2016-07-01

An Item Reduction Analysis of the Group Questionnaire

Jennifer Lynn Jensen
Brigham Young University

An Item Reduction Analysis of the Group Questionnaire

Jennifer Lynn Jensen

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Gary Burlingame, Chair
Bruce Carpenter
Jared Warren

Department of Psychology

Brigham Young University

July 2016

ABSTRACT

An Item Reduction Analysis of the Group Questionnaire

Jennifer Lynn Jensen
Department of Psychology, BYU
Master of Science

Introduction. The Group Questionnaire (GQ) was developed to measure group therapeutic processes—which are linked to successful prediction of patient outcome and therapeutic factors—across three qualitative dimensions (positive bond, positive work, and negative relationship) and three structural dimensions (member-leader, member-member, and member-group). The GQ model has been shown to be valid across 5 settings and 4 countries. As a clinical measure given after each session, length is of particular concern. Although shorter measures are more convenient for clients and therapists to use, fewer items necessarily means less information, a loss of psychometrics, and possible floor and ceiling effects. This study examined the effects of shortening the GQ on its clinical utility and psychometric integrity.

Methods. Archival data from 7 previous studies was used, with 2,594 participants in an estimated 455 groups gathered from counseling centers, non-clinical process groups, inpatient psychiatric hospitals, outpatient psychiatric hospitals, and an inpatient state hospital. Participants answered questions from the Group Questionnaire administered during the productive working phase of a group.

Analysis. Analysis was done using multilevel structural equation modeling in Mplus to account for the nested nature of groups. Items were selected using clinical judgment and statistical judgment considering inter item correlation and factor loading. Model fit was analyzed in comparison to the standards in the literature and in comparison to the full length GQ.

Discussion. The revised 12 item GQ has good model fit and acceptable reliability. Further assessment is needed to determine how the reduction affects clinical utility.

Keywords: group psychotherapy, feedback measure, Group Questionnaire, relationship

ACKNOWLEDGEMENTS

Many thanks to Dr. Gary Burlingame, who introduced me to the power of research; to my parents, who supported me in every way, every step of the way; and to my husband, the biggest cheerleader of them all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

The use of measures to track patient progress in routine clinical practice has become a topic of increasing interest in recent years. For instance, the journal *Psychotherapy* published a special issue on Progress Monitoring and Feedback (Hilsenroth, 2015) that described multiple systems to assess outcome and key therapeutic processes that predict treatment success (e.g., therapeutic relationship). Although there are differences between systems, the empirical research summarized in this special issue support their link to improved client outcomes. The routine use of standardized measures to monitor progress and provide feedback to therapists is known as "practice-based therapy," a form of practice-based evidence recognized by major professional organizations (e.g., American Psychological Association, National Academy of Sciences, etc.). In contrast to other evidence-based practice approaches which are guided by diagnosis, practice-based evidence monitoring systems assess the response of clients to the treatments they are receiving and provide feedback that can be used to modify treatment in real-time. Significant research has been done on progress feedback in the individual therapy literature, but research on feedback in group treatment is embryonic. The Group Questionnaire (GQ; Krogel 2013) is one of the few measures of therapeutic relationship in group psychotherapy that was designed to support practice-based therapy or feedback. As group therapy feedback research has progressed, a critical issue in measure development or selection is balancing the amount of information gathered against the time required to obtain such. This paper addresses the issue with the GQ by investigating the psychometric properties of an abbreviated version.

Feedback in Individual Therapy

To set a context for this study, we will first provide a brief introduction to feedback in individual therapy, and then link this to the limited research on feedback measures in group

treatment. Outcome measures when used in a routine fashion (session X session) allow us to track patient progress, informing both therapists and clients about the change process. Instead of relying on a clinician's experience and judgment to gauge client progress and make treatment decisions, feedback from outcome measures provides real-time quantitative information about therapy effectiveness (Lambert 1996). This is especially critical as research has shown that therapists are poor predictors of patient outcome. For example, Hannan and colleagues (2005) asked therapists to predict which of their patients were likely to become worse over the course of therapy. They used an outcome measure to assess client distress at each session and considered deterioration to be a reliable increase from the client's intake score. Hannan found that therapists were able to predict which clients would deteriorate less than 4% of the time, whereas the outcome measure algorithms correctly predicted deterioration 77% of the time. Thus, on their own, clinicians are poor predictors of client outcome.

Therapists are also poor predictors of how their patients perceive the therapeutic relationship. In studies in which therapists, patients, and clinical supervisors were asked to rate the empathy level of the therapist during the same session, results showed no significant agreement among patients, therapists, or supervisors on ratings of therapist empathy (Free, Green, Grace, Chernus, & Whitman, 1985; Squier, 1990). In both studies, only client ratings of therapist empathy were related to eventual outcome. The lack of agreement between therapists' and the clients' perception of therapist empathy is concerning since empathy is one of the best predictors of patient outcome (Elliot et al., 2011).

Outcome measures can not only give a clinician a better prediction of client outcome, but regular feedback can improve client outcome. Lambert and colleagues (2001) demonstrated that therapists who received outcome feedback using the Outcome Questionnaire-45 (OQ-45;

Lambert at al., 1996) had fewer clients ending therapy in a deteriorated state. Therapists who received feedback that a client was off-track for improvement were able to keep them in therapy longer compared to cases where no feedback was given. Additionally, off-track clients in the feedback condition demonstrated more improvement and reduced treatment failure at the end of treatment when compared to clients in the no-feedback condition. When feedback about alliance, motivation for change, and social support were added to the outcome feedback intervention, the proportion of at-risk clients who achieved clinically significant change doubled (Harmon et al., 2007; Hawkins et al., 2004; Lambert et al., 2001, 2002; Slade et al., 2008; Whipple et al., 2003).

