Psychotherapy Research
2009, 1–12, iFirst article

Generalizability theory in psychotherapy research: The impact of multiple sources of variance on the dependability of psychotherapy process ratings

RACHEL H. WASSERMAN, KENNETH N. LEVY, & ERIC LOKEN

The Pennsylvania State University

(Received 19 May 2008; revised 13 October 2008; accepted 20 October 2008)

Abstract
A central task of psychotherapy process measurement development is the assessment of reliability and validity. The convention of reporting intraclass correlations (ICCs) for coding procedures assumes that variance in scores can be adequately explained by differences between individuals and error resulting from differences in coders. Given the complex coding procedures that are common in psychotherapy process studies (multiple sessions may be rated by multiple coders on one or more multi-item scales), an ICC may fail to account for all of the relevant sources of variability in obtained scores. If process studies are to provide useful information about the mutative agents in psychotherapy, assessment procedures must be developed that dependably assess the constructs of interest. Generalizability theory provides a framework within which multiple sources of error can be simultaneously evaluated, thus improving the accuracy of reliability estimates and providing critical information for modification and improvement of coding procedures. To illustrate the applicability of generalizability theory to psychotherapy process research, the authors present the design and findings of a study investigating the generalizability of technique scales in the Psychotherapy Process Rating Scale for Borderline Personality Disorder. Implications for measurement development and procedural modifications are discussed.

Keywords: statistical methodology; process research; outcome research

A crucial task in the development of psychotherapy process measures is the assessment of reliability and validity. Although psychotherapy researchers have consistently reported intraclass correlations (ICCs) for their coding procedures, this statistic (based on classical test theory) only takes into account the variability associated with differences in ratings between coders (Tinsley & Weiss, 1975). Given the complex coding procedures that are common in psychotherapy process studies (multiple sessions may be rated by multiple coders on one or more multi-item scales), an ICC may fail to account for all of the relevant sources of variability in obtained scores. If process studies are to provide useful information about the mutative agent in psychotherapy, assessment procedures must be developed that dependably assess the constructs of interest. Generalizability theory (G-theory) provides a framework within which multiple sources of measurement error can be simultaneously evaluated (Shavelson, Webb, & Rowley, 1989; Shavelson & Webb, 1991), thus improving the accuracy of reliability estimates and providing critical information for modification and improvement of coding procedures.

Generalizability Theory

G-theory represents a set of techniques that can be used for assessing the extent to which a set of measurements generalize to a more extensive set of measurements. In this way, G-theory provides information regarding both the reliability and the validity of behavioral measures (Hintze & Matthews, 2004). Although G-theory has not been widely applied to observer ratings of psychotherapy process variables (see Hoyt, 2002, for discussion), many studies have been conducted using this technique for other types of observer ratings (Hintze & Matthews,

Rachel H. Wasserman and Kenneth Levy, Department of Psychology, and Eric Loken, Human Development and Family Studies, The Pennsylvania State University, University Park, Pennsylvania, USA.
Correspondence concerning this article should be addressed to Rachel H. Wasserman, Department of Psychology, 115C Moore Building, The Pennsylvania State University, University Park, PA 16802, USA. E-mail: rwasserm@psu.edu; and Kenneth N. Levy, Ph.D., Department of Psychology, 521 Moore Building, The Pennsylvania State University, University Park, PA 16802, USA. E-mail: klevy@psu.edu

ISSN 1050-3307 print/ISSN 1468-4381 online
DOI: 10.1080/10503300802579156

2004; O'Brian, O'Brian, Packman, & Onslow, 2003). In addition, G-theory has the potential to answer a number of important questions relevant to psychotherapy process research. Primarily one can answer the question, Does the coding procedure used dependably assess differences in the construct of interest? In much the same way that one would hesitate to use a single measurement of an assessment device with low test–retest reliability to assess a stable trait, if a coding procedure demonstrates low dependability, one should hesitate to draw substantive conclusions from such data. After answering this initial question, one may then ask what aspects of the coding procedure create the greatest proportion of error in measurement. By identifying sources of error, steps can be taken to minimize error through modification of coding procedures (e.g., inclusion of more coders or additional assessment points) or refinement of assessment measures themselves.

For a systematic treatment of G-theory and its applications, see Cronbach, Gleser, Nanda, and Rajaratnam (1972), Shavelson and Webb (1991), and Brennan (1992). The subsequent overview is meant to provide basic terminology and techniques of G-theory. After this brief review, G-theory techniques are applied to a complex and multifaceted psychotherapy process study. Procedural considerations, results, and conclusions are reviewed to illustrate the kinds of findings that may be derived from the application of G-theory to observational coding procedures. A broader discussion of the applications of G-theory then follows.

What Is G-Theory?

As described earlier, G-theory provides a framework within which multiple sources of error in a given set of measurements can be simultaneously estimated. As such, G-theory extends classical test theory in a similar way to how factorial analysis of variance (ANOVA) extends one-way ANOVA (Cranford et al., 2006). Within an item response theory framework, similar extensions to multifaceted designs have also been considered (Mellenbergh, 2001; Verhelst & Verstralen, 2001). By assessing reliability and error within a context of a multifaceted testing situation, the researcher can determine portions of error, which can be accounted for by various aspects of the assessment procedure. In this way, G-theory provides a researcher with information necessary for determining how many occasions, coders, questionnaire forms, or questions are needed to obtain dependable scores. *Dependability* refers to the accuracy with which one can generalize from a particular observed score on a given construct to the ideal mean score a person would have received across all acceptable observation contexts. In a psychotherapy process study, the *universe of admissible observations* includes all scoring contexts the researcher is willing to accept as interchangeable; such a universe may include observations of differing coders, varying assessment points, or alternate test forms. A universe is defined by its major sources of variation. Each discrete source of variation is referred to in G-theory as a *facet*. In a one-facet universe all systematic variance is assumed to come from one source, whereas in a multifacet universe multiple sources of systematic variance are present. Typically, an ICC is calculated to assess, for example, the reliability of coders or items in the context of a one-facet design. The task is to estimate measurement error resulting from coder variance, or item variance, but not both sources simultaneously. G-theory allows for the calculation an ICC for ANOVA designs that have more than one factor (facet).

