Psychotherapy
2018, Vol. 55, No. 2, 105–115

0033-3204/18/$12.00 http://dx.doi.org/10.1037/pst0000139

Developing a Therapeutic Relationship Monitoring System for Group Treatment

Rebecca A. Janis
Pennsylvania State University

Gary M. Burlingame and Joseph A. Olsen
Brigham Young University

The use of outcome monitoring systems to identify clients that are at-risk for treatment failure has now become part of daily clinical practice, shown in >25 empirical studies to improve client outcomes. These promising findings have led to outcome monitoring systems being recognized as evidence-based. Feedback systems based on client perception of therapeutic processes are recent additions to the monitoring literature, and the research suggests that these too lead to improved outcomes. Unfortunately, feedback systems and research have been primarily limited to individual therapy, creating a knowledge gap for multiperson treatment. This study reports on the development of a therapeutic relationship monitoring system for group treatment using results from 6 Group Questionnaire (GQ) studies conducted in 4 unique clinical populations: nonclinical process, counseling center, European inpatient, and seriously mentally ill inpatients. The GQ is a factor-analytically derived scale, which assesses a client's perception of 3 relationship quality constructs (positive bond, positive work, and negative relationship) across 3 structural domains (member–member, member–leader, and member–group). The first goal of the present study was to replicate the previously established factor structure across each clinical population. The second goal was to establish normative values and relevant feedback alerts for the GQ subscales in each population. Findings support the GQ factor structure across clinical populations, indicating that the constructs measured by the GQ bear similar relationships in each population. Further, findings support the implementation of unique norms and feedback alerts in each clinical population, reflecting the reality of meaningful differences between clinical populations.

Keywords: group psychotherapy, process monitoring, therapeutic relationship, Group Questionnaire

The monitoring of a client's progress in real-time with an outcome measure that is sensitive to change has clearly taken hold in both clinical practice and the research literature (Lutz, De Jong, & Rubel, 2015). The so-called patient-focused paradigm (Howard, Moras, Brill, Martinovich, & Lutz, 1996) has led to several quality assurance systems (Barkham et al., 2001; Beutler, 2001; Kordy, Hannöver, & Richard, 2001; Lambert, Hansen, & Finch, 2001; Locke et al., 2011) that use continuous monitoring of a patient's outcome status to alert therapists on whether a client is improving, deteriorating, or showing no reliable progress.

There are now at least 25 studies testing different monitoring systems, with the preponderance showing a positive effect of monitoring on outcome (Krägeloh, Czuba, Billington, Kersten, & Siegert, 2015). In a series of 12 randomized clinical trials (RCTs), Lambert (2015) has shown that providing progress feedback to individual therapists with the Outcome Questionnaire 45 (OQ-45; Lambert & Burlingame, 1996) improves client outcomes, especially for clients at risk for treatment failure. At-risk status is based on normative change trajectories based on a client's initial level of distress. Expected change trajectories statistically define boundaries within which a client is on-track for a successful outcome, and cut offs beyond which a client is off-track and at risk for treatment failure. The largest effects of feedback were for off-track cases at risk for treatment failure (cf. Shimokawa, Lambert, & Smart, 2010), cases which therapists have been shown to be unable to identify without feedback (Burns & Auerbach, 1996; Hannan et al., 2005). The effects of progress feedback have been replicated using other outcome instruments (Krägeloh et al., 2015), suggesting that measure-based feedback can improve client outcomes and add real value to the therapeutic enterprise.

In addition to the progress feedback research, there is a small but growing body of research that suggests that feedback on a client's perception of key therapeutic processes can also improve outcome. This research relies upon measures of therapeutic processes that have been shown to be reliable moderators of client improvement. For example, feedback to a therapist on a client's perception of the quality of the therapeutic alliance, motivation and readiness for change, as well as their perceived social support have been examined in a series of studies led by Lambert's team (Harmon et al., 2007; Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004; Slade, Lambert, Harmon, Smart, & Bailey, 2008; Whipple et al., 2003). Whipple and colleagues (2003) demonstrated a relationship between a therapist receiving process feedback and better out-

Rebecca A. Janis, Department of Psychology, Pennsylvania State University; Gary M. Burlingame, Department of Psychology, Brigham Young University; Joseph A. Olsen, College of Family, Home and Social Sciences, Brigham Young University.

Rebecca A. Janis and Gary M. Burlingame contributed equally to this work, and their lead authorship should be considered interchangeable.

Correspondence concerning this article should be addressed to Rebecca A. Janis, Department of Psychology, Pennsylvania State University, State College, PA 16801. E-mail: ruj134@psu.edu

comes when compared with both a no process or outcome feedback condition and a condition providing only outcome feedback. The positive effects of providing the individual therapist with feedback on the therapeutic process have been replicated twice with similar results (Hawkins et al., 2004; Slade et al., 2008) and once where mixed findings were produced (Harmon et al., 2007). Collectively, the use of therapy process feedback appears to reliably add to the effects of progress feedback (cf. Shimokawa et al., 2010). However, it is important to note that therapeutic process feedback in Lambert's study was only delivered when the outcome monitoring system indicated that a client was at risk for treatment failure. Other monitoring systems (Duncan & Reese, 2015) provide outcome and therapeutic process feedback for all clients at each session.

Operationalizing Alerts Used in Outcome and Process Monitoring System Research

The common mechanism of change in both outcome and therapeutic process feedback is therapist awareness of how a client's current score compares with normative values. Therapist awareness is anchored by using on- and off-track feedback messages that are triggered when a client's score is classified as either normative or exceptional, using one of three types of alerts: status, progress, or change. Status alerts typically classify a client's score on an outcome or process measure in reference to a specific normative population. The most common example is a cut score (typically the 50th percentile) that separates two distinct clinical population distributions (Jacobson & Truax, 1991), typically a clinical or community normal population. The second type of feedback message is a progress alert, used to classify whether a client's progress on an outcome measure from an expected progress trajectory associated with successful treatment (Howard et al., 1996). Progress alerts provide the clinician with information on whether a client is "on-track" or "off-track" for a successful treatment outcome. The final type of feedback message is a change alert, which is triggered when a client's score shows meaningful change from their baseline or previous score (e.g., last session). The most common change alert is the reliable change index (RCI; Jacobson & Truax, 1991) based on the standard error of measurement of the instrument.

