Published in final edited form as:

Infant Child Dev. 2008 June 1; 17(3): 269–284. doi:10.1002/icd.551.

What Sources Contribute to Variance in Observer Ratings? Using Generalizability Theory to Assess Construct Validity of Psychological Measures

Kimberley D. Lakes^{a,*} and William T. Hoyt^b

^a Child Development Center, Department of Pediatrics, University of California, Irvine, CA, USA

^b Department of Counseling Psychology, University of Wisconsin—Madison, WI, USA

Abstract

We illustrate the utility of generalizability theory (GT) as a conceptual framework that encourages psychological researchers to address this question and as a flexible set of analytic tools that can provide answers to inform both substantive theory and measurement practice. To illustrate these capabilities, we analyze observer ratings of 27 caregiver–child dyads, focusing on the importance of situational (contextual) factors as sources of variance in observer ratings of caregiver–child behaviors. Cross-situational consistency was relatively low for the categories of behavior analyzed, indicating that dyads vary greatly in their interactional patterns from one situation to the next, so that it is difficult to predict behavioral frequencies in one context from behaviors observed in a different context. Our findings suggest that single-situation behavioral measures may have limited generalizability, either to behavior in other contexts or as measures of global interaction tendencies. We discuss the implications of these findings for research and measurement design in developmental psychology.

Keywords

parent–child relationships; measurement; observer ratings; caregiver–child relationships; construct validity

INTRODUCTION

Perhaps the most complex judgments social scientists make concerning research design and interpretation relate to the validity of measurement procedures. Constructs of interest to developmental psychologists, for example, are usually not directly measurable and must be inferred from self-reports or other observable behaviors. The crucial insight that drives the search for both reliability and validity evidence is that only part of the variance in scores on these (indirect) measures is attributable to the construct they are intended to quantify. Other sources of variance include random error and systematic variance attributable to specific

facets of the measurement method (e.g. choice of items or timing of measurement, Schmidt, Le, & Ilies, 2003). Thus, the essential question for investigators seeking to establish the validity of a given measurement procedure for a particular use (more formally, the validity of their intended interpretations of scores on a measure; Messick, 1989) is ‘What constructs account for variance in test performance?’ (Cronbach & Meehl, 1955, p. 282).

Generalizability theory (GT; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is often presented as an extension of classical reliability theory (e.g. Brennan, 2001; Cronbach *et al.*, 1972; Shavelson & Webb, 1991). Indeed, the most common application of generalizability analyses is the examination of the reliability (dependability) of measurement when multiple sources (facets) contribute to error of measurement. However, as Cronbach *et al.* (1972) also foresaw, GT, which is based on the partitioning of observed score variance into components derived from multiple sources and their interactions, provides a powerful tool for addressing the broader issue of construct validity.

Our goal in this article is to provide a very brief introduction to GT, and to illustrate its utility for evaluating the construct validity of rating systems, by presenting analyses of a specific rating approach used in previous studies to assess characteristics of caregiver–child interaction. Other primers on conducting and interpreting generalizability analyses are available (Hoyt & Melby, 1999; Lakes & Hoyt, in press; Shavelson & Webb, 1991), and space constraints make a step-by-step tutorial out of the question. Instead, we focus on the variance partitioning process, which is the initial step of a generalizability (*G*) study, and show how interpretation of the resulting variance components sheds light on both methodological and substantive questions relevant to understanding caregiver–child relationships. Previous papers (e.g., Lakes & Hoyt, in press) have focused on the conventional application of GT as an extension of classical reliability theory; the purpose of the present article is to show its applications to substantive (i.e., validity) inquiries.

GENERALIZABILITY OF OBSERVER RATINGS

Lakes and Hoyt (in press) argue that observer ratings (when adequate numbers of observers are used) are more dependable than parent or teacher ratings, which are inherently limited in number of observers, and, therefore, corrupted by rater bias. However, they note that a strength of acquaintance (e.g. parent or teacher) ratings is that they are inherently based on a large behavior sample (i.e. the person completing a questionnaire has spent a great deal of time observing the child). In contrast, observer ratings are inherently based on a relatively small behavior sample, usually gathered through laboratory observations or classroom/field observations. Thus, in weighing the trade-off between these two measurement methods, it is crucial to have studies examining the generalizability of observer ratings over situations or contexts.

Lakes and Hoyt (in press) noted that raters can produce highly dependable scores when they observe the same brief behavioral episode, but this does not necessarily mean that these ratings will generalize to other observational occasions or to other situations. For some types of behavior (e.g. laughing), people may be relatively inconsistent in their behavior from one interaction to the next, so that observer ratings of brief behavioral episodes do not provide a

dependable index of a general behavioral tendency. Even more important, situational constraints may make it unlikely that a score on some characteristic (e.g. cooperativeness) will generalize from one observational situation (e.g. classroom) to another (e.g. playground).