Sources of Variability

Although there is no theoretical limit to the number of conditions under which observations can be made, Cone (1977) has proposed six universes that are most relevant to behavioral assessment: coder, time, method, setting, dimension, and item. Hintze and Matthews (2004) noted that each of these universes has links to traditional notions of reliability and validity: Coder generalizability is consistent with interobserver agreement; time generalizability roughly represents test–retest reliability; and item generalization is approximately equivalent to internal consistency or construct validity. The use of multiple measures or methods of quantifying the same construct also allows for assessment of construct validity. Halvorsen, Hagtvet, and Monsen (2006) noted that therapist characteristics and treatment sites may also be relevant facets for consideration in psychotherapy research. In addition to the main effects of each facet under consideration, interactions of facets provide further information about variance associated with particular combinations of facet levels. For example, let us take a two-facet universe defined by a coder facet and an occasion facet (analyzed by a three-factor random-effects ANOVA, where patient is the third random factor). Here, the patient factor represents the entire universe of score variability.[1] A main effect from the patient factor represents interindividual differences on the measure/construct of interest. The larger or more robust these differences are relative to other sources of variation, the greater the likelihood of dependable assessment. The main effect for the coder facet represents the constant effect for all persons resulting from the stringency or leniency of coders.

If coder variance is high, then some coders tend to perceive targets leniently, whereas others tend to perceive targets severely or stringently. The main effect for the occasion facet represents the constant effect for all persons resulting from the variability in behavior from one assessment occasion to another. If occasion variance is high, then there is a wide range in scores across time. In addition to these main effects, there are four interaction terms. The interaction of the patient facet with the coder facet represents inconsistencies of coder evaluations of a particular individual's behavior. This interaction is sometimes called dyadic variance (see Hoyt, 2002, for discussion) and can be understood as differences in a particular coder's perceptions of patients. This effect represents coders' reactions to certain types of patients and can, in some cases, be thought of as coder countertransference. The interaction of the patient facet with the occasion facet represents variability from one occasion to another in the assessment of a particular patient's behavior. This interaction reflects inconsistencies across patients in their degree of change over time. The interaction of the coder facet with the occasion facet represents differences in coders' stringency from one occasion to another (this is constant across patients). When this term is large, it may indicate coder drift. Last, the three-way interaction of the patient, coder, and occasion facets is here conflated with any unmeasured facets that may affect measurement and/or random events. In all measurement designs, the highest order interaction is always conflated with remaining unmeasured error.

Fixed Versus Random Effects

One feature of facets is their distinction as either random or fixed. Facets are considered random when the size of the sample is much smaller than the size of the universe and the sample is considered to be interchangeable with any other sample of the same size drawn from the universe. Shavelson and Webb (1991) suggest that, in deciding whether a set of conditions (within a given facet) should be considered random, one should ask whether one is willing to exchange the observed conditions for any other same-size set of conditions from that universe. If the answer is yes, then the facet may be treated as random. If not, the conditions should be treated as fixed. Within a psychotherapy process study, coders would be treated as a random effect if one seeks to make generalizations across a broader sample of individuals who might be coders in future studies. In contrast, if researchers only seek to draw conclusions about the existing coders (e.g., if only the creator of a treatment assesses adherence), the facet may be considered a fixed effect. Another type of fixed effect occurs when the measured conditions exhaust the universe of generalizability. If one was interested in variability associated with therapists in a small clinic in which only four therapists work (and are not anticipated to leave), this variable would be treated as a fixed effect because the number of observed conditions is equal to the number of conditions in the universe of generalization. For fixed effects, generalizability coefficients may be either estimated separately for each level of the facet or averaged across all levels. If one wishes to draw conclusions at each level of the facet (e.g., for men and for women independently instead of about people in general), then generalizability should be estimated separately. If one wishes to draw conclusions about overall performance across already specified and limited domains (e.g., total therapist activity as a composite of validating comments, questions, and interpretations), then generalizability should be aggregated across the levels of the fixed effect (here, specific techniques).

Crossed Versus Nested Design

A particular coding procedure is described as crossed or nested depending on the relationship of the facets to one another. In a crossed design, all levels of one facet must be assessed under all levels of the crossed facet. For example, for coders to be crossed with occasions, all coders must rate all occasions (psychotherapy sessions). Because of time, location, or other logistical considerations, it may only be possible for certain coders to code occasions. In this case, this design would be considered nested. A facet is nested within another facet when two or more of the levels of the nested facet appear with one and only one condition of another facet. For example, coders would be nested within occasions if coders A and B rate a subset of the occasions while coders C and D rate a separate subset of occasions (Shavelson & Webb, 1991). In a fully crossed design, every facet is crossed with every other facet (e.g., each person is scored by the same two coders on three occasions). When at least one facet is not crossed (e.g., each person is scored by two different coders at each of three occasions), the design is nested. In this case, occasions are crossed with persons, but coders are nested within occasions. Nested designs provide less specific information than full crossed designs because the effect of the nested variable cannot be differentiated from its interaction with the facet within which it is nested. In the prior example, the effect of the coder facet cannot be differentiated from its interaction with occasion of measurement. For this reason, it is desirable to use

fully crossed designs whenever possible or to maximize the number of crossed facets in order to estimate the greatest number of distinct sources of variability possible (Cronbach et al., 1972).