Operationalizing and distinguishing the different types of alerts sets up a path to integrate alerts into a formal structured feedback system for group treatment, which Krägeloh and colleagues (2015) conclude makes feedback more effective. Recent research demonstrates that the outcome trajectories for members in group treatment are indistinguishable from clients treated in individual therapy when progress is assessed by the OQ-45 (Burlingame, Gleave, Erekson, et al., 2016). This suggests that the OQ-45 progress alert system already in place is applicable to group treatment, so we focus here on a creating a monitoring system providing feedback for a key therapeutic process shown to be a trans-theoretical predictor of outcome in group treatment: the client's perception of the therapeutic relationship (Burlingame, McClendon, & Alonso, 2011; Burlingame, Strauss, & Joyce, 2013). We briefly review the group studies that test the effect of feedback on the therapeutic relationship, three new measures used to assess the therapeutic relationship, and two studies that have applied these measures in clinical practice. Finally, we describe the steps we undertook to create a formal structure to provide therapists with status and change alerts on the therapeutic relationship in group therapy using the Group Questionnaire (GQ).

Feedback on Therapeutic Processes in Group Treatment

Although the research on feedback using measures of therapeutic processes in individual therapy has recently grown (Krägeloh et al., 2015), few such studies exist in the group literature. The earliest feedback study we could locate was a study of intimacy training groups by Widra and Amidon (1987), in which members of the feedback group received their personal intimacy scores, as well as personalized written feedback from the group leader. The group receiving feedback showed the greatest improvement in scores, providing early support for the value of measure-based feedback supported by the leader.

The next group feedback study located was Davies (2004), which provided feedback using the Group Climate Questionnaire (GCQ; Mackenzie, 1981) to both leaders and members. The GCQ assesses a member's perception of the therapeutic relationship using three subscales—engagement, conflict, and avoidance. In addition to being one of the most frequently used group relationship measures (Burlingame, MacKenzie, & Strauss, 2004), it has proven to be one of the best transtheoretical predictors of outcome in group treatment (Burlingame et al., 2011). No effects for the GCQ item feedback condition were found on either the Curative Climate Inventory (CCI; Fuhriman, Drescher, Hanson, Henrie, & Rybicki, 1986) or the OQ-45, although both conditions did show change on the OQ-45. In explaining their results, they noted a lack of change over time on GCQ and CCI subscales for both the feedback and no-feedback conditions, which contradicts past GCQ research (McClendon & Burlingame, 2011) that shows an increase in engagement (GCQ) and cohesion (CCI) over time. They made several recommendations for future feedback research, including selecting a measure that provides personalized information for each member germane to the client's work in the group. They also recommended providing feedback to leaders and members that prompts action and selecting a measure that addresses a therapeutic process that is under the control of the member. Group climate was not viewed by members as something they could change, as it is a property of the group-as-a-whole. In sum, these limitations make an implicit call for a new measure to support feedback that addresses the weaknesses pointed out by Davies, Burlingame, Johnson, Gleave, and Barlow (2008).

Development and Study of Feedback Measures for Group Treatment

The limitations of the GCQ as a feedback tool coupled with the plethora of measures used to assess the therapeutic relationship in group (Burlingame, Fuhriman, & Johnson, 2002) led to the development of several measures intended to provide relationship or process feedback in group treatment. The GQ, discussed more comprehensively below, was derived after Davies' (2004) study by factor analyzing the responses of 662 members in 111 counseling center and personal growth groups on the four most frequently used group relationship measures (alliance, cohesion, climate, and empathy). This produced three latent factors—positive bond, pos-

itive work, and negative relationship (Johnson, Burlingame, Olsen, Davies, & Gleave, 2005). The Therapeutic Factors Inventory (TFI), originally designed to assess Yalom's 11 therapeutic factors, was shortened to a 23-item measure with four scales—hope, secure emotional expression, awareness of interpersonal impact, and social learning (MacNair-Semands, Ogrodniczuk, & Joyce, 2010). More recently,Tasca and colleagues (2016) shortened the TFI to eight items to support progress feedback. Finally, the Group Sessions Rating Scale (GSRS) is the most recently developed ultra-brief group alliance measure with four items that are completed using a visual analogue scale (Quirk, Miller, Duncan, & Owen, 2013). Although all three measures are intended to support monitoring, only three studies have tested these measures in clinical practice.

Chapman et al. (2012) conducted a study to determine if group therapists showed the same difficulty predicting treatment outcome as individual therapists using the OQ-45 (Hannan et al., 2005). Therapists were also asked to predict group members' perception of the therapeutic relationship using the three GQ subscales—Positive Bond, Positive Work, and Negative Relationship. Therapists were unable to predict the outcome status of any of the deteriorated clients, replicating findings from individual therapy (Hannan et al., 2005). Furthermore, no statistical relationship was found between therapist and member GQ scores in 15 of 18 comparisons. Chapman et al. argued that the absence of accurate prediction of outcome and therapeutic relationship supports the necessity of monitoring systems on group treatment.

A second study (Slone, Reese, Mathews-Duvall, & Kodet, 2015) randomly assigned groups to either an outcome feedback condition using Partners for Change Outcome Management System (Duncan, 2012) monitoring system and the 4-item GSRS or treatment as usual (TAU), in which outcome was tracked but no feedback provided. A significant effect ($d = .41$) was found for the feedback condition showing more improvement than the TAU condition.

Finally, Davidsen and colleagues (2017) tested progress (ORS) and group relationship (GSRS) feedback. Like Slone et al. (2015), outcome and relationship feedback were combined in the experimental condition. However, they did not find an effect for this combined feedback on attendance, outcome (ORS), eating disorder, or other measures of change. The combination of feedback on the ORS and GSRS in these two studies makes it impossible to test the independent effect of relationship feedback in group therapy. Thus, no conclusions regarding the independent effect of therapeutic relationship feedback in group treatment can be drawn.

Development of the Group Questionnaire

The GQ evolved out of programmatic research conducted by the Consortium for Group Research and Practice or C-GRP (http://cgrp.byu.edu) in the early 2000s (Burlingame, Gleave, Beecher, et al., 2016). C-GRP completed a review demonstrating strong empirical support for the therapeutic relationship predicting group member improvement across a wide variety of clinical settings and theoretical orientations (Burlingame et al., 2002) and found the GCQ to be the most frequently used measure in this literature (Burlingame et al., 2004). The null results of the Davies (2004) study, however, helped decide against the GCQ as an instrument to support therapeutic relationship feedback in group therapy. During this period, C-GRP was participating in the development of a core battery of process and outcome instruments with an international taskforce created by the American Group Psychotherapy Association (AGPA). This task force identified instruments with the greatest empirical support in group treatment that could be used to support measure-based feedback for group therapy (Burlingame et al., 2006; Strauss, Burlingame, & Bormann, 2008). The Davies (2004) study results led C-GRP to undertake a measurement development study (Johnson et al., 2005) to determine if there were underlying latent constructs within the common measures being used to assess the therapeutic relationship in group treatment. This study included four recommended therapeutic relationship measures in the AGPA core battery: Working Alliance Inventory (WAI; Horvath & Greenberg, 1989), Burns empathy measure (Burns & Auerbach, 1996), GCQ (MacKenzie, 1981), and the Cohesion subscale of the TFI (Lese & MacNair-Semands, 2000).