Thus, observer ratings are more dependable than parent or teacher ratings (Hoyt, 2000; Lakes & Hoyt, in press) in part because they standardize (or *fix*) both the time period (occasion) and the situation in which the observations occur. Because all raters observe the same sample of behavior, they are more likely to agree than, say, a math teacher and a homeroom teacher who observe the target child on different occasions and in different behavioral contexts. If the purpose of the measurement is to predict future behavior in a variety of contexts, it may be that the teacher ratings, although nominally of poorer quality, will be the better predictors. To put this differently, the more reliable ratings will not necessarily be more valid, given that reliability has been achieved by ‘fixing’ measurement facets (occasions and situations) over which we wish to generalize. Temporal stability (generalizability over occasions) is a property of scores that is frequently considered in both reliability and generalizability analyses. In this article, we show how GT can help investigators to address the issue of generalizability over situations (in this case, three differently structured, 5-min interactions) and discuss the substantive and psychometric implications of our findings.

METHOD

Participants

Dyads—Dyads in this study were parent–child dyads enrolled in the Community-University Initiative for the Development of Attention and Readiness (CUIDAR), San Bernardino, an early intervention project. Children in the dyads ranged in age from 3 to 5, and the sample was predominantly Latino.

Raters—Raters were psychology and child development research assistants who were trained using the Dyadic Parent–Child Interaction Coding System (DPICS) manual and practiced coding videotapes in group and individual training sessions. Each rater completed approximately 20 h of training in the DPICS.

Measure

The DPICS (Eyberg & Robinson, 1981) is an observational measure of parent–child interactions. The structure consists of 5 min of child-directed interaction (CDI), 5 min of parent-directed interaction (PDI), and 5 min of clean up (CU). These three situations provide coders with varied interactional contexts in which to observe degrees of parent control and child compliance to parental commands. Behaviors are tallied continuously throughout each 5-min section. On the DPICS, we analyzed 17 behavioral categories that are described in Table 1.

Procedure

On the DPICS, data consist of frequencies for each behavioral category in each of the three interactional contexts (situations). Frequencies not only reflect the relative importance of each behavior, but also the absolute number of behaviors observed, which may differ between dyads (also between raters). Accordingly, we converted these frequencies to proportions. These proportions were then subjected to a square root transformation, which reduced the expected positive skew (although there was still significant skewness for some categories). The DPICS rating system used in this study codes 45 categories and subcategories of behavior, but frequencies (and proportions) were near zero for most or all dyads on some of these categories. Because we were interested in categories that varied across dyads, we analyzed the 17 behavioral categories for which the maximum proportion was 0.20 or higher. This means that for each of these categories, at least one rater indicated that one dyad, on one of the three situations, devoted 20% of its categorizable interactions to this behavior.

Our sample analyses use data for parent–child dyads (or ‘dyads’; $n_d = 27$) rated by a group of trained observers (or ‘raters’; $n_r = 5$) on multiple behavior indices designed to measure behaviors in parent and child interactions in various behavioral situations (or ‘situations’; $n_s = 3$). Every dyad was rated by all 5 raters on 3 situations (CDI, PDI, and CU)—a fully crossed design denoted as dyads $\times$ raters $\times$ situations or *DRS*.

Analyses

Generalizability studies (or *G studies*) begin with a variance partitioning procedure similar to analysis of variance (ANOVA). Table 2 lists the components whose contributions to observed scores can be estimated in the *DRS* design. As in a factorial ANOVA, the score assigned to dyad d by rater r on situation s is conceptualized as a deviation from the grand mean (over all dyads, raters, and situations), with the degree of deviation determined by d 's dyad effect (called *universe score* in GT; similar to the true score of classical test theory), and two sources of error: r 's rater effect, and s 's situation effect. In addition to these main effects, each facet (R or S) interacts with the object of measurement (D) in a two-way interaction (DR and DS , respectively), as well as with the other facet (RS). Finally, there is also a three-way interaction, which is confounded with error (DRS , e). In most applications of the DPICS, the variance in dyad effects constitutes valid variance in scores, whereas the remaining main effects and interactions represent sources of error variance. The right column in Table 2 describes the interpretation of each effect as it contributes to a given rating. We used the program GENOVA (Crick & Brennan, 1982) to compute variance estimates.

RESULTS

Variance estimates for each of the seven components for 17 DPICS categories are shown in Table 3. Our interpretations of these results will focus on the question of whether a broader acquaintance (i.e. additional behavior samples from varied situations) with dyads enhances rating validity. If observer ratings *generalize* well over situations, this means that behaviors are relatively consistent over situations, and additional opportunities to observe behavior in

diverse situations will not make much difference; if they do not, then it is important to have a larger behavior sample across multiple situations, which strengthens the case for acquaintance ratings. To address this question, we focus primarily on variance attributable to *D* and *DS* components, and secondarily on *S* variance.

Dyadic (D) Variance

One hypothesized source of variance in DPICS ratings is between-dyad differences in behavioral tendencies that are consistent across raters (i.e. raters agree which dyads are high or low on these behaviors) and situations (i.e. dyads high in a given behavior in one situation also tend to be high in the other situations). This behavioral consistency is represented in Table 3 as dyadic variance or *D* variance, and accounted for between 1% (*Physical Positive*) and 26% (*Positive Affect*) of the total variance.

Positive Affect as coded on the DPICS represents a verbal expression from the child that expresses pleasure in something they have done, an activity, an object, or the caregiver. *D* variance was particularly large for *Positive Affect* (26%), which means that some children display consistently more of this quality (as perceived by different raters and across different situations) and others less. When there is substantial dyadic (or dyad-level) variance, we should suspect that this is related to enduring characteristics (and may predict other dyadic behaviors or outcomes).