Relative Versus Absolute Decision

In an analogous fashion to classical test theory and the application of ICCs, G-theory further distinguishes between decisions based on the relative standing or ranking of individuals (relative interpretations) and decisions based on the absolute level of their scores (absolute interpretations). Because correlations are affected by the relative standing of individuals, not by their absolute level of performance, relative decisions are often most relevant for psychotherapy process studies wishing to relate process variables to one another or to aspects of the outcome. In a situation where an absolute cutoff is being used for decision-making purposes (e.g., a minimum Beck Depression Inventory score required for admission to a depression treatment study), an absolute decision rule would be required.

For a relative decision, all variance components that influence the relative standing of individuals contribute to the error term; these are the interactions of each facet with the object of measurement (in this case, persons). In contrast, for absolute decisions all variance components except the object of measurement itself contribute to error (this includes the main effects of coder and occasion as well as all four interactions detailed previously). As described by Brennan (2001), these two ways of defining measurement error suggest two reliability-like coefficients. With respect to relative decisions, the generalizability coefficient estimates the extent to which consistency of scores is affected by relative error. In the prior two-facet design example, the generalizability coefficient would be the ratio of patient variability to the sum of patient variability and relative error: patient variance / (patient variance + relative error). In contrast, the dependability coefficient estimates the extent to which consistency of scores is affected by absolute error. In the same example, the dependability coefficient would be the ratio of patient variability to the sum of patient variability plus absolute error: patient variance / (patient variance + absolute error).

G-Studies and D-Studies

In G-theory a distinction is made between two types of studies: generalizability studies (G-studies) and decision studies (D-studies). The primary goal of a G-study is to estimate the effects of as many potential sources of error as possible. In this way, a G-study attempts to define the universe of admissible observations as broadly as possible (Shavelson & Webb, 1991). G-studies estimate variance components associated with the main effects and interactions as their primary results. Hoyt and Melby (1999) note that this is a contrast to the traditional emphasis of reliability studies, which tend to focus on summary statistics by reporting a single reliability coefficient. Although G-studies frequently report generalizability coefficients in their results, Cronbach et al. (1972) point out that the generalizability coefficient reported depends on the measurement procedures to be used in a D-study. A D-study, in turn, makes use of the information gathered in a G-study (in the form of variance component estimates) toward two ends: (1) to quantify the dependability of a set of measurement parameters and (2) to determine the best possible design to draw conclusions about the targets of measurement in a subsequent study.

Shavelson and Webb (1991) outline three primary steps in the application of G-theory techniques. First, the researcher must define the universe of generalization; this involves determining the number and levels of facets to be generalized over. In the ongoing example, the researcher would define two facets: coders and occasions. The number of patients, coders, and occasions to be used would then also be specified. In the first D-study, a researcher may use the number of facet levels that were actually used/collected in the study from which the G-study data came. Second, the researcher must specify the proposed interpretation of the results of the D-study. Here, the researcher must determine whether an absolute or relative decision rule is most applicable to the broader research questions. The choice of a relative versus absolute decision rule will determine the way measurement error is defined. Last, the researcher uses variance components estimates (the main effects and interactions of the facets) to evaluate the effectiveness of alternative designs. Here, the goal is to minimize error and maximize reliability by systematically varying the assessment design. In the ongoing example, the researcher used two coders on three occasions. In the D-study phase, the researcher would be able to determine whether having more coders and/or more assessment occasions would maximize the dependability of construct assessment. In summary, a G-study provides information needed to flexibly compute generalizability coefficients relevant to a wide range of potential D-study designs. A D-study is a substantive investigation that makes use of the results of a G-study to

optimize procedures for a specific application of a measure (Cronbach et al., 1972).

Applied Example

To illustrate the application of G-theory to psychotherapy process research, we conducted a G-study and D-study on data from a National Institute of Mental Health-funded (PI: John F. Clarkin) treatment development study examining pre–post changes observed in the 1-year outpatient treatment of 17 patients with borderline personality disorder (BPD) undergoing transference-focused psychotherapy (TFP; see Clarkin et al., 2001). As part of the treatment development study, psychotherapy sessions were videotaped. Sessions were coded using the Psychotherapy Process Rating Scale for Borderline Personality Disorder (PPRS-BPD; Levy, Wasserman, Clarkin, Eubanks-Carter, & Fisher, 2005). The PPRS-BPD was designed to assess specific observable key therapeutic techniques and facilitative behaviors in the psychotherapy process with patients specifically diagnosed with BPD so as to allow for the examination of the relationship between psychotherapy techniques and outcome. The PPRS-BPD is designed to be used with audio- or videotaped records of a single treatment session as the unit of observation. Items were designed to reflect the treatment techniques and patient–therapist process in TFP (Levy et al., 2006), as well as other common treatments for BPD such as dialectical behavior therapy (Linehan, 1993; Lynch, Chapman, Rosenthal, Kuo, & Linehann, 2006) and supportive psychotherapy (Appelbaum & Levy, 2002; Rockland, 1992). In addition, there are items to assess both nonspecific common factors and techniques specifically prohibited. Each item is rated on a 9-point Likert scale, from least to most characteristic of the session. Because the PPRS-BPD was applied to patients in TFP, we present findings for two scales specific to TFP: transference interpretation and maintenance of the treatment frame.