The four relationship measures were completed by nearly 700 group members participating in counseling center and nonclinical process groups. The resulting model (Johnson et al., 2005) identified a latent relationship quality dimension with three factors (positive bond, positive work, and negative relationship) and a latent relationship structure dimension (member–member, member–group, and member–leader). This model was then replicated with inpatient groups in Germany and Switzerland (Bormann & Strauss, 2007), and partially replicated with short and long-term analytic groups in Norway (Bakali, Baldwin, & Lorentzen, 2009). The first two studies led to a construct map of how the alliance, cohesion, climate, and empathy items mapped onto the "Johnson model" (see Table 1). The next two C-GRP studies reduced the 80+ items used in in the Johnson model to 30 items by identifying items that could lead to therapist action but also loaded highly on Johnson model. The 30-item GQ replicated the Johnson model with 485 members participating in nonclinical, outpatient, and inpatient groups in the United States (Krogel, 2008;

Table 1
Underlying Relationship Constructs Assessed by Crossing the GQ Quality and Structure Dimensions

GQ subscales	Member–Member	Member–Leader	Member–Group
Positive bonding	Cohesion	Alliance	Climate
Positive working	Task/goals	Task/goals	—
Negative relationship	Empathetic failure	Alliance rupture	Conflict

Note. Positive work is not measured on the Member–Group structural domain.

Krogel et al., 2013) and 424 members participating in German inpatient groups (Bormann, Burlingame, & Strauß, 2011). Finally, the GQ criterion validity was assessed to be high with the original four therapeutic relationship measures in groups conducted in both the United States and Germany (Bormann et al., 2011; Thayer & Burlingame, 2014).

The goal of the present study is to conduct a mega-analysis by pooling results from these six studies to simultaneously test the factor structure with a larger sample that includes all available clinical and nonclinical populations. Our second goal is to determine if the means of the three relationship quality subscales (positive bond, positive work, and negative relationship) vary by clinical population. After determining which normative populations are needed, our final goal was to determine GQ status alerts to support feedback to group leaders.

Methods

Participants

Data from six published studies using the GQ were included in this study. Each study's descriptive information is reported in Table 2. Participants were drawn from four distinct populations: university counseling centers, European inpatient, severely mentally ill (SMI) inpatient, and nonclinical AGPA. University counseling center data were taken from 16 sites (Chapman et al., 2012; Johnson et al., 2005; Krogel et al., 2013; Thayer & Burlingame, 2014). European inpatient data were taken from psychiatric hospitals in Germany and Switzerland (Bormann et al., 2011; Bormann & Strauss, 2007). Severely mentally ill inpatient data were taken from the Utah State Hospital (Chapman et al., 2012; Krogel et al., 2013). The AGPA data were taken from 2-day training groups at the 2002 annual meeting of the AGPA (Johnson et al., 2005; Krogel et al., 2013).

Data from 2,479 participants were included in the analyses. Three of the studies had information on group membership available (Chapman et al., 2012; Johnson et al., 2005; Thayer & Burlingame, 2014), and data from those three studies were included in the multilevel model to account for both between and within group variance in evaluating factor structure. Overall, data from 1,058 participants in 195 groups were included in the multilevel analysis, with an average of 5.43 (SD = 3.49) participants per group.

Instrument

The GQ is a 30-item self-report measure of the quality of therapeutic relationship in groups. It is measured on a 7-point Likert scale from 1 (*not true at all*) to 7 (*very true*). The scale assesses therapeutic relationship on three subscales: positive bonding relationship (PB; 13 items), positive working relationship (PW; 8 items), and negative relationship (NR; 9 items). The measure yields a score for each of the three factors, with no total score. All three subscales have good reliability, with positive bond ranging from .79 to .92, positive work ranging from .85 to .91, and negative relationship ranging from .86 to .87 (Chapman et al., 2012; Krogel et al., 2013; Thayer & Burlingame, 2014). Reliable change values have been calculated as follows for each subscale: PB = 10, PW = 9, and NR = 11 (Burlingame, Gleave, Beecher, et al., 2016). The GQ also assesses relationships on three structural dimensions (member–leader, member–member, and member–group), with questions specifically referring to the member's perception of the leader, the other members, and the group as a whole. Structural dimensions are assessed on all three subscales, with the exception of positive work, for which the member–group relationship is not assessed. Intraclass correlation coefficients (ICCs) suggest that the GQ adequately captures between and within group variance (Thayer & Burlingame, 2014). The GQ also shows criterion validity, with acceptable correlations with the WAI, GCQ, TFI, and Empathy Scale (ES; Thayer & Burlingame, 2014).

Analyses

Two studies included in this analysis (Chapman et al., 2012; Krogel et al., 2013) used the GQ-40 (an earlier version of the GQ), and two other studies (Bormann & Strauss, 2007; Johnson et al.,

Table 2
Descriptive Information About Included Studies

Study	Populations	*N*	Groups	Instruments
Johnson et al. (2005)	Counseling center; Nonclinical	662	120*	Group Climate Questionnaire; Therapeutic Factors Inventory; Working Alliance Inventory; Empathy Scale
Bormann and Strauss (2007)	European inpatient	438	67	Group Climate Questionnaire**; Therapeutic Factors Inventory**; Working Alliance Inventory**; Empathy Scale**
Bormann et al. (2011)	European inpatient	498	64	Group Questionnaire**
Chapman et al. (2012)	Counseling center; SMI inpatient	106	18*	Group Questionnaire; Severe Outcome Questionnaire; Outcome Questionnaire 45
Thayer and Burlingame (2014)	Counseling center	290	64*	Group Questionnaire; Group Climate Questionnaire; Therapeutic Factors Inventory; Working Alliance Inventory
Krogel et al. (2013)	Counseling center; Nonclinical; SMI inpatient	485	—	Group Climate Questionnaire; Therapeutic Factors Inventory; Working Alliance Inventory
Total		2,479	202*	

Note. Number of groups not available for Krogel et al. (2013). SMI = severely mentally ill.
* Group membership used in analysis. ** German translation.

2005) employed instruments used to create the GQ (GCQ, TFI, WAI, ES). From these instruments, only the items common to the GQ-30 were included in this analysis.

Data from Chapman et al. (2012) had GQ data from several time points. An ANOVA revealed no difference in mean GQ score by time point for any of the subscales (Positive Bond—$F(2, 239) = .011$, $p = .989$; Negative Relationship—$F(2, 239) = 1.519$, $p = .221$; Positive Work—$F(2, 241) = .530$, $p = .589$). As GQ scores did not differ by time point, we chose the time point for each group that had data from the greatest number of members and included only that time point for that group in the analysis to preserve the greatest amount of data.