D variance was low for *Physical Positive* (1%), indicating that for this behavior (positive or neutral physical contact from the caregiver to the child), few consistent between-dyad differences are observed. When *D* variance is low, the logical conclusion is that this behavior does not vary systematically among dyads. At any given time, some dyads will be higher and others lower on positive physical contact, but from time to time (or situation to situation) dyads do not tend to be consistent in terms of their frequency of using this behavior. (An alternative hypothesis is that there is stable dyadic variance on this category, but the rating system fails to operationalize the representative behaviors in such a way that it can be dependably scored. If theory dictates that this construct is pivotal, this can motivate researchers to take another look at the design of the rating system or at rater training.)

Dyad × Situation (DS) Variance

As in the case of ANOVA, in GT, main effects should be interpreted in the context of relevant interactions (e.g. *DS*). *DS* variance estimates the variance due to inconsistencies from one situation to another for a dyad, averaging over raters. *DS* variance accounted for between 10% (Indirect Command/ Noncompliance, Destructive Behavior) and 35% (*Physical Positive*) of total variance. This indicates that ordering of dyads differs somewhat in different situations. There are a number of behaviors where *DS* variance represents more than 25% of observed score variance. This could reflect that different dyads have different styles of engaging with specific tasks.

In some categories (e.g. *Physical Positive*) the *D* main effect is small (1%), but the *DS* interaction is large (35%). This suggests that although differences between dyads are not consistent across situations (small *D* variance) there are substantial between-dyad

differences *within* situations (large *DS* variance). In other words, *Physical Positive* shows strong between-dyad variance *within* situations, but a dyad that shows a high proportion of physical positive behaviors during CDI may not display high physical positive behaviors during PDI. Thus, we might ask ourselves whether one of these situations is more important than another, for the purpose of understanding positive behaviors in parent–child interactions. For example, we may want to examine physical positive behaviors during PDI, to assess whether a parent may engage in these behaviors to facilitate the child's involvement in the parent's selected activity (e.g. the parent gently places a hand on the child's shoulder to encourage the child to become involved with the parent).

Situations versus occasions—An important caveat for interpreting *DS* variance is that we do not know whether these between-dyad differences result from the nature of the situation (i.e. some dyads respond differently to child-directed interaction than others) or simply reflect the inconsistency of some behaviors over time (occasion variance). For example, if a dyad engages in many ‘physical positive’ behaviors in one situation, they may tend to have fewer such behaviors in the next situation, regardless of the structure of the situation, due to fatigue effects. Thus, *S* variance in this study may arise not because of systematically different responses to situations, but because of the temporal instability of the behavior under study. To unconfound situation-specific variance from occasion variance, it would be necessary to observe dyads on multiple occasions within each situation. Technically, this is a *DR(O:S)* design, with occasions nested within situations. (For a tutorial on conducting a three-facet *G* study, we refer readers to Lakes & Hoyt, in press.)

Cross-situational consistency—To better understand the generalizability of ratings over situations for each category, we can compare the proportions of *D* and *DS* variance. The ratio of $\text{var}(D)/[\text{var}(D) + \text{var}(DS)]$ is a useful index of cross-situational consistency for the categories considered on the DPICS. This ratio could be called the asymptotic *G* coefficient for single-situation ratings, in that it is the theoretical *G* coefficient for a study in which all dyads are observed in the same (single) situation, with an infinite number of raters (so that components other than *D* and *DS* contribute zero to error variance). It estimates the correlation between two sets of ratings taken in different situations, observed by the same infinite set of raters. The cross-situational consistency ratio therefore provides a ceiling on how strongly ratings from a finite set of observers in one situation will correlate with ratings from the same observers based on observations in a different situation.

The cross-situational consistency ratio for each DPICS category is shown in the rightmost column of Table 3. The findings suggest that, for most or all of these categories, generalizability across situations is very limited. To have generalizable ratings of most DPICS categories from trained observers, one would need to (a) observe the dyad in multiple situations and aggregate across situations, or (b) observe the dyad for more or longer time intervals. The goal is to reduce the extent to which *DS* variance contributes to observed score variance. (For an explanation of the principle of aggregation to improve generalizability, see Lakes & Hoyt, in press.) Whether more time by itself will be helpful (in the absence of increasing the number of situations) depends on whether *DS* variance in this study is due more to situations or more to simple temporal instability (occasion variance).

Situation (S) Variance

It may also be important to understand situational influences on behavior that are consistent across dyads. *S* estimates between-situation variance in ratings, averaged over dyads and raters. When *S* variance is large, this means that the behavior is rated consistently higher in some situations than others. Psychometrically, even a very high proportion of *S* variance is not a problem in most research applications unless some dyads are observed in one situation and others in a different situation. In this case, *S* variance will contribute to error of measurement. In other words, *S* variance can be ignored (from the standpoint of generalizability) as long as the researcher has control of the situation and all dyads are observed under similar situational constraints. In non-lab contexts, the researcher may have less control and will have to observe different dyads under somewhat different circumstances. In this case, situation variance contributes to error in the *G* coefficient, reducing generalizability of ratings.