Coders were four advanced clinical psychology doctoral students trained in a group format for 2 hr/week over a 4-month period to reach adequate prestudy reliability (an average measure two-way mixed ICC with absolute agreement, ICC(3,4)> .70). Levy et al. (2006) reported training procedures and data from the training phase, during which all four coders evaluated 10 "calibration" tapes and had an overall ICC of .93 across all items and scales. We selected six sessions from each patient's treatment: Two sessions were randomly selected from the first 3 months of therapy, two from the middle portion of the therapy (Months 5–7), and two from the latter portion of the therapy (Months 9–12) to ensure adequate representation of the process across the year. Two coders (of the four) were randomly selected to code each psychotherapy session. Each coder completed both scales for each session; each scale had invariant items.[2]

Based on this coding procedure, the G-study was designed as follows: Sessions, coders, scales, and items are facets of generalization, yielding a four-facet design. The five-factor random-effects ANOVA based on this design provides an overall estimate of the magnitude of variability across 1 year of treatment. Coders are treated as a random effect because the goal is to be able to select any same-size sample of trained coders to rate a particular session. Scales are treated as a fixed effect; we do not seek to generalize beyond these two scales. Last, items are treated as a random effect because similar items from a universe of possible indicators of each technique could replace the current items.

All random effects are estimated simultaneously, while fixed effects are treated and reported separately. That is, G- and D-studies are conducted on each scale individually. Variance components are, therefore, estimated for persons, sessions, coders, and items in subsequent G-studies. According to the present coding scheme, sessions are crossed with patients, scales are crossed with sessions, and items are nested within scales. Because scales are being treated separately, items are functionally crossed with sessions, because each item is assessed for all sessions and all persons. Last, coders are neither fully crossed nor fully nested in this design. A fully crossed design would have all four coders rate each session, with both scales, with all items. In the existent data set, two coders were randomly selected from the four trained coders because of monetary and time constraints. Because each person and session could be coded by any pair of coders, this does not represent a classic nested design, in which two pairs would consistently code a subset of persons or sessions each, without overlap. In the partially nested design in the current study, the main effect of coder is conflated with the interaction of coder with the facets within which it is nested.

As a first step in the G-study, the variances associated with each of the facets described previously are estimated from the data. Variance components are estimated using the VARCOMP procedure of SPSS 15.0 (SPSS Inc., 2006). Had there been no missing data, this procedure would have yielded identical estimates as the analysis of the expected mean squares for the three-way mixed-model ANOVA designs (Shavelson & Webb, 1991). The VARCOMP procedure uses as a default the minimum norm quadratic unbiased estimator

(MINQUE; Rao, 1971) method to make use of information from persons who have one or more missing sessions.[3]

In the second step, the variance component estimates from the G-study are used to make inferences about the quality of the measurements made based on the actual coding procedure used. In this initial potential D-study design, a relative decision rule will be applied because the intended use of these data is correlation. As discussed earlier, relative decision rules are most relevant for correlational designs in which the relative standing of individuals, not the absolute level of performance, is the type of data to be analyzed. In the third step, additional potential D-study designs are investigated to determine the effect of systematically varying the number of sessions, coders, and items on generalizability coefficients.

Variance Components Decomposition

The first step of the G-study involves estimating the variability associated with each facet of generalization. The purpose of this analysis was to calculate variance components associated with patients, sessions, coders, and items (within scales) in addition to the interaction among these effects for each scale. Table I shows the results of the variance decomposition analyses for transference interpretation and maintenance of treatment frame. Proportion of total variance for each effect is reported within parentheses as a standardized index of variability across scales.

Two effects account for the largest proportion of variation in scores across both scales. The first is the item facet, which reflects the variability of ratings from one item to another (similar to internal consistency). Accounting for, on average, 35% of the total variation in scores, this finding indicates that there was a great deal of variation of scores on items within scales. The second largest proportion of variance came from the four-way interaction term (which is also the error term), which, on average, accounted for 20% of variation of scores. This suggests that a substantial portion of the variation in scores is not accounted for by the four main effects and their higher order interactions and remains unexplained by the specified facets of generalization. The main effects for patient, session, and coder were next considered. The patient variance component represents the entire universe of score variability and is the desired target of measurement. On average, the patient effect accounted for 7.5% of variation in scores.

Table I. Variance Decomposition Analyses for Transference Interpretation and Maintenance of Treatment Frame

Variable	Transference interpretation	Treatment frame
Patient	.394 (.08)	.431 (.07)
Session	.000 (.00)	.011 (.00)
Coder	.005 (.00)	.000 (.00)
Item	1.12 (.22)	3.15 (.48)
Patient × Session	1.08 (.21)	.383 (.06)
Patient × Coder	.038 (.01)	.000 (.00)
Patient × Item	.187 (.04)	.241 (.04)
Session × Coder	.038 (.00)	.033 (.00)
Session × Item	.054 (.01)	.000 (.00)
Coder × Item	.035 (.01)	.419 (.06)
Patient × Session × Coder	.301 (.06)	.074 (.01)
Patient × Session × Item	.501 (.10)	.688 (.10)
Patient × Coder × Item	.097 (.02)	.139 (.02)
Session × Coder × Item	.022 (.00)	.054 (.01)
Residual (Patient × Session × Coder × Item)	1.32 (.25)	.994 (.15)

Note. Percentage of total variance for each effect is reported within parentheses as a standardized index of variability across scales.

Variability in scores as a function of coders has traditionally been approached by the use of ICCs as a measure of interobserver agreement.[4] The coder facet represents the constant effect for all persons resulting from the stringency of different coders. In the present analyses, coder variance accounted for less than 1% of the total variability in scores; this finding is consistent with the high ICCs obtained during the training period.[5] Session variability, on average, accounted for less than 1% of total variability in scores.