Feedback in Group Therapy

Many of the same constructs assessed by feedback measures in individual therapy have a parallel in group therapy. For instance, a common mechanism of change in individual and group therapy is the client's relationship with the therapist (Fuhriman & Burlingame, 1990). However, there are a number of mechanisms of change unique to group therapy, such as the interpersonal environment in group (Fuhriman & Burlingame 1990). Indeed, the group therapeutic relationship has been identified as one of the most significant clinical mechanisms of change in group psychotherapy (Johnson et al., 2005; Johnson, Burlingame, Strauss, & Bormann, 2008) and as the best predictor of treatment outcome (Burlingame, McClendon & Alonso, 2011) making it an ideal candidate for feedback monitoring studies.

The importance of the therapeutic relationship as a mechanisms of change coupled with the empirical support for it predicting outcome led to Davies and colleagues' (2008) feedback study where therapists and clients received weekly feedback from the Group Climate Questionnaire (GCQ; MacKenzie, 1981). The GCQ is a 12-item member-completed measure composed of three subscales: Engagement captures the positive working environment of group,

Conflict assess member anger and rejection, and Avoidance indexes a member's personal responsibility for group work. This team was interested in determining if weekly GCQ feedback to both therapists and members would improve scores on three therapeutic processes (cohesion, insight, and catharsis) assessed by the Curative Climate Inventory (CCI; Fuhriman et al., 1986) and pre-to-post outcome assessed by the OQ-45. There was no association between GCQ feedback and either measure. Davies and his colleagues explained these findings by noting that the GCQ focused on the group as a whole rather than an individual's relation to a group and that previous feedback studies in individual therapy had provided client-specific information (rather than group-specific). They opined that personalized information might be needed for feedback to be more relevant to the client and lead to a reliable effect. Finally, it is important to note that this study did not provide outcome or progress feedback to therapists, leaving the impact of progress feedback in group treatment an open question for future research.

A subsequent study replicated Hannan and colleagues (2005) individual therapy study, testing group clinicians' ability to predict client outcome and the therapeutic relationship. Chapman and colleagues (2012) asked ten group leaders to predict their clients' outcome and perception of therapeutic relationship to see if group leaders were any better or worse than individual therapists. Replicating results from Hannan and colleagues, they found that group leaders were unable to predict member outcome or therapeutic relationship beyond chance. These two studies raise two important considerations: First, not all measures of group processes are good candidates for generating feedback for therapists and clients, and clinicians in a group setting are equally unable to predict client outcome and therapeutic relationships as clinicians in an individual setting.

The Group Questionnaire

The importance of therapeutic relationship in predicting outcome in group treatment (Burlingame et al., 2011), the inability of therapists to accurate predict a client's perception of the therapeutic relationship (Chapman et al., 2012) and the failure of the one of the most popular measures (GCQ) of the therapeutic relationship in group treatment (Burlingame et al., 2004) to be useful in providing therapists with feedback (Davies et al., 2010) created, in part, the theoretical and empirical context for an international cooperation to develop a composite measure of the therapeutic relationship in group treatment. More than 20 measures are available to assess the therapeutic relationship in group treatment (Burlingame, Fuhriman & Johnson, 2002) and in an attempt to assist both clinicians and researchers, the American Group Psychological Association (AGPA) created an international task force in 2002 to recommend group process measures that had the best empirical support in the group treatment literature. The result was the CORE-R battery (Burlingame et al., 2006) which recommended four measures to assess the multi-faceted relationship in group treatment: the Working Alliance Inventory (WAI; Horvath & Greenberg, 1989) and Burns Empathy Scale (ES; Burns & Auerbach, 1996) were recommended to assess the member-leader relationship; the engagement and conflict scales from Group Climate Questionnaire (GCQ; MacKenzie, 1983) were recommended to assess the member-group relationship and; the cohesion scale from the Therapeutic Factors Inventory (TFI; Lese & MacNair-Semands, 2000) was recommended to assess the member-member relationship. The three relationship structures (member-leader, member-group and member-member) assessed by these four measures also represent a dominant conceptualization of the therapeutic relationship in group treatment (Yalom & Leszcz, 2005).

Johnson and colleagues (2005) undertook a study to ascertain if there were underlying latent factors in the four AGPA recommended measures since the 80+ items generated by these measures were too numerous for use in routine clinical practice. They had nearly 700 group members from more than 100 groups assessing both clinical (university counseling centers) and non-clinical (AGPA Institute groups) populations items from all four group measures during the working phase of the group (mid-treatment). Johnson and her colleagues used exploratory factor analysis to identify three latent factors that accounted for the majority of the variance produced by items from the four measures: positive working relationship, positive bonding relationship, and negative working relationship. Positive Bond is the sense of belonging or attraction that a member has to the group, its members, and its leader(s). Positive Work is the ability of the group to agree upon and work toward treatment goals. Negative Relationship is a lack of trust, conflict, and empathic failure that might exist between the group, its members, and leaders. These same measures of the group therapeutic relationship (the "Johnson Model) were subsequently tested on 453 patients from 67 Swiss and German inpatient therapy groups, producing the same three latent factors (Bormann & Strauss, 2007), with further support coming from a study of 424 patients in short- and long-term analytic groups in Norway (Bakali, Baldwin, & Lorentzen, 2009).

The Johnson Model was refined into a shorter, practice-friendly measure by Krogel and colleagues (2013). Empirically redundant items were removed and others were altered to improve clarity, avoid copyright conflict, and eliminate highly correlated items. The resulting 30 item GQ measures the same three quality factors of therapeutic relationship in Johnson's model (positive bonding—PB, positive working—PW, and negative relationship—NR) across the three structural parameters of the group therapeutic relationship (i.e., member to leader, member to

member, and member to group). There are no fewer than three items for the cells created when the quality (PB, PW & NR) and structure dimension (member-member, member-leader & member-group) are crossed (see Figure 1) and sample items for each subscale and cell can be found in Table 1.

