<td>.73</td>
<td>.74</td>
<td>.72</td>
</tr>
<tr>
<td>13</td>
<td>.72</td>
<td>.73</td>
<td>.74</td>
<td>.73</td>
<td>.74</td>
<td>.74</td>
<td>.73</td>
<td>.74</td>
<td>.75</td>
<td>.74</td>
</tr>
<tr>
<td>14</td>
<td>.74</td>
<td>.75</td>
<td>.75</td>
<td>.74</td>
<td>.75</td>
<td>.76</td>
<td>.75</td>
<td>.76</td>
<td>.77</td>
<td>.76</td>
</tr>
<tr>
<td>15</td>
<td>.75</td>
<td>.76</td>
<td>.77</td>
<td>.75</td>
<td>.76</td>
<td>.77</td>
<td>.76</td>
<td>.77</td>
<td>.78</td>
<td>.77</td>
</tr>
<tr>
<td>16</td>
<td>.76</td>
<td>.77</td>
<td>.78</td>
<td>.77</td>
<td>.78</td>
<td>.78</td>
<td>.77</td>
<td>.78</td>
<td>.79</td>
<td>.78</td>
</tr>
<tr>
<td>17</td>
<td>.77</td>
<td>.78</td>
<td>.79</td>
<td>.78</td>
<td>.79</td>
<td>.79</td>
<td>.78</td>
<td>.79</td>
<td>.80</td>
<td>.79</td>
</tr>
<tr>
<td>18</td>
<td>.78</td>
<td>.79</td>
<td>.80</td>
<td>.79</td>
<td>.80</td>
<td>.80</td>
<td>.79</td>
<td>.80</td>
<td>.80</td>
<td>.79</td>
</tr>
<tr>
<td>19</td>
<td>.79</td>
<td>.80</td>
<td>.81</td>
<td>.80</td>
<td>.81</td>
<td>.81</td>
<td>.80</td>
<td>.81</td>
<td>.82</td>
<td>.81</td>
</tr>
<tr>
<td>20</td>
<td>.80</td>
<td>.81</td>
<td>.81</td>
<td>.80</td>
<td>.82</td>
<td>.82</td>
<td>.81</td>
<td>.82</td>
<td>.83</td>
<td>.82</td>
</tr>
</tbody>
</table>

Note. $N_{\text{therapists}} = 3$. 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 (Perpletchikova & 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


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