Among the interaction terms, three accounted for substantial portions of total score variance. The Patient × Session interaction accounted for, on average, 13.5% of variance. Specifically, a large proportion of variance (21%) within transference interpretation was accounted for by this effect. The Patient × Session × Item interaction term represents variability in scores for different patients across time. Although both the main effects for patients and session were, on average, quite small, the more substantial variances associated with this interaction term suggest that for some individuals transference interpretations varied as a function of time, whereas for others no such systematic variation occurred. The Coder × Item interaction term represents inconsistencies in coders' scoring of particular items. For the transference scale, less than 1% of variability was accounted for by this interaction compared with 6% for maintenance of the treatment frame. The low variance component for transference interpretations may reflect the particular attention paid to training coders on this scale, because it was the technique hypothesized to be unique and specific to TFP. The Patient × Session × Item interaction accounted for 10% of variability in scores. The magnitude of this

effect indicates that for certain patients, at certain sessions, items were a substantial source of variability.

Generalizability Coefficients

The second step of the G-study involved calculating generalizability coefficients, which are analogous to the reliability coefficients in classical test theory, based on the obtained variance component estimates. (For a full discussion of the selection of appropriate variance components, see Brennan, 1992; Shavelson & Webb, 1991) As discussed earlier, a relative decision rule was used in the estimation of the error term and the generalizability coefficients. The variance components included in the relative error were the interaction terms, including the person effect. Relative error involves only those effects that impact the standing of individuals with respect to one another and does not take into account those effects that affect the absolute score observed. For example, even though the variance component for item is large, the effect is assumed to affect all patients and is not considered because the impact is felt across all patients, and it is, therefore, not relevant to the relative error or the dependability of measurement (as indexed by the generalizability coefficient). For each variance parameter included in the error term, its contribution to the error is a function of not only the variance but the number of levels of the facet. For example, in the Patient $\times$ Session interaction, the variance component for that interaction is divided by the number of sessions for which data are collected to reflect the expected reduction in total variance resulting from the aggregation across multiple sessions. The generalizability coefficient represents the expected between-person dependability estimate for a coding scheme in which six sessions are rated by two randomly selected coders on all items for both scales of the PPRS-BPD. For a relative decision, the generalizability coefficient is as follows:

$$= \frac{\sigma^2_p}{\sigma^2_p + \left(\frac{\sigma^2_{ps}}{n_s}\right) + \left(\frac{\sigma^2_{pc}}{n_c}\right) + \left(\frac{\sigma^2_{pi}}{n_i}\right) + \left(\frac{\sigma^2_{psc}}{n_s n_c}\right) + \left(\frac{\sigma^2_{psi}}{n_s n_i}\right) + \left(\frac{\sigma^2_{pci}}{n_c n_i}\right) + \left(\frac{\sigma^2_{psci}}{n_s n_c n_i}\right)}$$

The generalizability coefficient from the obtained study design for transference interpretation was .591. The generalizability coefficient for maintenance of the treatment frame was .736. Although the thresholds for dependability vary as a result of the goals of a particular study, a minimum of .7 is generally considered adequate for interpersonal and observationally coded constructs (Allen & Yen, 1979). As such, modifications to the coding procedures for both scales are warranted and are strongly indicated for transference interpretations. To understand the implications of the obtained generalizability coefficients, it is useful to evaluate them in the context of the variance components estimates. Because a generalizability coefficient is a ratio of person variance to total variance (the sum of person variance and relative error), a low generalizability estimate results whenever error is large relative to person variance. In more concrete terms, this may occur if there is a lot of error in measurement or if there is minimal variation across individuals. In the present context, both of these issues are present. Variance associated with patients accounted for, on average, only 7.5% of variance. Variance components decomposition also demonstrated that a number of affects accounted for large amounts of variation in scores. Although both the main effect of item and the interaction of item with coder accounted for large proportions of variation in scores, neither of these variance components are used to compute relative error and, therefore, do not affect dependability estimates in the present study. Although neither coder nor session variance accounted for large proportions of variance, higher order interactions (Patient $\times$ Session and Patient $\times$ Session $\times$ Item) were significant contributors to overall variance and relative error. To the extent that these sources of error were the cause of a lowered generalizability coefficient, modifications to the measurement procedures may improve dependability. In the next section, the effects of modifications to the number of sessions, coders, and items (within scales) are explored with respect to their impact on dependability of measurement.

Decision Study

A unique advantage of generalizability theory is its application in decision studies. In addition to examining the dependability of measurement based on an observed study design, it is also possible to isolate

Table II. Potential Decision Study Designs for Transference Interpretations

	Number of sessions			
Variable	1	3	6	12
1 coder	0.193	0.397	0.541	0.661
2 coders	0.219	0.441	0.591	0.713
3 coders	0.229	0.458	0.610	0.732
6 coders	0.241	0.476	0.630	0.752

and systematically vary individual or multiple aspects of the overall procedure in order to maximize dependability of measurement. Decision studies are analogous to the Spearman–Brown prophecy formula in classical test theory (Brennan, 2001; Hintze & Matthews, 2004), except that they allow for multiple sources of prediction rather than only one. In the present study, the effects of increasing the number of sessions, coders, and items are investigated in a series of hypothetical D-study designs with the goal of not only maximizing dependability but also reaching adequate reliability for the measure to be used for its intended purpose in a psychotherapy process study.

Transference interpretations. Although session and coder variability accounted for small proportions of total variation in scores, a number of higher order interactions, including session and coder facets, did significantly contribute to variation in scores. For this reason, these facets were systematically varied to determine whether adequate dependability could be reached with an alternate coding procedure. Table II shows the generalizability coefficients obtained when number of sessions and coders per sessions are systematically varied while maintaining the same number of items on the scale. At the top left corner of the table, the dependability of transference interpretation when assessed by one coder on one session is presented. Reading across the upper row, the effect of increasing the number of sessions assessed while maintaining the use of one coder can be seen. Here, we notice substantial gains; however, dependability never reaches an adequate level, even when 12 sessions are assessed. Table II also shows the effect of increasing the number of coders while only coding one session. There is little increase in dependability even when the number of coders is sextupled. If generalizability coefficients of .7 are considered the accepted minimum, then the table allows for the determination of coding procedures that would meet or exceed this threshold. Because increasing sessions is more effective (for the same cost) than increasing coders, increasing dependability of coding procedures here requires assessment of more sessions (12 rather than six). With 12 sessions being coded, marginal gains for adding coders are evident.