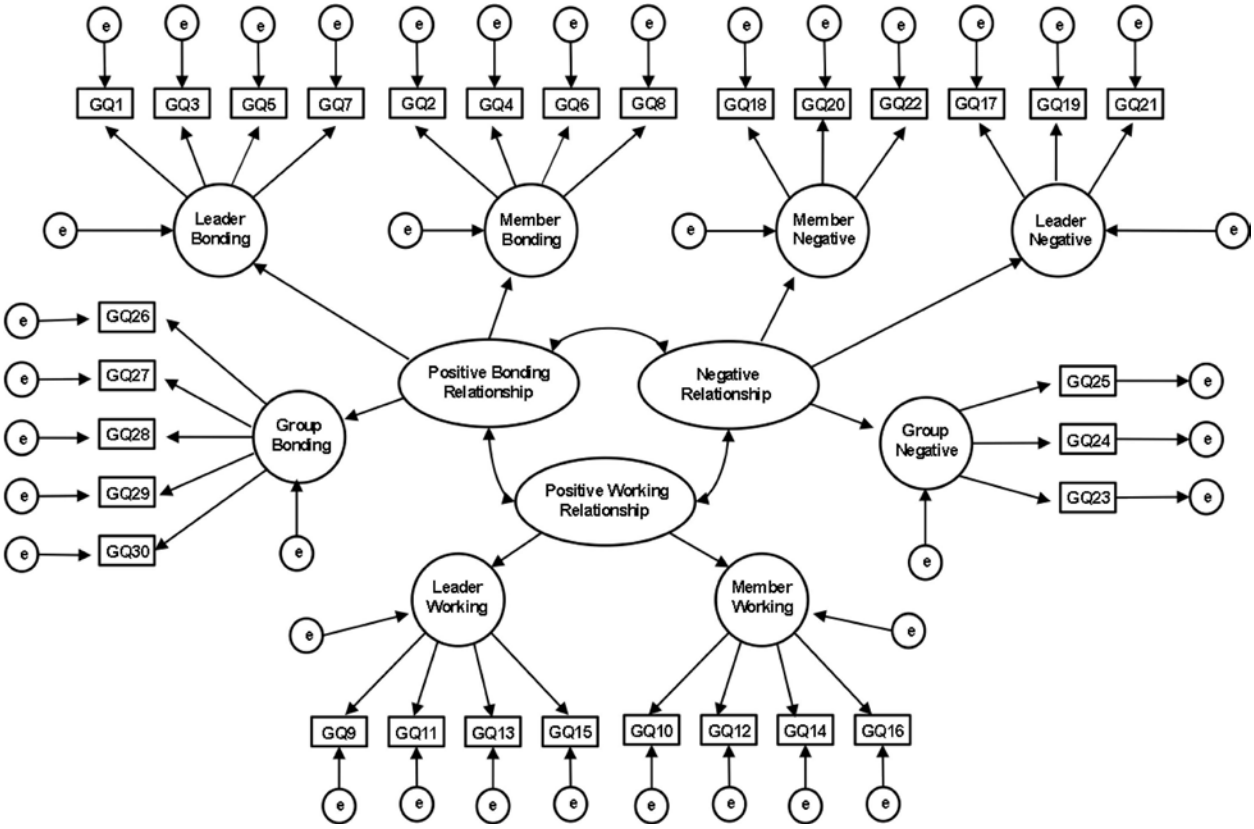

Figure 1 CFA model for the analysis of the GQ's construct validity. Ellipses and circles represent unobserved latent factors. Rectangles and squares represent observed variables. Single-headed arrows represent the impact of one variable on another. Double-headed arrows represent covariances or correlations between pairs of variables.

Table 1 *Sample Questions for the Three Quality Subscales*

Factor	Example Questions
Positive Bond	
Member to Member	I felt that I could trust the other group members during today's session
Member to Leader	The group leaders and I respect each other
Member to Group	We cooperate and work together as a group
Positive Work	
Member to Member	The other group members and I agree on what is important to work on
Member to Leader	The group leaders and I agree about the things I will need to do in therapy
Member to Group	N/A
Negative Relationship	
Member to Member	The other group members did not always understand the way I felt inside
Member to Leader	The group leaders did not always seem to care about me
Member to Group	There was friction and anger between the members

The process of selecting items for the GQ-30 involved two equally important considerations. Items were included/excluded to reduce the redundancy found in the 80+ items in the Johnson model. An equally important consideration was the use of the GQ-30 to support practice-based group therapy. Accordingly, items were selected or modified to be action-

oriented, enabling a clinician to act upon troublesome feedback. For instance, if a client scores a leader as low on a “respect” or “care” item, empirically supported interventions associated with alliance rupture (Safran & Muran, 2011) are recommended. Rather than an overall score, the feedback on low or high scores are provided in the form of a 2-color alert for the three GQ subscales (PB, PW, & NR). One type of GQ alert—absolute alert—enables the therapist to quickly interpret whether a client’s score is in a normative range (green alert) or if it falls outside normative values (red alert; Burlingame et al., 2016). A second GQ alert—relative alerts—indicates that reliable change (Jacobson & Truax, 1991) has occurred in a subscales score since the last GQ administration—typically a session. Relative alerts are viewed as an early warning of a potential problem on a GQ subscale (Burlingame et al., 2016) since past research has shown that reliable negative change temporally precedes red alerts. Both alerts (absolute & relative) are presented in a report each time the GQ is administered with graphs displaying multiple weeks for each subscale (Burlingame & Janis, 2014).