Kenny, Mohr, and Levesque (2001) reviewed studies examining interaction partner as one important species of social situation and found little evidence of *S* variance on a variety of behavioral ratings. The DPICS varies situations by altering the instructional set, and this produces greater evidence of general situational response sets. As shown in Table 3, *S* variance was high for *Questions* (60%), *Indirect Commands* (34%), and *Reflective Statements* (31%). Caregivers asked the most questions in the situations where they were told to exert the least control (CDI). They used the most *indirect commands* during CU, indicating that as the need for parental control increases, these parents used more commands. Parents used the most reflective statements during CDI and virtually no reflective statements during CU. These situational differences are important for planning observational studies, and this issue will be discussed further in the following section.

DISCUSSION

According to Cronbach and Meehl (1955, p. 282), the key question in construct validity inquiry is ‘What sources contribute to variance in test performance?’ GT provides a conceptual framework that encourages researchers to address this question, and a flexible set of analytic tools that can provide nuanced answers to inform both substantive theory and measurement practice. To illustrate these capabilities, we analyzed DPICS data for 27 caregiver–child dyads to examine the importance of situational (contextual) factors as sources of variance in observer ratings of parent–child behaviors. Cross-situational consistency was relatively low for these 17 categories of behavior, indicating that dyads vary greatly in their interactional patterns from one situation to the next, so that it is difficult to predict behavioral frequencies in one context (say, PDI) from behaviors observed in a different context (say, CDI). In the terminology of GT, single-situation measures (whether expressed as frequencies or proportions) can be expected to have very limited generalizability, either to behavior in other contexts or as measures of global interaction tendencies.¹

As we consider how these findings may generalize to other observational measures, several features of the DPICS are noteworthy. First, observers are restricted to relatively brief (5 min) behavior samples, which raises the possibility that occasion-specific variance is large

for these ratings, particularly for low-frequency behavioral categories. As discussed above, it is likely that some of what we categorize as *DS* variance in Table 3 is actually *DO* variance (i.e. is attributable to instability of behaviors over time, rather than to differences in family interaction patterns in response to particular situational constraints).

Second, the interaction context is varied with three instructional sets (CDI, PDI, and CU) that may be expected to evoke different behavior patterns. (The substantial proportion of *S* variance for several of these categories suggests that this was indeed the case for dyads in our sample.) Thus, some inconsistency of behavior across situations is to be expected. Note, however, that the index of cross-situational consistency in Table 3 ignores *S* variance (which does not contribute to observed score variance when all dyads are observed in the same set of situations) and still indicates limited consistency in behavior for most DPICS categories.

Finally, the DPICS might be described as a low-inference rating scale. Coders classify each behavior into mutually exclusive categories that are presumed to be directly observable. This approach may be contrasted with rating scales requiring greater inference on the part of the coders. For example, coders could give each dyad a global rating on characteristics such as cooperativeness or positive affect, based on their overall impressions from a 5-min behavior sample. Low inference rating systems have been found to show lower proportions of variance attributable to rater biases (Hoyt & Kerns, 1999), relative to rating systems requiring greater inference. However, the proportion of bias variance can be reduced by aggregating over a suitable number of raters (Lakes & Hoyt, in press), and it is possible that global (inferential) rating systems are less sensitive to temporal and situational variations in particular behaviors, and so would show greater cross-situational consistency than the DPICS.²

Implications for Validity of Observer Ratings

Our findings are preliminary and are based on a single rating scale with three types of situations and, therefore, may not generalize to all observational rating scales. However, we have illustrated how GT can be used to study the validity of observer ratings and future research should apply GT to measurement studies with other observational measures used in developmental psychology. The low cross-situational consistency in our sample study suggests that observers may need to observe dyads in multiple contexts in order for ratings to be valid measures of the construct of interest; therefore, researchers conducting measurement studies should include multiple contexts in their research design.

Generalizability will be enhanced by aggregating scores across these multiple contexts. In

¹Although our focus here was on the heuristic value of the variance components analysis, a typical *G* study would use the variance estimates in Table 3 to estimate the generalizability of composite measures that aggregate over varying numbers of raters and situations. Such aggregate measures (averaging over multiple situations, e.g.) are typically much more reliable than single-situation ratings. In point of fact, however, generalizability of proportion scores averaging over 5 raters and 3 situations is still quite low, with *G* coefficients of 0.7 or higher for only three categories (noncompliance to indirect communication, positive effect, and destructive behavior). Most other categories had *G* coefficients below 0.5.

²There is some evidence that this may be the case for global personality ratings of adults based on brief (median observation time = 3 min) behavior samples. Borkenau, Mauer, Riemann, Spinath, and Angleitner (2004) describe ratings from the German Observational Study of Adult Twins, in which participants were videotaped in 15 diverse situations (e.g. solving a logic puzzle, telling a joke, introducing a new acquaintance to the experimenter). Trained judges used these videotapes as the basis for global ratings of the Big 5 personality domains, which showed cross-situational consistencies ranging from 0.35 to 0.51—somewhat higher than those we found for the DPICS (Table 3).

addition, further study of the relative strengths and weaknesses of low-inference and high-inference rating systems is needed.