Maintenance of the treatment frame. Because item variability accounted for such a large proportion of total variability, scale revision is a likely next step to be considered if increased dependability is desired. One way to deal with high item variability is to increase the number of items on the scale. This indirectly decreases the impact of the variability by increasing the N in the denominator of effects, including the item facet. Similar to the transference interpretation scale, a number of higher order interactions, including session and coder facets, also contributed significantly to variation in score. For this reason, three facets were systematically varied to determine whether adequate dependability could be reached. Table III shows the generalizability coefficients obtained when number of sessions, coders, and items are systematically varied. At the top left corner of the table, the dependability of maintenance of treatment frame if one session was rated by one coder using the current number of items on the scale (six) can be seen. Table III also shows the effect of increasing the number of sessions, while maintaining the use of one coder and the six-item scale. Here, we note substantial gains for each increase in sessions, with adequate dependability being reached between six and 12 sessions (likely much closer to six given the estimates). In addition, the effect of increasing the number of coders, while only considering one session with the six-item scale, is evident. There is little increase in dependability even when the number of coders is sextupled. Last, Table III demonstrates the impact of maintaining any combination of sessions and coders while doubling the number of

Table III. Potential Decision Study Designs for Maintenance of Treatment Frame

	1 session		3 sessions		6 sessions		12 sessions	
No. coders	6 items	12 items	6 items	12 items	6 items	12 items	6 items	12 items
1 coder	0.350	0.407	0.582	0.651	0.698	0.767	0.776	0.841
2 coders	0.392	0.442	0.626	0.684	0.736	0.793	0.807	0.862
3 coders	0.408	0.455	0.642	0.696	0.749	0.803	0.818	0.867
6 coders	0.426	0.469	0.659	0.708	0.763	0.812	0.829	0.876

items on the scale. For example, with three sessions and six coders, on the six-item scale the dependability is 0.659. In addition, the dependability when a 12-item revised scale is used is 0.708. At all levels of sessions and coders, improvements in dependability can be seen when the number of items on the scale is increased. Last, because one of the primary aims of investigating hypothetical D-study designs is to determine a maximal coding procedure, the table is inspected for values above .7 (which was set in this example as the desired threshold). When six or 12 sessions are rated, dependability is consistently above .7; even higher dependability estimates result when two or more coders are used.

To determine an optimal coding procedure, the relative cost (in terms of time, human resources, and money) of various combinations should be considered. For illustrative purposes, let us consider three potential D-study designs. First, six sessions are rated by two coders with 12 items. Second, six sessions are rated by three coders with 12 items. Last, 12 sessions are rated by two coders with six items. Table III indicates that these three designs are nearly equivalent in their dependability. A direct comparison of the first and second design reveals that a 50% increase in cost would result from choosing the second design because of the increase in coders; therefore, the first design would, under most circumstances, be preferable. Next, the first and third designs may be compared. The first design uses twice as many items but half as many sessions as the third. In most situations, it is faster, easier, and less costly to have an individual rate additional items than to have another individual rate the same smaller set of items. In particular, with observational coding systems in which the majority of the time needed for coding is spent reviewing the videotaped (or audiotaped) material rather than scoring the individual items, an increase in items can be a cost-effective way to increase dependability without drastically increasing the cost of a project. Design 1 is preferable in this case because it requires only about half the resources of Design 3; specifically, by increasing items before sessions, substantial savings in time and money result while maintaining an equivalent degree of dependability. Broadly, it should be noted that by comparing designs with similar (or different but acceptable) degrees of dependability, decision studies can be used to inform modifications to coding procedures in future studies. Importantly, when a pilot study has been conducted and relevant facets have been investigated, major savings can be obtained by using procedures that maximize dependability and minimize collection of unnecessary data through the strategic and empirically determined selection of coding parameters and levels.

Example Summary

G-theory techniques were applied to illustrate how dependability of measurement can be derived from ratings of a psychotherapy process measure. This illustration was conducted in three major steps. First, the magnitude of variance components associated with each facet of measurement was estimated and evaluated in a series of G-studies. Second, dependability of each scale was calculated based on the actual coding procedure used in the existent data set. Last, a series of potential D-study designs were presented to assess the impact of modifications to the existent measurement procedure and to maximize dependability of measurement. Given the data available, a five-facet design was selected that included persons, sessions, coders, scales, and items as the facets of generalization. Each scale was investigated separately because generalization across constructs was not desired. A relative decision rule was applied in line with the intended applications of the measure and associated relevant sources of error.

Results of the variance components decomposition revealed a number of facets and interactions accounting for large proportions of total variance. In particular, item variance and residual variance were large for both scales. Large item variability indicates the presence of substantial variation in scores on items within scales. Coder variance was low for both scales, indicating high levels of interrater reliability. Variability across sessions was also quite small, suggesting a relatively stable use of these techniques across the year of treatment. A number of interactions accounted for considerable variation in scores.

An initial potential D-study design, in which dependability was estimated based on the coding procedure used to collect the data used in the G-studies, was first considered. The generalizability coefficient for transference interpretation was .591 and for maintenance of the treatment frame, .736. Because of the low to moderate dependability of the scales using the existent coding scheme, modifications were next considered. Subsequent potential designs investigated the impact of increasing sessions, coders, and items on dependability. For both scales, the effect of increasing sessions rated was greater than the effect of increasing the number of coders per session and yielded substantial increases in dependability. When considering scale revision through the addition of new items, this approach was found to be a cost-effective way of increasing dependability for maintenance of treatment frame.