A recent randomized clinical trial by Burlingame, Woodland, & Whitcomb (2014) showed the positive effects of feedback from the GQ-30 on client outcome. Initial findings focusing on red absolute alerts (i.e., patients considered not on track) showed that the proportion of relationship problems (red absolute alerts) decreased over time in feedback groups compared to no-feedback groups. Members in the feedback groups produced fewer alerts over time on all three subscales of the GQ-30 compared to members in groups led by the same leader who did not receive feedback. Thus, receiving GQ feedback appears to improve the therapeutic relationship over time. A session-by-session analysis of red absolute alerts revealed the strongest effects for the negative relationship subscale, followed by positive bond and positive work. More specifically, when therapists received a red alert on the negative relationship subscale, it took one

session for the average members' score to return to green in 100% of the sessions analyzed. In contrast, red alerts persisted in the non-feedback condition in 100% of the sessions analyzed. A similar result was found for positive bond and positive work, but the return to a green alert status took two sessions to be realized for positive bond in 75% of the sessions and in 63% of the sessions for positive.

Current Study

As with all measures administered to clients, group feedback measures experience a tension between being long enough to reliably and validly capture the desired information and being short enough to be practically useful as a repeated measure. Clinical experience, as well as general experience, indicates that most clinicians and patients prefer shorter measures. Most practitioners will not use any measure that takes more than 5 minutes to complete, score, and interpret (Brown et al., 1999). To this end, various ultra-short measures of group therapeutic processes have recently been created, such as the four-item Group Session Rating Scale by Quirk and colleagues (2013) and the eight-item Therapeutic Factor Inventory-8 by Tasca and colleagues (2014). Although shorter measures are more convenient, reliability and validity cannot be sacrificed. The goal of this study was to use a portion of the data from Burlingame and Janis' mega-analysis (2014) to assess the effect of retaining the quality subscales (PB, PW, & NR), eliminating the structure subscales (member-member, member-leader, & member-group) on the factor structure and reliability of the GQ. In order to maintain clinical utility, the shorter GQ must retain a sufficient number of items to provide alerts. Thus, a shortened scale that maintains the Johnson model factor structure and acceptable reliability but loses sufficient information leading to deleterious effects on the alert system is unacceptable given that the GQ

was primarily designed as a practice-based therapy measure. Given these considerations, the research question at hand is two-fold:

- Question 1: Can the GQ be shorter and maintain psychometric integrity?
- Question 2: If a shortened version does not sacrifice psychometric integrity relative to the GQ-30, what effect does shortening have on absolute and relative alerts?

The current study attempts to answer only the first question. Another study will be used to follow up with the second question as necessary.

Hypothesis

Based on the aforementioned research and observations, the following hypothesis was generated:

- Can redundant (high inter-item correlations) GQ items be eliminated without deleterious effects on psychometric properties?

Methods

Participants

Archival data was used from 3 previous studies (Table 2; Johnson et al, 2005; Chapman et al., 2012; and Thayer & Burlingame, 2014). This data contains a total of 1,087 participants in 196 groups. Data was gathered from counseling centers, non-clinical process groups at conferences for the American Group Psychological Association (AGPA), outpatient psychiatric hospitals, and an inpatient state hospital in the United States. University counseling center data was taken from various sites (Johnson et al, 2005; Chapman et al, 2012; Thayer, 2012). Severely mentally ill inpatient data was taken from the Utah State Hospital (Chapman et al, 2012). American Group Psychotherapy Association data was taken from two-day training groups at the 2002 annual meeting of the American Group Psychotherapy Association (Johnson et al, 2005).

Table 2 *Archival data from Three Previous Studies*

Research team	Clinical Setting	Members/Groups
Johnson et al., 2005	- 14 US counseling centers	326/81
	- US nonclinical process groups	336/30
Chapman et al., 2012	- 1 US counseling center	135/20
	- Inpatient state hospital	
Thayer &	- 4 US counseling centers	290/65
Burlingame, 2014	- 1 CMHC	
TOTALS	US counseling centers; nonclinical	1087/196
	process, outpatient & state hospital	

Measures

Group Questionnaire. The Group Questionnaire (GQ; Krogel et al., 2009) is a 30-item self-report questionnaire designed to measure the quality of the therapeutic relationship in group treatment. Items are scored on a 7-point Likert scale ranging from not at all true (1) to very true (7) with some reverse scored items. No overall score is given. Item from the three quality subscales are summed to produce three subscale scores assessing the quality of the therapeutic relationship: Positive Bonding Relationship (13-items; e.g., "I felt that I could trust the group leaders during today's session"), Positive Working Relationship (8-items; e.g., "The other group members and I agree on what is important to work on"), and Negative Relationship (9-items; e.g., "There was friction and anger between the members"). These subscales are measured across three structural dimensions (member to leader, member to member, and member to group). The factorial validity of these subscales has been supported across several studies using inpatient,

outpatient, and nonclinical groups in the United States, Norway, Switzerland, and Germany (Bakali, Baldwin, & Lorentzen, 2009; Bormann & Strauss, 2007; Bormann et al., 2011; Krogel et al., 2013). The reliability estimates (Cronbach's alpha) of the three subscales are .92 for Positive Bond, .90 for Positive Work, and .80 for Negative Relationship. The reliability for Negative Relationship was found to be attenuated by a restriction in range. Using Ghiselli's formula to estimate the reliability of a measure when attenuated by a restriction in range, Negative Relationship rises to a reliability of .90 (Krogel et al., 2013). These reliability estimates were confirmed by Thayer and colleagues (2014).