Question 1-Does the factor model provide adequate fit for the data? To first confirm that the GQ factor structure is generally replicated in a different sample of data from the one on which it was established, a single level confirmatory factor analysis was conducted. Although clients were nested within groups, resulting in dependency indicating the use of a multilevel model, group membership was not available for some of the data, so an initial single level model was conducted to maximize the data included. Analyses were done in Mplus using maximum likelihood estimation. This model (see Figure 1) comprised three second-order factors representing the GQ subscales (positive bond, positive work, and negative relationship) and eight first-order factors representing the group therapy relationships within each of the QG subscales (member–group bond, member–leader bond, member–member bond, member–leader work, member–member work, member–group negative relationship, member–leader negative relationship, and member–member negative relationship). To determine how well the proposed model fit the data, several goodness-of-fit indices were used, including the Comparative Fit Index (CFI), the Tucker–Lewis index (TLI), the Root Mean Square Error of Approximation (RMSEA), the Standardized Root Mean Square Residual (SRMR), and chi-square. When evaluating CFI and TLI, values above .90 indicate acceptable model fit. For RMSEA, values below .05 indicate acceptable model fit, and for SRMR, values less than .08 indicate acceptable fit (Hu & Bentler, 1999). In evaluating chi-square, nonsignificant values, or values less than twice the model's degrees of freedom indicate acceptable model fit. Additionally, a multigroup confirmatory factor analysis was performed on data grouped by population to ensure that the proposed factor structure fit the data derived from each population, and not just at the aggregate level. Factor invariance of first and second-order factor loadings was tested.

After performing single-level analyses with data from all six of the studies providing complete GQ data, we moved to data providing group membership. These data were used to calculate ICCs, which provide a measure of how similar observations are within groups of clustered data (Kenny, Mannetti, Pierro, Livi, & Kashy, 2002). Given the nested nature of the data and dependency observed in previous studies using the GQ, a multilevel confirmatory factory analysis was then performed to account for within and between group variance (Janis, Burlingame, & Olsen, 2016). Data from three studies providing group membership information were included in the calculation of ICCs and the estimation of the multilevel model (Chapman et al., 2012; Johnson et al., 2005; Thayer & Burlingame, 2014). As clients are nested within psychotherapy groups, single-level modeling risks violating the "indepen-

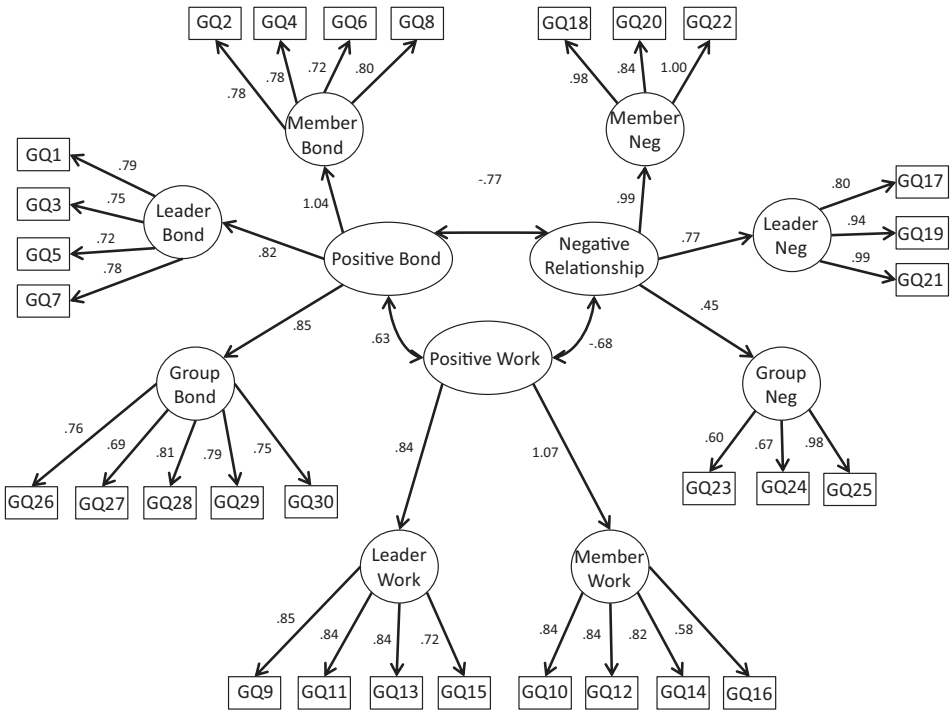

Figure 1. Single-level confirmatory factor analysis. This figure illustrates the between-level factor structure and parameters. Model fit for the between level is as follows: $\chi^2(380) = 1571.37$, RMSEA = .036, CFI = .96, TLI = .95, SRMR = .038.

dence of observations" assumption, and increases the risk of Type I error. Additionally, ICCsreported in Thayer and Burlingame (2014) were all close to or above .10 (.07–.49), indicating the need for multilevel analysis. Unlike a single-level modeling, which estimates the total covariance structure, multilevel modeling allows for the fitting of a between-groups covariance structure, which reflects differences between groups, as well as a pooled within-groups covariance structure, which reflects differences of individual clients within groups. Factor loadings at the between and within levels were constrained to be equal. Additionally, at the within level, measurement errors for the member–leader items and member–member items with corresponding wording were correlated. We used the procedure outlined in Ryu and West (2009) to obtain level specific fit by estimating partially saturated models at the between and within levels. Chi-square, RMSEA, and CFI were then calculated at both the within and between level. The multilevel analysis in this study used data from 1,058 participants in 205 groups.

Question 2-Are different normative values needed for different populations? As status alerts are based upon descriptive statistics (means, standard deviations, and outlier values), we calculated average subscale scores on each of the three subscales (Positive Work, Positive Bond, Negative Relationship) for each population (counseling center, nonclinical, European inpatient, SMI inpatient). We evaluated differences between populations using effect sizes. A Cohen's *d* effect size was calculated to indicate the standardized difference between the means for each population on each subscale. An effect size greater than .5 indicates a medium effect and .2 a small effect according to Cohen's benchmarks (Cohen, 1988). These values were used to evaluate the magnitude of differences between populations.

On the basis of the findings of normative values, positive and negative cut scores for generating GQ status alerts were calculated for each population only if it was determined that different normative values were needed. Negative and positive cut scores indicate the score at which 10% of the clients are likely to fall below on the low end and at which 10% of clients are likely to fall above on the high end. Negative and positive cut scores for positive bonding and positive working are set at the 10th and 90th percentile respectively, whereas negative and positive cut scores for negative relationship are set at the 90th and 10th percentiles respectively. Status alerts are generated for a given client when his or her scores meet or exceed the cut score. Status alerts indicate that a client is at risk for treatment failure based on low (or high, for negative relationship) levels of empirically supported relationship constructs.

Results

Question 1-Does the Model Provide Adequate Fit for the Data?