Comparing observer and acquaintance ratings—Observer and acquaintance ratings have distinct advantages and disadvantages, and researchers designing a study should consider these carefully as they determine how to best design or select a measure for their particular study. As noted in Lakes and Hoyt (in press), observer ratings tend to be strong in terms of interrater reliability (due to rater training and the fact that ratings are based on identical laboratory or field observations). However, this advantage can also be a limitation, if (a) the goal is to construct a measure of general behavioral tendencies, and (b) ratings are based on a distorted or unrepresentative sample of behaviors. In the terminology of GT, this is a question of generalizability (stability) of behavior over time and across situations. To understand the relative merits of observer and acquaintance ratings for different behaviors (or traits) and for different populations, it is important to conduct G studies such as the one we report on here.

Value of GT for Psychological Research

GT is usually presented as an extension of classical reliability theory, in that it allows researchers to consider simultaneously the contribution of several sources of error to variance in observed scores. In this article, we have shown that it is also a natural framework within which to address the fundamental issue for construct validation: ‘What constructs contribute to variance in test performance’ (or, more generally, to variance in observed scores on psychological measures)? Finally, we consider the utility of GT for addressing substantive issues relevant to psychological theory.

Reliability and error of measurement—Measurement error normally attenuates effect sizes, resulting in underestimates of observed correlations between scores, and increasing the likelihood of Type II errors (Schmidt & Hunter, 1996). However, the effects of measurement error on observed effect sizes are complex, and correlated measurement error can inflate effect sizes in some research studies (Hoyt, 2000; Hoyt, Warbasse, & Chu, 2006). Thus, it is vital for researchers to understand the contributions of multiple sources of error to score variance and to consider the impact of this nuisance variance in interpreting observed effect sizes. Typically, traditional reliability coefficients underestimate error variance, because they consider only one source (Schmidt *et al.*, 2003). When important sources of error are overlooked, reliability (or generalizability) coefficients provide inaccurate indices of score dependability, and complicate interpretation of observed correlations (or other effect sizes). As we have illustrated here, this is the case with many G studies of observer ratings, which focus on observers (and often items) as a source of error, but ignore error variance due to occasions and situations. So GT provides an important tool for researchers seeking to take multiple sources of error into account and to avoid the perils of ‘hidden’ error variance concealed by traditional reliability coefficients (Lakes & Hoyt, in press).

Construct validity and the meaning of method variance—Classical reliability theory proceeds from a simple partitioning of variance into true score and error components, with the implication that error variance is simply random noise without substantive

significance. Construct validity theory, in contrast, adopts a more nuanced view. Construct-irrelevant variance is omnipresent in psychological measures and contains both systematic and random components. As suggested by Cronbach (1995), one researcher's method variance may constitute a substantive finding for another investigator. For example, researchers wishing to understand caregiver–child behavior within a particular situational context (e.g. to predict child responses to the caregiver in a free-play setting) may find that situation-specific variance on the DPICS (*S* and *DS* components in Table 3) constitute construct-relevant variance in this application. For this intended application, it makes sense to observe behavior in a single situation, and to consider *D* and *DS* components as representing variance in universe scores for the (situation-specific) construct of interest.³ Thus, generalizability analyses invite researchers to consider more carefully the components that contribute to score variance, and to determine for themselves which components constitute error in their particular applications.

Using GT to address substantive research questions—In another sense, ‘giving method variance its due’ (Cronbach, 1995, p. 145) may mean considering how method effects in one study may constitute the foundation for substantive research questions in a different investigation. The *DS* interactions in our example study provide an illustration of this approach. From the point of view of a researcher using DPICS ratings to quantify general behavioral tendencies in caregiver–child dyads, *DS* interactions represent nuisance variance that lowers the reliability (and validity) of DPICS scores. From the point of view of a psychologist interested in the person–situation debate (e.g. Ross & Nisbett, 1991), this same component represents a valuable source of data on the importance of person–situation interactions for this population, and the data on cross-situational consistency in Table 3 provide a means to quantify stability of behavioral tendencies for these dyads.

Another class of research questions of interest to developmental psychologists pertains to the level of analysis for data from studies of dyads and families (Aksan, Kochanska, & Ortmann, 2006). When a construct, such as forgiveness, occurs in the context of a close relationship, it is worth considering whether forgiving is mostly a function of the victim (actor effect or dispositional inclination to forgive) of the transgressor (partner effect or forgivability) or of the relationship between the two (dyad effect, McCullough, Hoyt, & Rachal 2000). The Social Relations Model (Kenny, 1994) represents a special case of GT developed to study actor, partner, and relationship effects in social perceptions and social behavior and can assist researchers in clarifying the level at which such dyadic constructs operate, by examining the sources of variance that most strongly contribute to observed construct variance. For example, Hoyt, Fincham, McCullough, Maio, and Davila (2005) studied forgiveness tendencies in three-person families and found that variance partitioning differed by family member (mother, father, child) and by dyad. Partner effects (evidence of forgivability) were mostly nonsignificant, indicating little consistent evidence of variance across families in how readily particular members are forgiven by others. Relationship effects were often significant for parental dyads, and tended to be strongest and most

³*S* variance is also construct relevant in this investigation. However, when all dyads are observed in the same situation, the *S* main effect does not contribute to variance in observed scores (because it is constant for all dyads).

consistent for mothers forgiving fathers, suggesting that relationship-level factors were an important determinant of wives' willingness to forgive their husbands.