Procedures

Archival data was used from three previous studies (Table 2; Johnson et al, 2005; Chapman et al., 2012; and Thayer & Burlingame, 2014). Two studies administered the Group Questionnaire (Chapman et al., 2012; Thayer & Burlingame, 2014). Two studies administered the Group Climate Questionnaire, Therapeutic Factors Inventory, and Working Alliance Inventory, which contain all of the items from the GQ-30 (Johnson et al, 2005; Thayer & Burlingame, 2014). One study also included an additional measure not used in this study (Thayer & Burlingame, 2014 included the Severe Outcomes Questionnaire). Data collection occurred part-way through the course of group treatment when it was reasonable to assume the group had entered the productive working phase of the group. During the productive working phase, intimacy, engagement, and cohesion peak and questions about relationship quality become relevant (Yalom & Leszcz, 2005). For most groups this meant that at least three sessions (4.5-6 hours) had passed. For the non-clinical process groups who met all day for two consecutive days, the measures were given at the end of the first full day, after approximately 10.5 hours of group. The measures were administered by a group leader at the end of a session. Most participants

were not incentivized. However, participants from the state hospital were given small snacks as an incentive (Chapman et al., 2012; Krogel et al., 2013). Participants from Thayer and Burlingame (2014) were incentivized with a $20 Amazon.com gift card. Some studies administered measures in multiple waves. In these cases, data was selected from the wave providing the most data so that there was only one set of GQ-30 data per participant. Johnson and colleagues (2005) tested for differences based in measures of group therapeutic relationship based on time and found no differences.

Analysis

This study examined whether the GQ continued to have good fit with the Johnson Model when some subscale items are removed. It is recommended that measures maintain a minimum of four to six items per subscale in order to have acceptable psychometric values (Yang 2009). In its current form, the GQ-30 already had as low as three items in the eight different subcategories created by looking at the three quality subscales across three structural, and it was therefore impossible to reduce in its current structure. To create a shorter measure, the structural subscales (i.e., "member to member," "member to leader," and "member to group") were dropped, leaving a simplified structure with only the three quality subscales (i.e., positive bond, positive work, and negative relationship). Each of the quality subscales was reduced to four to six items for the revised questionnaire. Items in the subscales were considered for removal using a combination of both clinical judgment and statistical consideration of inter-item correlation. Then, the new model was tested for fit using statistical modeling.

Item selection. Item selection involved a balancing act between the inter-item correlations and the judgment of six clinicians with 10 years of experience using the GQ-30 in therapy groups. Clinical judgment was included, rather than relying purely on statistical decision

making, to maintain the same action-oriented item process used to create the GQ-30 (See Figure 2).

The goal of the statistical consideration was to reduce common variance by looking at whether items were highly inter-correlated and to preserve factor loading on the three quality subscales by considering each item's current loading on the GQ-30 subscales. If two items were highly intercorrelated with similar factor loading, clinical judgment was used to select the item that was most useful in telling clinicians what step to take next. If two items were highly intercorrelated, but one had a notably higher factor loading, before discarding the item with lower factor loading clinical judgment was used to ensure that removing the item would not result in a significant loss of value to clinicians. Using this method, several versions of a shortened questionnaire were created with 4, 5, or 6 items per subscale to be tested for model fit.

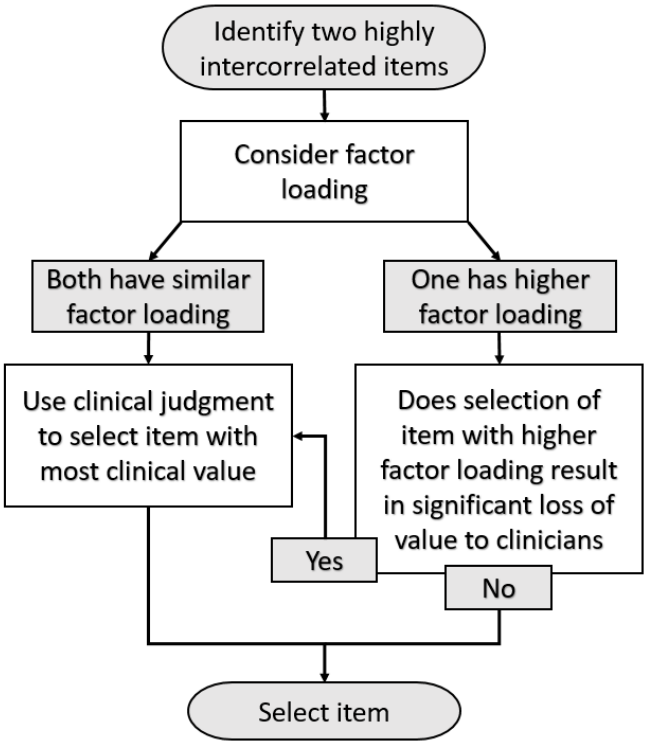

Figure 2 Decision process for item selection

Model fit. Analysis of model fit was done using multilevel structural equation modeling (SEM) in Mplus (used by Johnson et al., 2005 and Thayer & Burlingame, 2014). Multilevel SEM differs from regular or single-level SEM in that it does not assume data are statistically independent and explicitly addresses the degree and effects of intragroup dependency. This is necessary because having participants nested within groups means there is intragroup dependency and therefore measurements are not independent. Ignoring intragroup dependency increases the risk for Type I Error inflation (Baldwin, Murray, & Shadish, 2005; Baldwin et al., 2011; Baldwin, Stice, & Rohde, 2008). Regular SEM estimates the total covariance structure, whereas multilevel SEM allows separate fitting of the between-groups and within-groups covariance structure to take intragroup dependency into account. If the model fits at the within-groups level, then the model adequately describes individual differences from the group means (i.e., individual differences within groups). If the model fits at the between-groups level, then it adequately describes differences among aggregated data (i.e., group means). Past literature suggests that acceptable model fit has a chi-square value of less than twice the model's degrees of freedom and a significant *p*-value (indicating the probability that the observed and estimated matrices are the same); a Root Mean Square Error of Approximation (RMSEA) of .05 or below; a Standardized Root Mean Square Residual (SRMR) for both the within- and between-groups models of .05 or smaller; a Comparative Fit Index of .90 or greater; and a Tucker Lewis Index (TLI) of .90 or greater (Johnson et al., 2005, Krogel et al., 2013, Thayer & Burlingame, 2014).