Discussion

This article discusses G-theory as a framework for estimating the reliability and dependability of psychotherapy process measures in an observational coding context. G-theory provides a framework within which multiple sources of error in a given set of measurements can be simultaneously estimated. As such, G theory extends classical test theory in a similar way to how factorial ANOVA extends one-way ANOVA (Cranford et al., 2006). By measuring reliability and error within a context of a multifaceted coding situation, one can determine sources of error that can be accounted for by various aspects of the assessment procedure. In this way, G-theory provides the information necessary for determining how many occasions, coders, questionnaire forms, or questions are needed to obtain dependable scores. G-theory has a number of additional features that, although not unique and in the context of multifaceted assessment designs, are strengths for use in psychotherapy research studies; the distinction between fixed versus random effects and relative versus absolute decision rules are both major strengths. The specification of fixed effects yields generalizations to only the investigated variable parameters, whereas the specification of random effects yields data relevant to a wider range of values of the same variable. Where the goals of a study are prospective and study parameters are likely to change, the selection of random-effects models provides the benefit of results that are not bound to the existing data set but that can inform future decisions about study parameters as personnel and/or study cites change. The selection of a relative versus absolute decision rule similarly allows for control and specificity in determinations of dependability. The application of an absolute decision rule yields estimates of dependability that take into account sources of error, which affect the absolute standing of each individual in a study. Where inclusion, exclusion, or termination may be dependent on specific cutoff scores, an absolute decision rule yields appropriate estimates. In contrast, the application of a relative decision rule is appropriate when the relative standing of individuals is the index to be used in subsequent analytic steps.

The application of G-theory to psychotherapy process research provides valuable information in the domains of measurement development, application, and training. With respect to measurement development, variance component decomposition (in the G-study) and the generalizability coefficient obtained for the initial (observed) decision study directly address the question of whether the existing measure and coding scheme were able to adequately assess the construct of interest. When low generalizability coefficients obtain, the conclusion that individuals may not be sufficiently differentiable from one another in the given sample on the given construct may be warranted. In this case, one would need to be careful when using data from this study in other applications. In particular, with low to moderate confidence in the ability of a procedure to detect differences across individuals, when correlating scale scores with psychotherapy outcomes, a low correlation could indicate no relationship (the normal interpretation) between the construct and the outcome of interest, or it may simply reflect inadequate ability to differentiate people from one another and, thus, inadequate ability to assess how their rank ordering on one measure (a technique) relates to their rank ordering on another measure (the outcome).

Beyond the estimate of dependability itself (as indexed by the generalizability coefficient), inspection of the magnitude of various sources illuminates areas in need of modification. Large item variance estimates provide evidence for low internal consistency of the measure and empirical evidence of the need for measurement revision as a means of reducing error. Here, decision studies can be used to determine whether increasing items will sufficiently improve dependability. In cases where item addition is not possible or decision studies suggest it will not significantly improve dependability, other approaches to measurement revision may be relevant. Variance component decomposition also provides information critical to procedural modifications and the allocation of resources.

Inspection of variance component estimates also provides information relevant to training and coder drift. When a large Patient $\times$ Coder interaction occurs, the possibility of coder reactions to particular patients interfering with objective assessment is present and warrants training and/or a support structure to minimize this effect. When a substantial Coder $\times$ Item interaction is present, increased attention to standardization of coding procedures and detailed item descriptions may be necessary. Similarly, the higher order interaction of these three effects may also draw attention to the need for greater standardization of coding procedures or more in-depth or structured ongoing supervisory coding meetings in order to minimize the potential of coders' differential use of items and scales across persons. Last, when interactions including coder and sessions are large, this may be an indication of coder drift or fatigue. To investigate this hypothesis directly, session (time) may be indexed as the order in which coders rated each session (for a given patient) rather than as a chronological index of the

patient's time in treatment. A large interaction in this case indicates that for at least one coder across sessions scores have begun to deviate. At this point, direct inspection of raw scores will indicate the individual who has begun to drift and subsequent retraining or supervision can be delivered.

Limitations

Despite the numerous benefits of applying G-theory to psychotherapy research, some limitations should be noted. G-theory makes a number of assumptions, including random sampling, normal distributions, and large measurement samples, that often do not hold in practice. The greater the violations of these assumptions, particularly with respect to unbalanced data, the more biased estimates of variance components become. In such cases, single estimates may be unstable and computation of standard errors and/or confidence intervals may be necessary. Several alternatives for estimating variance components have been articulated by Searle (1987), including maximum likelihood, restricted maximum likelihood, MINQUE, and minimum variance quadratic unbiased estimation. Bootstrapping and jackknife procedures (Brennan, 2001) can also be used to estimate variance components and standard errors.

An additional limitation in the use of G-theory has to do with its application to data in which change occurs as a function of time. A primary assumption of G-theory is that all variance in scores is due to error and not to change. For this reason, if there is systematic change over time and it does not represent error, this must be taken into account before seeking to determine the dependability of an assessment procedure. Within the scope of the techniques discussed, one approach is to treat time as a fixed effect. If it is expected that there will be systematic variability in the construct of interest over time, then one would not be willing to substitute the existent sample of occasions for a random same-size sample of occasions, hence the use of a fixed effect. To the extent that change over time can be predicted a priori, time can be separated into discrete epochs of time in which change is not expected to occur. In this case, epochs would be treated as fixed effects and investigated separately, whereas sessions within a given epoch would be treated as a random effect and the assumption of a static rather than dynamic process would apply. However, if a researcher's primary goal is to model the change in a process rather than determine the parameters necessary to achieve a stable estimate of the process, then G-theory may not be applicable.