In fitting the single level CFA, there was one Heywood case, a negative residual variance (member–member work), which was constrained to be nonnegative. Results from the single-level CFA (see Figure 1) indicated that the model provided adequate fit for the data on all tests of model fit except for χ^2 ($\chi^2(380) = 1571.37$, RMSEA = .036, CFI = .96, TLI = .95, SRMR = .038). Given that all other indices indicated good model fit and χ^2 is overestimated as sample size increases, we concluded that the model provided acceptable fit.

All items loaded significantly on their intended factors ($p < .001$), and subscales representing positive and negative aspects of the group relationship were correlated as expected. Positive bond and positive work were positively correlated with each other, $r = .65$, $p < .001$, whereas negative relationship was negatively correlated with both positive bond, $r = -.69$, $p < .001$, and positive work, $r = -.39$, $p < .001$. Each of the second-order factors had a clearly dominant first order factor (member–member relationship) with which it was almost perfectly related. Additionally, the multigroup CFA model across all four populations provided adequate fit on several tests of model fit (CFI = .92, TLI = .91, SRMR = .058) and borderline acceptable fit on one other (RMSEA = .054).Similarly to the single-level analysis, χ^2 did not indicate adequate fit ($\chi^2(1601) = 4460.89$). Taken together, the fit indices provide evidence that the GQ factor structure replicates across populations.

ICCs for the GQ items were based on data from 1,058 members from 202 groups. ICCs ranged from .06to .49, with a mean of .17, closely replicating the ICCs found in Thayer and Burlingame (2014) and confirming the need for a multilevel model to account for this shared variance. The items least affected by group membership (smallest ICCs) were from the negative relationship subscale tapping perceptions of the leader's and other members' genuineness (.09/.10, respectively), caring (.07/.10) and understanding (.06/.08). Interestingly, the items most affected by group membership (largest ICCs) were from the same scale, but tapping the perceived conflict, distance, and tension (.49, .30, and .26, respectively) in the group. Average ICCs for positive bond, positive work, and negative relationship were .18, .18 and .16, respectively, and ICCs for member–leader, member–member, and member– group relationships were .13, .13, and .28, respectively.

For the multilevel model, the RMSEA (.032), CFI (.95), TLI (.94), and within-level SRMR (.044) all indicated acceptable model fit. However, $\chi^2(1672.13, df = 798)$ and between-level SRMR (.140) fell short of the criteria for acceptable model fit. Factor loadings and correlation coefficients between-groups and within-groups were similar to those from the single-level CFA. In fitting the initial multilevel CFA, there were four negative residual variances, and models setting those residual variances to zero or constraining them to be non-negative gave results that were virtually identical to the original model.

In addition to the overall model fit, we also examined level-specific within-groups and between-groups fit measures (Ryu & West, 2009). The within-groups, $\chi^2(380) = 803.33$, $p < .001$ and between-groups, $\chi^2(391) = 514.24$, $p < .001$. Chi-square values were both statistically significant, though this is again expected with such a large sample. The adjusted within-groups (.030) and between-groups (.040) RMSEA values both indicated good fit, but the within-groups CFI (.97) showed substantially better fit than the between-groups CFI (.81). We fit several other models at the between level in search for a good fitting model. Modification indices suggested the addition of covariances between structural components and between individual items. Although the addition of these covariance parameters improved model fit, they did not result in a model meeting criteria for acceptable model fit at the between level ($\chi^2(774) = 1219.85$, $p < .001$, SRMR between =

.127). A between-level model was also estimated in which items loaded only on the three relationship subscales, removing the second-order structural components; however, this model would not converge.

Question 2-Are Norms Needed for Different Populations?

Means and standard deviations were calculated for each population for each of the three relationship quality subscale scores, which were created as unit-weighted linear composites of the items associated with each of the three factors (see Table 3). Inspection of the table shows that the counseling center population was highest on positive bonding and positive working relationships and lowest on negative relationship. The European and SMI inpatient populations were both lower on positive bonding relationship, but interestingly differed on negative relationship, with the European inpatient group having the highest mean. The nonclinical group had the lowest level of positive working relationship but was close to the grand mean for the other two subscales. Six effect sizes indicated a moderate effect, and 13 of 18 (72%) indicated either a small or moderate effect (see Table 4). Although on some subscales, certain populations were not statistically or clinically different, on others, they are different. None of the four populations could be considered equivalent, as there was at least one subscale that differed between each of them using our moderate effect size rule. We took this, in conjunction with the statistical and clinical differences, as an indication that different norms were required for each population. Because unique norms were created for each population, unique cut scores were also required to support status alerts.

Although differences were evident between the normative values, the differences become even more apparent in the cut scores, which represent the extremes of each subscale in each population (see Table 5). The difference in cut-off values is most striking for the positive bonding subscale, on which the negative cut score ranges from 44 for an SMI inpatient population to 63 for a counseling center population, a 19-point difference that almost doubles the change alert RCI of 10 for that scale. For positive work and negative relationship, the differences between populations are less drastic, but still evident. For positive work, there is a 9-point difference between the nonclinical (17) and European inpatient population (26), which is equivalent to the change, alert RCI for that subscale. For negative relationship, the cut scores range from 31 in the counseling center population to 39 in both the SMI and European inpatient populations. This 8-point difference is also close to the change alert RCI of 11 for the negative relationship subscale. Taken together, these differences provide further support for the establishment of different normative values and cut scores for each population.

Discussion

The previously established factor structure of the GQ (Bormann et al., 2011; Bormann & Strauss, 2007; Johnson et al., 2005; Krogel et al., 2013; Thayer & Burlingame, 2014) generally provided good fit for a single level model that aggregated data across all the populations. Further, the factor structure was independently replicated across all four populations tested, indicating that the constructs are operating in the same ways across different populations. The aggregate replication is not surprising, as it included nearly 700 members from Johnson et al. (2005). However, the value of this mega-analysis is that it also included data from populations not represented in the Johnson model (i.e., SMI and European inpatient). By combining data from six separate studies, we were able to replicate a common factor structure for four distinct clinical populations with a larger and more diverse sample than found in any single study, confirming a structure that was previously established in individual studies with smaller sample sizes. Taken together, these results provide further support for the factorial validity of the GQ and confidence in its use across an array of nonclinical and clinical group applications.

The GQ provides feedback at the subscale level (there is no total score), so a fair question is how much overlap there is between subscales. Positive bond correlates fairly strongly with the positive work and negative relationship subscales (r = .65 and −.69, respectively). However, it is important to remember that the shared variance or information between these subscales is still below 50% (42–48%), suggesting that they tap related, but somewhat independent constructs. Greater subscale independence is found between the negative relationship and positive work subscales, which shared 15% of the variance in our samples. These patterns follow clinical intuition, as the positive bond and negative relationship scales tap the affective climate within the group while positive work assesses satisfaction with the tasks and goals of the group. Taken together, we'd expect a higher alert correspondence for the subscales that correlate highly compared with the lower correlation between the negative relationship and positive work subscales.