Results

Item Reduction

Factor loadings and inter-item correlations for the complete 30 items of the GQ-30 were identified by building a two-level model in Mplus and using the archival data set. Taking these

statistical relations into consideration, a panel of researcher clinicians who collaborated on several previous GQ projects (e.g. Krogel et al., 2013; Chapman et al., 2012) and who have used the GQ-30 in their clinical practice from its inception collaborated to identify which items they found most informative and essential. Three theoretical models were created with six, five, and four items per subscale. Starting with the model with six items per subscale, Mplus was used to evaluate model fit and factor loadings. The model with six items per subscale had poor fit, even with added covariances. This was partially due to a cross-loading of item GQ 7 on both the Positive Bond and Positive Work subscales. The model with five items per subscale dropped some of the items with the lowest factor loadings, but still did not achieve satisfactory model fit. However, the model with four items per subscale, dubbed the GQ-12 (Table 3), had good factor loadings and model fit.

GQ-12 Model Fit

The 12 items of the GQ-12 all had good factor loadings on the three subscales. Factor loadings ranged from 0.50 to 0.84 (Table 4). Factor loadings over 0.40 are generally considered acceptable. In the GQ-12, all of the items have factor loadings of 0.50 and above, and two thirds of them have factor loadings of .7 and above. This suggests that the subscale factor structure of the GQ-30 is maintained in the item-reduced GQ-12.

Table 3 *GQ-12 Items by Subscale*

Item	Question Text
Positive Bond	
1	I felt I could trust the leaders during today's session
4	The other members and I respect each other
28	We cooperate and work together in group
30	The group members accept one another
Positive Work	
10	The other members and I agree about the things I will need to do in therapy
11	The leaders and I agree on what is important to work on
14	The other members and I have established a good understanding of the kind of changes that would be good for me
15	The leaders and I are working together toward mutually agreed upon goals
Negative Relationship	
17	Sometimes the leaders did not seem to be completely genuine
22	The other members did not always understand the way I felt inside
24	The members were distant and withdrawn from each other
25	There was tension and anxiety between the members

Table 4 *GQ-12 Factor Loadings by Subscale*

Positive Bond		Positive Work		Negative Relationship	
Item	Factor Loading	Item	Factor Loading	Item	Factor Loading
GQ 1	0.659	GQ 10	0.840	GQ 17	0.575
GQ 4	0.755	GQ 11	0.765	GQ 22	0.502
GQ 28	0.795	GQ 14	0.838	GQ 24	0.586
GQ-30	0.748	GQ 15	0.738	GQ 25	0.499

The GQ-12 had good model fit based on a variety of tested goodness-of-fit indices (Table 5). The GQ-12 exceeded the recommended threshold for the Comparative Fit Index (CFI; 0.94 with a goal of .0.90), Tucker-Lewis Index (TLI; 0.919 with a goal of .0.90), Root Mean Square Error of Approximation (RMSEA; 0.048 with a goal of <0.05), and Standardized Root Mean Square Residual at the within level (SRMR; 0.038 with a goal of 0.05). However, the chi-square value is greater than two times the degrees of freedom, which exceeds the recommended standard. Taken together, these suggest acceptable model fit. This indicates that the data used fits the predicted model well and the three subscale factor is supported for the GQ-12.

Table 5 *Goodness of Fit Indicies – GQ-30 and GQ-12*

GOF Indices	Goal	GQ-12	GQ-30
Chi Square	< 2x degrees of freedom	337.972 for 98 df	775.4, df = 381
	$p < 0.05$	$p = 0.000$	
CFI	> 0.90	0.94	0.957
TLI	>0.90	0.919	--
RMSEA	< 0.05	0.048	0.046
SRMR – within	< 0.05	0.038	--

Reliability

Reliability for the three subscales of the GQ-12 was calculated using Cronbach's alpha, as is common practice in the literature (Table 6). Alpha scores were calculated with Positive Bond at 0.77, Positive Work at 0.86, and Negative Relationship at 0.58. The lowest acceptable threshold for research use allowed by any of the sources reviewed by Ponterotto and Ruckdeschel (2007), with the majority of authors recommending .7 or .8 as the lower level (see Table 7). The standard is even stricter for use with individuals in a clinical setting, where the lowest acceptable threshold is .7, and the majority of authors reviewed recommend a minimum of .85 to .9. This is particularly concerning for the GQ-12, as its primary purpose would be not in a research setting, but a clinical setting. Based on these standards, reliability for the Positive Bond and Positive Work subscales fall within an acceptable range for research purposes, but reliability for the Negative Work subscale is unacceptably low. For clinical purposes, both the Positive Bond and Negative Relationship subscales fail to meet the minimum threshold.