Conclusions

In summary, the techniques of G-theory can also be used in a wide range of contexts. It can be used to determine the power to detect differences between individuals in psychotherapy process and/or outcome studies. This application may be prospective or retrospective. In cases where results have already been published, G-theory can provide information relevant to determining the ceiling of the observable effect as a result of the error associated with a given coding procedure (Hoyt & Melby, 1999). Prospectively, researchers can determine whether a planned procedure will have adequate power to detect the effect of interest. As previously discussed, G-theory can also be used to provide information about particular sources of error in an existent coding procedure. That is, inspection of variance component estimates provides direction for allocation of resources to procedural or training modifications necessary for increased dependability of measurement. These techniques would be fruitfully applied to a wide range of measures in their development stages. Both technique-specific and common-factors measures would benefit from attention to determination of dependability of assessment. Measures aimed at assessment of common factors may also use G-theory to determine whether adequate variability in these factors (e.g., alliance, empathy) is present to detect between-person differences and, therefore, can be predictive of outcome. Finally, for both measures of process and outcome, decision studies can be used to determine the number of assessments necessary to obtain a stable estimate of the targeted construct.

Notes

[1] Throughout this article, "patient" is used to denote the primary target of assessment. There are situations in which therapist or dyad (patient and therapist together) would be the more relevant target of measurement. "Patient" is here used to designate the techniques used in the therapy session of a particular patient, despite the fact that the therapist is clearly involved in this process. It would be equally feasible to call the target a dyad effect. The effect, regardless of labeling, represents interindividual differences in the construct of interest.

[2] Process data from two patients were not obtained because of missing videotapes from the psychotherapy sessions. Thus, videotaped psychotherapy sessions from 15 patients were coded with the PPRS-BPD.

[3] Two patients had missing sessions; one had available data for five sessions out of six and the other had only four sessions available.

[4] An appropriate ICC is selected based on the study design and goals. ICC(1,1) may be used when targets are the only source of variability of interest. ICC(2,1) and (2,k) may be used when a two-way random-effects model is being considered (i.e., when both targets and coders are random). ICC(3,1) and (3,k) may be used when a two-way mixed-effects model is being

considered (i.e., when targets are random but coders are fixed). The use of ICC(2,k) and (3,k) provide an estimate of reliability when data averaged across coders is to be used, whereas ICC(2,1) and (3,1) provide as estimate or reliability when one coder data will not be averaged (McGraw & Wong, 1996).

[5] It is important to note that, because coders are not fully crossed with sessions, the absolute contribution to error cannot be determined from this design. [0]

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.

Appelbaum, A. J., & Levy, K. N. (2002). Supportive psychotherapy for borderline patients: A psychoanalytic research perspective. *American Journal of Psychoanalysis*, *62*(2), 201–202.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Clarkin, J. F., Foelsch, P. A., Levy, K. N., Hull, J. W., Delaney, J. C., & Kernberg, O. F. (2001). The development of a psychodynamic treatment for patients with borderline personality disorders: A preliminary study of behavioral change. *Journal of Personality Disorders*, *15*, 487–495.

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, *8*, 411–426.

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, *32*, 917–929.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral assessment measurements*. New York: Wiley.

Halvorsen, M. S., Hagtvet, K. A., & Monsen, J. T. (2006). The reliability of self-image change scores in psychotherapy research: An application of generalizability theory. Psychotherapy: Theory, Research, Practice. *Training*, *43*, 308–321.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, *33*(2), 258–270.

Hoyt, W. T. (2002). Bias in participant ratings of psychotherapy process: An initial generalizability study. *Journal of Counseling Psychology*, *49*(1), 35–46.

Hoyt, W. T., & Melby, J. N. (1999). Dependability of measurement in counseling psychology: An introduction to generalizability theory. *The Counseling Psychologist*, *27*(3), 325–352.

Levy, K. N., Clarkin, J. F., Yeomans, F. E., Scott, L. N., Wasserman, R. H., & Kernberg, O. F. (2006). The mechanisms of change in the treatment of borderline personality disorder with transference focused psychotherapy. *Journal of Clinical Psychology*, *62*, 481–501.

Levy, K. N., Wasserman, R. H., Clarkin, J. F., Eubanks-Carter, C., & Fisher, C. (2005). *The Psychotherapy Process Rating Scale for Borderline Personality Disorder (PPRS-BPD)*. Unpublished rating scale, The Pennsylvania State University.

Linehan, M. M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. New York: Guilford Press.

Lynch, T. R., Chapman, A. L., Rosenthal, M. Z., Kuo, J. R., & Linehann, M. M. (2006). Mechanisms of change in dialectical behavior therapy: Theoretical and empirical observations. *Journal of Clinical Psychology*, *62*(4), 459–480.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Mellenbergh, G. J. (2001). Outline of a faceted theory of item response data. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders (Eds), *Essays on item response theory* (pp. 415–432). New York: Springer.

O'Brian, N., O'Brian, S., Packman, A., & Onslow, M. (2003). Generalizability theory: I. Assessing reliability of observational data in the communication sciences. Journal of Speech, Language. *and Hearing Research*, *46*, 711–717.

Rao, C. R. (1971). Estimation of variance and covariance components-MINQUE theory. *Journal of Multivariate Analysis*, *1*, 257–275.

Rockland, L. H. (1992). *Supportive therapy for borderline patients*. New York: Guilford Press.

Searle, S. R. (1987). *Linear models for unbalanced data*. New York: Wiley.

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

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922–932.

SPSS Inc. (2006). *SPSS 15.0 user's guide*. Englewood Cliffs, NJ: Prentice Hall.

Tinsley, H. E., & Weiss, D. J. (1975). Inter-rater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*, 358–376.

Verhelst, N. D., & Verstralen, H. H. F. M. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders (Eds), *Essays on item response theory* (pp. 89–108). New York: Springer.