Such findings can be important for understanding properties of dyadic measures. (For example, it is important to recognize that when we measure women's forgiveness toward their husbands for some offense, variance in these scores is not just attributable to their general forgivingness, but has a substantial dyadic, or relationship-specific, component.) Understanding what sources contribute to variance in scores is important for interpreting findings from any multimodal investigation (Hoyt & McCullough, 2005; Hoyt *et al.*, 2006). In addition, investigations of level of analysis can have implications for theory development (because they direct our attention to the likely level of relevant causal processes; Aksan *et al.*, 2006) and for intervention development (because the nature of these causal processes has implication for the appropriate level of intervention; Hoyt & McCullough, 2005).

CONCLUSION

In this article, we suggest that the utility of GT goes beyond its undeniable utility as an extension of classical reliability theory. We note that GT is a natural framework for inquiry into construct validity of measurement, in that it focuses on the question that is central to such investigations, namely, what constructs contribute to variance in scores on a given measure? Although GT appears to be relatively infrequently used by researchers in developmental psychology, we believe that it holds great promise for both improving researchers' psychometric sophistication and addressing complex substantive research questions.

Appendix

APPENDIX

GENOVA syntax for G study:

```
STUDY      DPICS DRS DESIGN: t1p2 STATEMENT
COMMENT
COMMENT    # RECORDS = 405 (27 dyads; 5 raters; 3 situations)
COMMENT    # VALUES PER RECORD = 1 (situation rating)
COMMENT    D = DYAD; R = RATER; S = SITUATION (1 = CDI; 2 = PDI; 3 = CU)
COMMENT
OPTIONS    RECORDS ALL
EFFECT     * D 27 0
EFFECT     + R 5 0
EFFECT     + S 3 0
FORMAT     (T32, F5.3)
PROCESS
```

```
1100  0  1  .450
1100  0  2  .338
1100  0  3  .350
```

1100	1	1	.358
1100	1	2	.318
1100	1	3	.383
1100	2	1	.386
1100	2	2	.424
1100	2	3	.387
1100	3	1	.493
1100	3	2	.494
1100	3	3	.561
1100	4	1	.497
1100	4	2	.360
1100	4	3	.440

Note: We include only the first 15 records of data (for the first of 27 dyads) for illustrative purposes. GENOVA reads only the decimal values, which are transformed proportions for the 'statement' category. (The FORMAT line uses standard FORTRAN formats: T32 = tab 32 spaces; F5.3 = 5-space numeral with three decimal places.) Preceding the data on each line are codes for dyad, rater, and situation (not read by GENOVA). The full data set contains 405 such records.

REFERENCES

- Aksan N, Kochanska G, Ortmann MR. Mutually responsive orientation between parents and their young children: Toward methodological advances in the science of relationships. *Developmental Psychology*. 2006; 42:833–848. DOI: 10.1037/0012-1649.42.5.833. [PubMed: 16953690]
- Borkenau P, Mauer N, Riemann R, Spinath FM, Angleitner A. Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*. 2004; 86:599–614. DOI: 10.1037/0022-3514.86.4.599. [PubMed: 15053708]
- Brennan, RL. Generalizability theory. Springer; New York: 2001.
- Cronbach, LJ. Giving method variance its due.. In: Shrout, PE.; Fiske, TS., editors. *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske*. Lawrence Erlbaum; Hillsdale, NJ: 1995. p. 145-157.
- Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychological Bulletin*. 1955; 52:281–302. [PubMed: 13245896]
- Cronbach, LJ.; Gleser, GC.; Nanda, AN.; Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. John Wiley; New York: 1972.
- Crick, JE.; Brennan, RL. *GENOVA: A generalized analysis of variance system* (FORTRAN IV computer program and manual).. Computer Facilities; University of Massachusetts; Dorchester: Boston: 1982. Free download: <http://www.uiowa.edu/itp/pages/SWGENOVA.SHTML>
- Dennis T. Emotional self-regulation in preschoolers: The interplay of child approach reactivity, parenting, and control capacities. *Developmental Psychology*. 2006; 42:84–97. DOI: 10.1037/0012-1649.42.1.84. [PubMed: 16420120]
- Eyberg, SM.; Robinson, EA. Dyadic parent–child interaction coding system: A manual. The Parenting Clinic Department of Family and Child Nursing, School of Nursing, University of Washington; Washington: 1981.
- Hoyt WT. Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*. 2000; 5:64–86. [PubMed: 10937323]
- Hoyt WT, Fincham F, McCullough ME, Maio G, Davila J. Responses to interpersonal transgressions in families: Forgiveness, forgivability, and relationship-specific effects. *Journal of Personality and Social Psychology*. 2005; 89:375–394. DOI: 10.1037/0022-3514.89.3.375. [PubMed: 16248720]
- Hoyt WT, Kerns MD. Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*. 1999; 4:403–424. DOI: 10.1037/1082-989X.4.4.403.