The GQ was designed as a feedback support tool for group treatment and high item and subscale ICCs suggest that it is indeed sensitive to the effects of group membership. Although there is no consensus on how large the ICC needs to be to indicate that a

Table 3
Means and Standard Deviations for Group Questionnaire Subscales by Population

	Positive bonding			Positive working			Negative relationship		
Population	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Counseling center	850	77.37	10.56	832	39.66	10.34	845	20.18	7.61
Nonclinical	451	74.35	10.27	406	34.11	11.60	444	24.07	7.52
European inpatient	864	68.91	12.39	871	38.44	9.2	873	26.78	8.66
SMI inpatient	157	66.70	16.57	153	38.47	12.95	147	22.01	11.00
All	2,322	72.91	12.37	2,262	38.11	10.52	2,311	23.54	8.74

Note. *N*s differ between subscales due to missing data. SMI = severely mentally ill.

Table 4
Cohen's D Effect Sizes Between Populations on Group Questionnaire Subscales

GQ subscales	Counseling center	SMI inpatient	Nonclinical
Positive bonding			
SMI inpatient	.77	—	—
Nonclinical	.29	−.55	—
European inpatient	.73	−.13	.39
Positive working			
SMI inpatient	.10	—	—
Nonclinical	.51	.35	—
European inpatient	.12	.00	−.41
Negative relationship			
SMI inpatient	−.19	—	—
Nonclinical	−.51	−.21	—
European inpatient	−.81	−.48	−.33

Note. Group in left column treated as control group in calculating effect sizes. SMI = severely mentally ill.

group is having a significant effect on a member's score (Janis et al., 2016), most would agree that values at or above .05 reflect at least moderate group influence. The ICCs for the three GQ quality subscales were over three times this value (.16–.18) in our study. Moreover, the member–group subscale was over five times this value (.28), suggesting the GQ is sensitive to the effect a group has on its members. ICC values can be interpreted as percent of variance estimates (Janis et al., 2016), which means that in our sample, 28% of the differences between members' scores on the member–group subscale is explained by the group to which the member belonged. Furthermore, the items with the highest ICC values (.26–.49) were those in the member–group negative relationship cell, which taps group conflict and is a clear group-specific phenomenon. Collectively, these results argue for the sensitivity of the GQ in capturing group-specific effects on the therapeutic relationship.

The multilevel model examining both within and between group fit revealed some model misfit. The three relationship quality (positive bond, positive work, and negative relationship) and structure (member–member, member–leader, and member–group) latent constructs provide acceptable fit when using the GQ to explain differences between members within the same group. However, poorer fit exists when this model is used to explain differences between groups. In other words, the model better captures the relationships between the quality and structure subscales at the member level, but less well at the whole group level. This allows for the measure to be used confidently to describe how individuals within groups differ from each other, which is the most common use of a relationship inventory for group psychotherapy. Indeed, this was the primary purpose for the creation of the GQ (Burlingame, Gleave, Beecher, et al., 2016). More empirical work is suggested by our results in which the GQ is used to capture differences between groups. This model misfit at the between level may be a result of the heterogeneity in group type represented in the data. Even within one setting, multiple types of groups were included in our data set. For example, the university counseling center setting included both process groups and structured, disorder specific groups. Indeed, Burlingame and colleagues (2018) found that both average GQ scores and GQ change over treatment differed between these two types of groups. We cannot determine with the present data whether the between group model fit might be improved if there were greater homogeneity within the groups (e.g., same theoretical orientation applied to a diagnostically homogenous set of members with a fixed dose). Clearly, this is an area for future factor analytic research to explore.

In evaluating the norms and cut scores across the four populations, we found that none of the populations were equivalent across all the GQ subscales. This indicates that although the same constructs are operating across populations, the normative levels of those constructs differ. As a function of these differences, the cut scores for each population differed as well, indicating that status alerts may differ between populations. The fact that different populations produced different normative values supports construct validity of the GQ. Clinically, we would expect outpatient groups (nonclinical and counseling center) to report higher levels of bonding than inpatient populations (SMI and European inpatient) who are experiencing more psychiatric impairment. This is the pattern we found. The lowest levels of positive work were reported by the nonclinical population made up of process groups held at the annual meetings of the AGPA. This makes sense, as they meet for two days, are made up of mental health professionals attending the conference, and have no common focus. However, the clinical groups in our sample were more often diagnostically homogenous (e.g., depression, anxiety, etc.) and followed an evidence-based protocol, which explains the higher levels of positive work reported. The normative values for negative relationship are mixed. The university counseling center population reported

Table 5
Group Questionnaire Subscale Alert Cut Scores by Population

	Positive bonding		Positive working		Negative relationship	
Population	10%	90%	10%	90%	10%	90%
Counseling center	63	89	25	52	11	31
Nonclinical	60	86	17	48	14	34
European inpatient	51	84	26	50	16	39
SMI inpatient	44	87	18	53	9	39
All	56	88	23	51	13	35

Note. Negative alerts for positive bonding and positive working are generated at the 10% cut score and for negative relationship at the 90% cut score. Positive alerts for positive bonding and positive working are generated at the 90% cut score and for negative relationship at the 10% cut score. SMI = severely mentally ill.

the lowest levels. We would expect the SMI inpatient sample to report the highest level of negative relationship, as they are in treatment against their will (civil or criminal commitment); however, both the European inpatient and nonclinical populations reported higher levels. The SMI inpatients may report lower levels, as discharge decisions are partly related to their level of cooperativeness. Future research with this population may need to include measures tapping demand characteristics to test this explanation.

The differences in normative values were amplified in the differences found between cut off values. For instance, differences between populations' cut scores often met or doubled the change alert RCI for the same subscale. Overall, the pattern of results suggests that there are different normative levels of latent constructs present in the different populations, but that the constructs are relating to each other in the same way across populations.

The results of this study suggest several avenues for extension and future directions. First, the misfit of several possible between level models and heterogeneity suggests that future research should examine the factor structure of the GQ in a sample of more homogeneous groups to determine whether a good fitting between model can be developed in that context. Developing an acceptable between level model would allow the GQ to be used at the group-as-a-whole level, which is valuable in both research and clinical settings. The development of a group-level model would also allow for the development of group-level status alerts indicative of potential problems at a group level.