Table 6 *Internal Consistency for GQ-12 and GQ-30 as Cronbach's Alpha*

Subscale	GQ-12	GQ-30
Positive Bond	0.77	0.92
Positive Work	0.86	0.90
Negative Relationship	0.58	.80

Table 7 *Recommended Guidelines for Minimally Acceptable Levels of Internal Consistency*

Purpose	Minimal	Mediocre	Good
Research	0.6—0.7	0.7—0.8	0.8 or above
Clinical	Below 0.85	0.85—0.9	0.9 or above

Recommendations from Ponterotto & Ruckdeschel, 2007

When the GQ-30 was originally created (Krogel et al., 2013) and criterion validity estimated (Thayer & Burlingame, 2014), both studies encountered a lower reliability estimate for the Negative Relationship subscale. Analyses separated by clinical setting revealed discrepancies in the reliability of Negative Relationship. These were attributed to a restriction of range in the responses given by group members treated in university counseling centers. Krogel and Thayer both used the Ghiselli et al. (1981) formula (see Figure 3) to estimate what the reliability of the Negative Relationship subscale would have been had the university counseling center had the same unrestricted range of the inpatient state hospital setting. The same procedure was applied in the present study for the university counseling center members, resulting in an adjusted reliability estimate of 0.82 for the Negative Relationship subscale. The AGPA population was not corrected since it is non-clinical and our study intent is to explore shortening the GQ for clinical populations (i.e. practice-based group therapy).

$$r'_{xx} = 1 - \left[\left(\frac{\sigma x}{\sigma'_x}\right)^2 (1 - r_{xx})\right]$$

Figure 3 Ghiselli et al. (1981) equation to correct for restriction of range in reliability estimates

Table 8 *Adjusted Reliability for the GQ-12 as Cronbach's Alpha*

Subscale	Non-adjusted	Adjusted
UCC		
Negative Relationship	0.58	0.82

As an alternative reliability measure to Cronbach's alpha, omega was calculated for the three subscales using the standardized factor loadings. These were calculated to be 0.83 for Positive Bond, 0.87 for Positive Work, and 0.62 for Negative Relationship. These estimations were based on Bayesian statistics and are therefore the median reliability. However, since the extant literature uses Cronbach's Alpha, these omega values are reported, but not used in analyzing the GQ-12 at this time.

Discussion

The GQ previously existed only in a 30-item format. The recent development of several ultra-short measures of group process prompted the question of whether it were possible to create a much shorter version of the GQ that retained its clinical validity. The present study used statistical analysis and clinical judgment to reduce the number of items from 30 to 12 and then evaluated its statistical properties using two-level confirmatory factor analysis.

This study successfully identified a 12 item model (four items for each of the three subscales) that had good factor loadings and model fit. This indicated that despite the reduced number of items, the quality factors (e.g. positive bond, positive work, negative relationship)

were still supported by the data, although the structural factors (e.g. member to member, member to group, member to leader) dropped out.

Calculation of Cronbach's alpha for internal reliability for the three subscales of the GQ-12 revealed both Positive Bond and Negative Relationship failed to meet the minimum recommended standard for clinical use, and Positive Work barely met the mediocre standard. It is therefore recommended that the GQ-12 not be used in a clinical setting.

The reliability estimates for Positive Bond and Positive Work were acceptable for a research setting, but the estimate for Negative Relationship was unacceptably low. Further analysis by population revealed this low reliability was likely the result of a restriction in range in the university counseling center and AGPA populations. However, when the restriction of range was accounted for, the reliability estimate increased. This suggests that the GQ-12 may in fact be suited for research purposes, particularly when the clinical population does not suffer from restriction of range.

Future Research

It is possible that the low reliability of the GQ-12 could be addressed by substituting highly correlated items eliminated in the vetting process, or through language changes in the existing GQ-12 items. However, an essential but missing piece of the current study is determining how much information was lost in the GQ-12 with respect to absolute and relative alerts. In short, before one invests resources in increasing the reliability of the GQ-12, it is critical to compare the number of absolute and relative alerts produced by the GQ-30 and GQ-12 and how this might affect its value as a practice-based group therapy tool.

Limitations

This study has a risk of experimenter bias due to the selection of which items to eliminate being a combination of empirical and clinical judgment. Clinical judgment was used because the goal of this study is to create a measure that is not only statistically sound but is also empirically useful and action-oriented. As is always the case with clinical judgment, the experimenters bring certain expectancies on which items to choose.

References

Bakali, J. V., Baldwin, S. A., & Lorentzen, S. (2009). Modeling group process constructs at three stages in group psychotherapy. *Psychotherapy Research*, *19*(3), 332–343. http://doi.org/10.1080/10503300902894430

Baldwin, S. A., Murray, D. M., & Shadish, W. R. (2005). Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology*, *73,* 924–935. doi: 10.1037/0022-006X.73.5.924

Baldwin, S. A., Stice, E., & Rohde, P. (2008). Statistical analysis of group-administered intervention data: Reanalysis of two randomized trials. *Psychotherapy Research*, *18*, 365–376. doi:10.1080/10503300701796992

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), 1–20.

Bormann, B., Burlingame, G. M., & Strauß, B. (2011). Der gruppenfragebogen (GQ-D): Instrument zur messung von therapeutischen beziehungen in der gruppenpsychotherapie. *Psychotherapeut*, *56*(4), 297–309. http://doi.org/10.1007/s00278-011-0841-4

Brown, J., Dreis, S., & Nace, D. K. (1999). What really makes a difference in psychotherapy outcome? Why does managed care want to know? In M. A. Hubble, B. L. Duncan, & S. D. Miller (Eds.), *The heart and soul of change: What works in therapy* (pp. 389–406). Washington, DC, US: American Psychological Association.

Burlingame, G., Gleave, R., Beecher, M., Griner, D., Hansen, K., Worthen, V., Jensen, J. (2014). *Administration and scoring manual Group Questionnaire.* Salt Lake City: OQMeasures.

Burlingame, G. & Janis, R. (2014, March). *Differences in the therapeutic relationship on the GQ: Are population specific norms and cut scores necessary?* Paper presented at the annual meeting of the American Group Psychotherapy Association, Boston, MA

Burlingame, G. M., McClendon, D. T., & Alonso, J. (2011). Cohesion in group therapy. *Psychotherapy*, *48*(1), 34–42. http://doi.org/10.1037/a0022063

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., Woodland, S., Whitcomb, K., & Beecher, M. (2014, June). *The effect of therapeutic relationship feedback on group leaders and members: Year 1 results from a GQ/OQ RCT*. Paper presented at the Society for Psychotherapy Research, Copenhagen, Denmark.