- Hoyt, WT.; McCullough, ME. Issues in the multimodal measurement of forgiveness.. In: Worthington, EL., Jr., editor. *Handbook of Forgiveness*. Brunner-Routledge; New York: 2005. p. 109-123.
- Hoyt WT, Melby JN. Dependability of measurement in counseling psychology: An introduction to generalizability theory. *The Counseling Psychologist*. 1999; 27:325–352.
- Hoyt WT, Warbasse RE, Chu EY. Construct validity in counseling psychology research. *The Counseling Psychologist*. 2006:769–805.
- Kenny, DA. *Interpersonal perception: A social relations analysis*. Guilford; New York: 1994.
- Kenny DA, Mohr C, Levesque M. A social relations partitioning variance of dyadic behavior. *Psychological Bulletin*. 2001; 127:128–141. [PubMed: 11271751]
- Lakes KD, Hoyt WT. Applications of generalizability theory to clinical child psychology research. *Journal of Clinical Child and Adolescent Psychology*. in press.
- McCullough, ME.; Hoyt, WT.; Rachal, KC. What we know (and need to know) about assessing forgiveness constructs.. In: McCullough, ME.; Pargament, K.; Thoresen, C., editors. *Forgiveness: Theory, research, and practice*. Guilford; New York: 2000. p. 65-88.
- Messick, S. Validity.. In: Linn, RL., editor. *Educational measurement*. 3rd ed.. Macmillan; New York: 1989. p. 13-103.
- Ross, L.; Nisbett, RE. *The person and the situation: Perspectives of social psychology*. McGraw-Hill; New York: 1991.
- Schmidt FL, Hunter JE. Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*. 1996; 1(2):199–223.
- Schmidt FL, Le H, Ilies R. Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*. 2003; 8(2):206–224. [PubMed: 12924815]
- Shavelson, RJ.; Webb, NM. *Generalizability Theory: A primer*. Sage; Newbury Park, CA: 1991.

Table 1

DPICS categories analyzed

Behavioral category	Description of behavior
<i>P2: Statement</i>	Coded when a phrase gives an account of the immediate situation or of unstated feelings. It must also be free of praise or criticism (e.g. 'I bet you're feeling sad because that toy broke')
<i>P3: Question</i>	A question is a phrase that gives an account of the situation in question form. A question may require a verbal response, but not a behavioral response (e.g. 'I made a tall tower, didn't I?')
<i>P4: Irrelevant verbalization</i>	Coded when the caretaker introduces a topic in either question form or as a comment that is irrelevant to the immediate activity of the parent or child (e.g. Child: 'Look at my pretty picture, mommy.' Parent: 'We need to go buy some more milk on our way home.')
<i>P6: Indirect command</i>	An implied or vague order that requires a behavioral response. It may also be a command stated in question form (e.g. 'Why don't you sit down next to me?')
<i>P7: Indirect command/ no opportunity</i>	When a child is given a command and is not given enough time (less than 5 s) to complete the task or if the command is too vague for the child to comply, it is coded as no opportunity (e.g. 'Be good' which is too vague for compliance)
<i>P8: Indirect command/compliance</i>	Coded when the child obeys within 5 s of a command. The child must at least make an attempt to obey within those 5 s (e.g. Parent: 'Can you bring me all those crayons?' Child: Starts to hand parent crayons one at a time as 5 s elapse)
<i>P9: Indirect command/non compliance</i>	Coded when child fails to comply after 5 s of a command (e.g. Parent: 'Let's put all the toys away now, okay?' Child: Proceeds to play with toys for 5 s)
<i>P10: Direct command</i>	An order from the caretaker to the child that gives enough information so that child can begin to obey (e.g. 'Put your hands on your lap')
<i>P11: Direct command/no opportunity</i>	When a child is given a direct command and is not given enough time (less than 5 s) to complete the task or parent completes the task for them (e.g. Parent: 'Hand me the red crayon.' After two seconds, parent reaches out and grabs the crayon)
<i>P12: Direct command/compliance</i>	Coded when the child obeys within 5 s of a command. The child must at least make an attempt to obey within those 5 s (e.g. Parent: 'Bring me the doll.' Child: Starts to walk over to the doll and picks it up as five second elapse)
<i>P14: Negative command</i>	A critical statement that tells the child what not to do (e.g. Parent: 'Stop running around!')
<i>P17: Reflective statement</i>	Coded when the caretaker repeats the basic content of a child's verbalization. It may contain all or part of what the child said and can be elaborated (e.g. Child: 'The girl is driving the car.' Parent: 'The girl is driving the car to the store.')
<i>P19: Descriptive comment/encouragement</i>	A phrase or statement that focuses on what the child is doing in the immediate present. It is descriptive of the child's activity or feelings (e.g. Parent: 'You are jumping up and down.'). A statement or phrase of encouragement expresses a positive feeling for the child's activity. It is almost sounds like a compliment but it does not include an evaluative word (e.g. Parent: 'You are so quick!')
<i>P34: Physical Positive</i>	Coded when caretaker has bodily contact with child, the contact may be positive or neutral (e.g. kiss, hug or accidental touch such as brushing of the hands while playing with blocks)
<i>P37: Positive Affect (verbal)</i>	Coded for a verbal expression on the part of the child that expresses pleasure in an activity, object, in someone else, or in something they have done (e.g. Child: I like playing this game with you, mom)
<i>P42: Positive Affect nonverbal</i>	An expression of pleasure or warmth that is directed at the parent. The behavior must be seen or heard by the parent to be counted (e.g. Child: Laughing with or smiling at parent while making eye contact)
<i>P45: Destructive</i>	Behavior on the part of the child is coded as destructive when it means to damage objects. Self-destructive behaviors are also coded as destructive (e.g. Child: Beats dolls head on table)

Table 2

Component definitions for two-facet generalizability analysis (DRS)