Our proposal to create GQ norms and status alerts for different clinical populations is a departure from outcome monitoring systems that typically recognize only clinical and nonclinical cut scores. Our proposal to create population-specific norms raises a number of important questions such as—When should a new population be considered? Which are the right populations? How do we avoid fragmentation and proliferation of cut-scores? We are acutely aware of these concerns and offer our proposal with some trepidation. As we've trained group clinicians to use the GQ over the past decade, differences in therapeutic relationship between clinical settings emerged as potential obstacles to implementation. Outcome monitoring systems tap a core reason for a client to enter therapy—psychiatric distress—and most therapists easily grasp the distinction between clinical and nonclinical cut scores. Indeed, the construct validity of many outcome measures predicts higher scores for more severe populations. There is greater therapist consternation and less acceptance of a parallel continuum for the therapeutic relationship groups as they are applied to different clinical settings. Indeed, the findings herein on the GQ negative relationship provides some support for therapist concerns we've encountered. Perhaps a large instead of moderate effect size difference could be used to create distinct norms, and if so, the data herein support an omnibus status alert system. As more GQ data are accumulated over time, greater clarity on this matter will undoubtedly emerge.

A final consideration is how the two types of GQ alerts impact clinical practice. For example, do relative alerts, based on changes meeting or exceeding the RCI for a given scale, frequently precede absolute alerts, offering an early warning sign? Further, how do therapists respond to alerts in practice? Do alerts offer an opportunity to prevent and reverse potential relationship failures? Finally, and perhaps most importantly, how are alerts related to attendance, dropout, member engagement, and outcome?

References

Bakali, J. V., Baldwin, S. A., & Lorentzen, S. (2009). Modeling group process constructs at three stages in group psychotherapy. *Psychotherapy Research, 19,* 332–343. http://dx.doi.org/10.1080/10503300902894430

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., . . . the Clinical Outcomes in Routine Evaluation-Outcome Measures. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology, 69,* 184–196. http://dx.doi.org/10.1037/0022-006X.69.2.184

Beutler, L. E. (2001). Comparisons among quality assurance systems: From outcome assessment to clinical utility. *Journal of Consulting and Clinical Psychology, 69,* 197–204. http://dx.doi.org/10.1037/0022-006X.69.2.197

Bormann, B., Burlingame, G. M., & Strauß, B. (2011). Der Gruppenfragebogen (GQ-D). *Psychotherapeut, 56,* 297–309. http://dx.doi.org/10.1007/s00278-011-0841-4

Bormann, B., & Strauss, B. (2007). Group climate, cohesion, alliance, and empathy as components of the therapeutic relationship within group psychotherapy-Test of a multilevel model. *Gruppenpsychotherapie und Gruppendynamik, 43,* 1–20.

Burlingame, G., Fuhriman, A., & Johnson, J. (2002). Cohesion in group psychotherapy. In J. C. Norcross (Ed.), *A guide to psychotherapy relationships that work* (pp. 71–88). New York, NY: Oxford University Press.

Burlingame, G., Gleave, R., Beecher, M., Griner, D., Hansen, K., & Jensen, J. (2016). *Administration and scoring manual for the Group Questionnaire—GQ*. Salt Lake City, UT: OQ Measures, LLC.

Burlingame, G. M., Gleave, R., Erekson, D., Nelson, P. L., Olsen, J., Thayer, S., & Beecher, M. (2016). Differential effectiveness of group, individual, and conjoint treatments: An archival analysis of OQ-45 change trajectories. *Psychotherapy Research, 26,* 556–572. http://dx.doi.org/10.1080/10503307.2015.1044583

Burlingame, G. M., MacKenzie, K. R., & Strauss, B. (2004). Small group treatment: Evidence for effectiveness and mechanisms of change. In M. J. Lambert (Ed.), *Bergin & Garfield's Handbook of psychotherapy and behavior change* (5th ed., pp. 647–696). New York, NY: Wiley.

Burlingame, G., McClendon, D., & Alonso, J. (2011). Cohesion in group psychotherapy. In J. C. Norcross (Ed.), *A guide to psychotherapy relationships that work* (2nd ed., pp. 110–131). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199737208.003.0005

Burlingame, G., Strauss, B., & Joyce, A. (2013). Change mechanisms and effectiveness of small group treatments. In M. J. Lambert (Ed.), *Bergin & Garfield'shandbook of psychotherapy and behavior change* (6th ed., pp. 640–689). New York, NY: Wiley.

Burlingame, G., Strauss, B., Joyce, A., MacNair-Semands, R., MacKenzie, K., Ogrodniczuk, J., & Taylor, S. (2006). *CORE battery: A revision and update*. New York, NY: American Group Psychotherapy Association.

Burlingame, G. M., Whitcomb, K. E., Woodland, S. C., Olsen, J. A., Beecher, M., & Gleave, R. (2018). The effects of relationship and progress feedback in group psychotherapy using the group questionnaire and outcome questionnaire–45: A randomized clinical trial. *Psychotherapy, 55,* 116–131. http://dx.doi.org/10.1037/pst0000133

Burns, D. D., & Auerbach, A. (1996). Therapeutic empathy in cognitive-behavioral therapy: Does it really make a difference? In P. M. Salkovskis (Ed.), *Frontiers of cognitive therapy* (pp. 135–164). New York, NY: Guilford Press.

Chapman, C. L., Burlingame, G. M., Gleave, R., Rees, F., Beecher, M., & Porter, G. S. (2012). Clinical prediction in group psychotherapy. *Psychotherapy Research, 22,* 673–681. http://dx.doi.org/10.1080/10503307.2012.702512

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Davidsen, A. H., Poulsen, S., Lindschou, J., Winkel, P., Tróndarson, M. F., Waaddegaard, M., & Lau, M. (2017). Feedback in group psychotherapy for eating disorders: A randomized clinical trial. *Journal of Consulting and Clinical Psychology, 85,* 484–494. http://dx.doi.org/10.1037/ccp0000173

Davies, D. R., Burlingame, G. M., Johnson, J. E., Gleave, R. L., & Barlow, S. H. (2008). The effects of a feedback intervention on group process and outcome. *Group Dynamics: Theory, Research, and Practice, 12,* 141–154. http://dx.doi.org/10.1037/1089-2699.12.2.141

Davies, R. (2004). *The effects of a feedback intervention on group process and outcome*. (Unpublished doctoral dissertation). Brigham Young University, Provo, UT.