Burlingame, G.M., Furhiman, A., & Johnson, J. (2002). Cohesion in group psychotherapy. In J. Norcross (Ed.), *A guide to psychotherapy relationships that work* (pp. 71-88). Oxford: Oxford University Press.

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: 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*(6), 673–681. http://doi.org/10.1080/10503307.2012.702512

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*(2), 141–154. http://doi.org/10.1037/1089-2699.12.2.141

Free, N. K., Green, B.L., Grace, M. D., Chernus, L. A., & Whitman, R. M. (1985). Empathy and outcome in brief, focal dynamic therapy. *American Journal of Psychiatry*, *142*, 917-921.

Fuhriman, A., Drescher, S., Hanson, E., & Henrie, R. (1986). Refining the measurement of curativeness: An empirical approach. *Small Group Behavior, 17,* 186–201.

Fuhriman, A., & Burlingame, G. M. (1990). Consistency of matter: A comparative analysis of individual and group process variables. *The Counseling Psychologist*, *18*, 6–63. doi:10.1177/ 0011000090181002

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. WH Freeman.

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: In Session*, *61*, 155-163.

Harmon, S. C., Lambert, M. J., Smart, D. M., Hawkins, E., Nielsen, S. L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist–client feedback and clinical support tools. *Psychotherapy Research*, *17*(4), 379–392. http://doi.org/10.1080/10503300600702331

Hawkins, E. J., Lambert, M. J., Vermeersch, D. A., Slade, K. L., & Tuttle, K. C. (2004). The therapeutic effects of providing patient progress information to therapists and patients. *Psychotherapy Research*, *14*(3), 308–327. http://doi.org/10.1093/ptr/kph027

Hilsenroth, M. (Ed.). (2015). Progress monitoring and feedback [Special issue]. *Psychotherapy*, *52*(4).

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, *36*, 223–233.

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.

Johnson, J. E., Burlingame, G. M., Olsen, J. A., 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*(3), 310–321. http://doi.org/10.1037/0022-0167.52.3.310

Johnson, J. E., Burlingame, G. M., Strauß, B., & Bormann, B. (2008). Die therapeutischen beziehungen in der gruppenpsychotherapie. *Gruppenpsychotherapie Und Gruppendynamik*, *44*(1), 52–89. http://doi.org/10.13109/grup.2008.44.1.52

Krogel, J. (2008). The Group Questionnaire: A new measure of the group relationship. *All theses and dissertations*. Retrieved from http://scholarsarchive.byu.edu/etd/1732

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*(3), 344–354. http://doi.org/10.1080/10503307.2012.729868

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy*, *3*(4), 249–258. http://doi.org/10.1002/(SICI)1099-0879(199612)3:4<249:AID-CPP106>3.0.CO;2-S

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research*, *11*(1), 49–68. http://doi.org/10.1080/713663852

Lambert, M. J., Whipple, J. L., Vermeersch, D. A., Smart, D. W., Hawkins, E. J., Nielsen, S. L., & Goates, M. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical Psychology & Psychotherapy*, *9*(2), 91–103. http://doi.org/10.1002/cpp.324

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.

Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G. V., & Finch, A. E. (2002). Comparison of empirically derived and rationally derived methods for identifying clients at risk for treatment failure. *Clinical Psychology and Psychotherapy, 9*, 149–164.

Lese, K. P., & MacNair-Semands, R. R. (2000). The Therapeutic Factors Inventory: Development of a scale. *Group, 24*, 303–317.

MacKenzie, K. R. (1981). Measurement of group climate. *Journal of Group Psychotherapy, 31,* 287–296.

MacKenzie, K. R. (1983). The clinical application of group measure. In R. R. Dies & K. R. MacKenzie (Eds.), *Advances in group psychotherapy: Integrating research and practice* (pp. 159–170). New York: International Universities Press.

Ponterotto, J. G., & Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills*, *105*(3), 997-1014.

Quirk, K., Miller, S., Duncan, B., & Owen, J. (2013). Group Session Rating Scale: Preliminary psychometrics in substance abuse group interventions. *Counselling and Psychotherapy Research*, *13*(3), 194-200.

Safran, J. D., Muran, J. C., & Eubanks-Carter, C. (2011). Repairing alliance ruptures. *Psychotherapy*, *48*(1), 80.

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*(3), 298–311. http://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 electronic feedback and revised clinical support tools. *Journal of Clinical Psychology and Psychotherapy, 15*(5), 287-303.

Squier, R. W. (1990). A model of empathic understanding and adherence to treatment regimens in practitioner-patient relationships. *Social Science and Medicine*, *30*(3), 325-339.

Tasca, G. A., Cabrera, C., Kristjansson, E., MacNair-Semands, R., Joyce, A. S., & Ogrodniczuk, J. S. (2014). The therapeutic factor inventory-8: Using item response theory to create a brief scale for continuous process monitoring for group psychotherapy. *Psychotherapy Research*, *26*(2), 131–145. http://doi.org/10.1080/10503307.2014.963729

Thayer, S. D., & Burlingame, G. M. (2014). The validity of the Group Questionnaire: Construct clarity or construct drift? *Group Dynamics: Theory, Research, and Practice*, *18*(4), 318–332. http://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*(1), 59–68. http://doi.org/10.1037/0022-0167.50.1.59

Yalom, I.D., & Leszcz, M. (2005). *The theory and practice of group psychotherapy* (5th ed). New York: Basic Books.

Yang, C., Nay, S., & Hoyle, R. H. (2010). Three approaches to using lengthy ordinal scales in structural equation models: Parceling, latent scoring, and shortening scales. *Applied Psychological Measurement*, *34*(2), 122–142. http://doi.org/10.1177/0146621609338592.