Dyads (D)	Universe score for dyad d (deviation from grand mean, averaged over raters and situation)
Raters (R)	Rater effect for rater r (rater leniency, averaged over dyads and situations)
Situations (S)	Situation effect for situation s (deviation from grand mean, averaged over dyads and raters)
DR	Idiosyncratic perception of dyad d by rater r (averaged over situations)
DS	Idiosyncratic response of dyad d on situation s (averaged over raters)
RS	Idiosyncratic leniency of rater r on situation s (averaged over dyads)
DRS,e	Idiosyncratic perception of dyad d by rater r on situation s , confounded with random error

Note: Components contributing to each score X_{DRS} , representing the rating for dyad d by rater r on situation s of the DPICS. In this table, components are defined as *effects*—deviations from the grand mean for all persons, raters, and situations. In GT, we will be interested in the *variance* of each component, which provides information about the importance of the corresponding effect in determining observed variance in ratings. Adapted from Lakes and Hoyt (in press).

Table 3

Raw and percentage variance estimates (DRS)

	Variance estimates (% total variance)							CSC index	
	<i>D</i>	<i>R</i>	<i>S</i>	<i>DR</i>	<i>DS</i>	<i>RS</i>	<i>DRS, e</i>	<i>Var(D)/[var(D)+var(DS)]</i>	
<i>P2: Statement</i>	0.0018 (11%)	0.0031 (19%)	0.0018 (11%)	0.0011 (7%)	0.0037 (22%)	0.0003 (2%)	0.0047 (29%)	0.3238	
<i>P3: Question</i>	0.0034 (9%)	0.0011 (3%)	0.0217 (60%)	0.0003 (1%)	0.0054 (15%)	0.0000 (0%)	0.0045 (12%)	0.3890	
<i>P4: Irrelevant verbalization</i>	0.0009 (7%)	0.0036 (26%)	0.0002 (1%)	0.0012 (9%)	0.0024 (17%)	0.0002 (1%)	0.0054 (39%)	0.2786	
<i>P6: Indirect command</i>	0.0019 (9%)	0.0010 (5%)	0.0069 (34%)	0.0008 (4%)	0.0030 (15%)	0.0006 (3%)	0.0060 (30%)	0.3826	
<i>P7: Indirect command/ no opportunity</i>	0.0024 (14%)	0.0028 (16%)	0.0031 (18%)	0.0011 (6%)	0.0017 (10%)	0.0002 (1%)	0.0061 (35%)	0.5890	
<i>P8: Indirect command/compliance</i>	0.0003 (2%)	0.0015 (11%)	0.0024 (18%)	0.0010 (8%)	0.0021 (16%)	0.0002 (2%)	0.0057 (43%)	0.1150	
<i>P9: Indirect command/ noncompliance</i>	0.0023 (23%)	0.0002 (2%)	0.0010 (10%)	0.0007 (7%)	0.0013 (13%)	0.0000 (0%)	0.0045 (45%)	0.6307	
<i>P10: Direct command</i>	0.0022 (10%)	0.0022 (10%)	0.0051 (23%)	0.0003 (2%)	0.0062 (28%)	0.0000 (0%)	0.0061 (27%)	0.2643	
<i>P11: Direct command/no opportunity</i>	0.0012 (10%)	0.0009 (7%)	0.0012 (10%)	0.0009 (8%)	0.0032 (27%)	0.0000 (0%)	0.0044 (37%)	0.2679	
<i>P12: Direct command/compliance</i>	0.0014 (9%)	0.0012 (7%)	0.0025 (16%)	0.0006 (4%)	0.0042 (27%)	0.0003 (2%)	0.0054 (35%)	0.2523	
<i>P 14: Negative command</i>	0.0010 (12%)	0.0005 (6%)	0.0000 (0%)	0.0003 (4%)	0.0021 (25%)	0.0000 (0%)	0.0043 (52%)	0.3299	
<i>P17: Reflective statement</i>	0.0022 (15%)	0.0003 (2%)	0.0046 (31%)	0.0004 (3%)	0.0027 (18%)	0.0007 (5%)	0.0040 (27%)	0.4425	
<i>P19: Descriptive comment/ encouragement</i>	0.0011 (8%)	0.0026 (19%)	0.0000 (0%)	0.0015 (12%)	0.0035 (26%)	0.0002 (2%)	0.0045 (34%)	0.2342	
<i>P34: Physical Positive</i>	0.0001 (1%)	0.0013 (13%)	0.0008 (8%)	0.0006 (6%)	0.0034 (35%)	0.0004 (4%)	0.0033 (33%)	0.0398	
<i>P37: Positive affect (verbal)</i>	0.0027 (26%)	0.0005 (5%)	0.0006 (6%)	0.0006 (6%)	0.0021 (20%)	0.0001 (1%)	0.0038 (36%)	0.5643	
<i>P42: Positive affect (nonverbal)</i>	0.0011 (11%)	0.0002 (2%)	0.0012 (12%)	0.0009 (8%)	0.0024 (23%)	0.0000 (0%)	0.0045 (44%)	0.3185	
<i>P45: Destructive</i>	0.0006 (24%)	0.0001 (3%)	0.0000 (0%)	0.0003 (13%)	0.0002 (10%)	0.0000 (0%)	0.0012 (50%)	0.7024	

Note. *D* = dyad, *R* = rater, *S* = situation, CSC = cross-situational consistency.