Duncan, B. (2012). The partners for change outcome management system (PCOMS): The heart and soul of change project. *Canadian Psychology/Psychologiecanadienne, 53,* 93–104. http://dx.doi.org/10.1037/a0027762

Duncan, B. L., & Reese, R. J. (2015). The partners for change outcome management system (PCOMS) revisiting the client's frame of reference. *Psychotherapy, 52,* 391–401. http://dx.doi.org/10.1037/pst0000026

Fuhriman, A., Drescher, S., Hanson, E., Henrie, R., & Rybicki, W. (1986). Refining the measurement of curativeness: An empirical approach. *Small Group Research, 17,* 186–201. http://dx.doi.org/10.1177/104649648601700204

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61,* 155–163. http://dx.doi.org/10.1002/jclp.20108

Harmon, S., Lambert, M., Smart, D., Hawkins, E., Nielsen, S., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist-client feedback and clinical support tools. *Psychotherapy Research, 17,* 379–392. http://dx.doi.org/10.1080/10503300600702331

Hawkins, E., Lambert, M., Vermeersch, D., Slade, K., & Tuttle, K. (2004). The therapeutic effects of providing patient progress information to therapists and patients. *Psychotherapy Research, 14,* 308–327. http://dx.doi.org/10.1093/ptr/kph027

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of Counseling Psychology, 36,* 223–233. http://dx.doi.org/10.1037/0022-0167.36.2.223

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy. Efficacy, effectiveness, and patient progress. *American Psychologist, 51,* 1059–1064. http://dx.doi.org/10.1037/0003-066X.51.10.1059

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19. http://dx.doi.org/10.1037/0022-006X.59.1.12

Janis, R. A., Burlingame, G. M., & Olsen, J. A. (2016). Evaluating factor structures of measures in group research: Looking between and within. *Group Dynamics: Theory, Research, and Practice, 20,* 165–180. http://dx.doi.org/10.1037/gdn0000043

Johnson, J. E., Burlingame, G. M., Olsen, J., Davies, D. R., & Gleave, R. L. (2005). Group climate, cohesion, alliance, and empathy in group psychotherapy: Multilevel structural equation models. *Journal of Counseling Psychology, 52,* 310–321. http://dx.doi.org/10.1037/0022-0167.52.3.310

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology, 83,* 126–137. http://dx.doi.org/10.1037/0022-3514.83.1.126

Kordy, H., Hannöver, W., & Richard, M. (2001). Computer-assisted feedback-driven quality management for psychotherapy: The Stuttgart-Heidelberg model. *Journal of Consulting and Clinical Psychology, 69,* 173–183. http://dx.doi.org/10.1037/0022-006X.69.2.173

Krägeloh, C. U., Czuba, K. J., Billington, D. R., Kersten, P., & Siegert, R. J. (2015). Using feedback from patient-reported outcome measures in mental health services: A scoping study and typology. *Psychiatric Services, 66,* 224–241. http://dx.doi.org/10.1176/appi.ps.201400141

Krogel, J. (2008). *The Group Questionnaire: A new measure of the group relationship*. Unpublished dissertation, Brigham Young University, Provo, UT.

Krogel, J., Burlingame, G., Chapman, C., Renshaw, T., Gleave, R., Beecher, M., & MacNair-Semands, R. (2013). The Group Questionnaire: A clinical and empirically derived measure of group relationship. *Psychotherapy Research, 23,* 344–354. http://dx.doi.org/10.1080/10503307.2012.729868

Lambert, M. J. (2015). Progress feedback and the OQ-system: The past and the future. *Psychotherapy, 52,* 381–390. http://dx.doi.org/10.1037/pst0000027

Lambert, M., & Burlingame, G. (1996). *Outcome Questionnaire* 45.2. Salt Lake City, UT: American Professional Credentialing Services, LLC.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69,* 159–172. http://dx.doi.org/10.1037/0022-006X.69.2.159

Lese, K. P., & MacNair-Semands, R. R. (2000). The therapeutic factors inventory: Development of a scale. *Group, 24,* 303–317. http://dx.doi.org/10.1023/A:1026616626780

Locke, B. D., Buzolitz, J. S., Lei, P.-W., Boswell, J. F., McAleavey, A. A., Sevig, T. D., . . . Hayes, J. A. (2011). Development of the counseling center assessment of psychological symptoms-62 (CCAPS-62). *Journal of Counseling Psychology, 58,* 97–109. http://dx.doi.org/10.1037/a0021282

Lutz, W., De Jong, K., & Rubel, J. (2015). Patient-focused and feedback research in psychotherapy: Where are we and where do we want to go? *Psychotherapy Research, 25,* 625–632. http://dx.doi.org/10.1080/10503307.2015.1079661

MacKenzie, K. R. (1981). Measurement of group climate. *International Journal of Group Psychotherapy, 31,* 287–295. http://dx.doi.org/10.1080/00207284.1981.11491708

MacNair-Semands, R. R., Ogrodniczuk, J. S., & Joyce, A. S. (2010). Structure and initial validation of a short form of the therapeutic factors inventory. *International Journal of Group Psychotherapy, 60,* 245–281. http://dx.doi.org/10.1521/ijgp.2010.60.2.245

McClendon, D., & Burlingame, G. (2011). Group climate: Construct in search of clarity. In R. Conyne (Ed.), *Oxford handbook of group counseling* (pp. 164–181). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780195394450.013.0010

Quirk, K., Miller, S., Duncan, B., & Owen, J. (2013). Group session rating scale: Preliminary psychometrics in substance abuse group interventions. *Counselling & Psychotherapy Research, 13,* 194–200. http://dx.doi.org/10.1080/14733145.2012.744425

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling, 16,* 583–601. http://dx.doi.org/10.1080/10705510903203466

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78,* 298–311. http://dx.doi.org/10.1037/a0019247

Slade, K., Lambert, M. J., Harmon, S. C., Smart, D. W., & Bailey, R. (2008). Improving psychotherapy outcome: The use of immediate elec-

tronic feedback and revised clinical support tools. *Clinical Psychology & Psychotherapy, 15,* 287–303. http://dx.doi.org/10.1002/cpp.594

Slone, N., Reese, R., Mathews-Duvall, S., & Kodet, J. (2015). Evaluating the efficacy of client feedback in group psychotherapy. *Group Dynamics: Theory, Research, and Practice, 19,* 122–136. http://dx.doi.org/10.1037/gdn0000026

Strauss, B., Burlingame, G. M., & Bormann, B. (2008). Using the CORE-R battery in group psychotherapy. *Journal of Clinical Psychology, 64,* 1225–1237. http://dx.doi.org/10.1002/jclp.20535

Tasca, G. A., Cabrera, C., Kristjansson, E., MacNair-Semands, R., Joyce, A. S., & Ogrodniczuk, J. S. (2016). The therapeutic factor inventory-8: Using item response theory to create a brief scale for continuous process monitoring for group psychotherapy. *Psychotherapy Research, 26,* 131–145. http://dx.doi.org/10.1080/10503307.2014.963729

Thayer, S., & Burlingame, G. (2014). The validity of the Group Questionnaire: Construct clarity or construct drift? *Group Dynamics: Theory, Research, and Practice, 18,* 318–332. http://dx.doi.org/10.1037/gdn0000015

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment and problem solving strategies in routine practice. *Journal of Counseling Psychology, 50,* 59–68. http://dx.doi.org/10.1037/0022-0167.50.1.59

Widra, J., & Amidon, E. (1987). Improving self-concept through intimacy group training. *Small Group Behavior, 18,* 269–279. http://dx.doi.org/10.1177/104649648701800209

Received April 18, 2017
Revision received August 16, 2017
Accepted August 18, 2017 ■